

Univerzita Karlova

Filozofická fakulta

Ústav anglického jazyka a didaktiky



Diplomová práce

bc. Tomáš Savčenko

Effects of Semantic Network Structure of English on Word Processing

Efekty struktury sémantické sítě angličtiny na zpracování slov

Praha 2024

vedoucí práce: doc. Dr. phil. Eva Maria
Luef, Mag. phil.

Acknowledgements

I would like to express my sincere thanks to my supervisor, Eva Luef, for all her support, insightful advice, and prompt responses throughout my writing process. Not only that but also her introduction to the topic of language networks in her course made the present diploma thesis possible.

Prohlašuji, že jsem diplomovou práci vypracoval/a samostatně, že jsem řádně citoval/a všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 12. 8. 2024

.....

Souhlasím se zapůjčením diplomové práce ke studijním účelům.

I have no objections to the MA thesis being borrowed and used for study purposes.

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Abbreviations

BERT - Bidirectional Encoder Representations from Transformers

LLM – large language model

MALD database - Massive Auditory Lexical Decision database

NLP – natural language processing

PMI - pointwise mutual information

Abstract

This diploma thesis explores the intersection of computational approaches to language, network science, and psycholinguistic research of word production. The thesis introduces network science together with its formalism and application in linguistic research as phonological and semantic networks. It introduces relevant psycholinguistic experimental research of word processing, namely lexical decision task that is indicative of processing efficiency. Finally, large language models and word vectors are introduced. The aim of this thesis is to construct a semantic network of English based on word vectors computed by BERT language model from a sample of the TV Corpus. A structure of the resulting semantic network is analysed in the light of results from lexical decision task drawn from the MALD database that reflect word processing efficiency. The resulting semantic network has small-world structure implying that word vectors transformed into a semantic network can capture cognitively salient semantic relationships between words. Multiple linear regression analysis between degree centrality, closeness centrality, and clustering coefficient of words within the semantic network and reaction time for the same words from the MALD database did not show statistically significant relationship. Clustering coefficient appears to have slightly negative relationship to the reaction time that was approaching statistical significance implying that words from denser parts of the network are processed faster. Current results allow careful optimism for the use of semantic networks based on word vectors for the research of cognitive processes underlying language.

Keywords: semantic network, word processing, word vectors, machine learning

Abstrakt

Tato diplomová práce zkoumá průnik komputačních přístupů k jazyku, vědy o sítích a psycholinguistického výzkumu produkce slov. Práce představuje vědu o sítích spolu s jejím formalismem a aplikací v lingvistickém výzkumu ve formě fonologických a sémantických sítí. Představuje relevantní psycholinguistický výzkum zpracování slov, konkrétně *lexical decision task*, který vypovídá o efektivitě zpracování slov. Nakonec jsou představeny velké jazykové modely a vektory slov. Cílem této práce je zkonstruovat sémantickou síť angličtiny na základě slovních vektorů vytvořených jazykovým modelem BERT ze vzorku z The TV Corpus. Struktura výsledné sémantické sítě je analyzována ve světle výsledků z *lexical decision task* čerpaných z databáze MALD, které odrážejí efektivitu zpracování slov. Výsledná sémantická síť má strukturu malého světa (*small-world structure*), což znamená, že slovní vektory transformované do sémantické sítě mohou zachytit kognitivně salientní sémantické vztahy mezi slovy. Lineární regresní analýza mezi síťovými proměnnými *degree centrality*, *closeness centrality*, and *clustering coefficient* pro jednotlivá slova v sémantické síti a reakčním časem pro stejná slova z databáze MALD neprokázala statisticky významný vztah. Zdá se, že *clustering coefficient* má mírně negativní vztah k reakční době, který se blížil statistické významnosti, což znamená, že slova z hustších částí sítě jsou zpracovávána rychleji. Současné výsledky dovolují opatrný optimismus pro využití sémantických sítí založených na slovních vektorech pro výzkum kognitivních procesů, které jsou základem jazyka.

Klíčová slova: sémantická síť, zpracování slov, slovní vektory, strojové učení

Contents

1. Introduction	1
2. Theoretical background	2
2.1. Network Science	2
2.2. Psycholinguistics and Word Processing.....	9
2.3. Phonological networks	13
2.4. Semantic networks.....	19
2.5. Large Language Models and Word Vectors	27
3. Material and Method	32
3.1. Computation of the Semantic Network.....	33
3.2. Word Processing Data.....	35
4. Research.....	36
5. Conclusion.....	40
6. References	41
7. Résumé	45
8. Apendix.....	48

List of Figures

Figure 1 A comparison between small-world network and scale-free network	8
Figure 3 Example phonological network from Vitevitch et al. (2023)	15
Figure 4 Visualisation of a portion of the network made by Lakhzoum et al. (2021) where the node size and colour intensity reflect its degree	22
Figure 5 An example of a visualization of Wordnet's network for the word 'cat' made by Visuwords educational project	23
Figure 6 T-SNE two-dimensional projection of the word2vec vectors representing a selection of the most frequent words in its training corpus. Credit: Gastaldi (2021).....	29
Figure 7 Offset representing the gender relation, revealed in the embedding space by a PCA projection. Credit: Gastaldi (2021).....	30
Figure 8 Pattern in the embedding space (word2vec) corresponding to the comparative category (base, comparative and superlative forms). Credit: Gastaldi (2021).....	31
Figure 9 Pattern in the embedding space (word2vec) corresponding to conjugation of irregular verbs. Credit: Gastaldi (2021)	31
Figure 10 Visualization of semantic change based on word2vec. Credit: Hamilton et al. (2016).	32
Figure 11 Cosine similarity distribution.....	35
Figure 12 Semantic network based on cosine similarity measures of word vectors computed by BERT from a sample of The TV Corpus	37
Figure 13 Box plots of mean word reaction times and network measures.....	38
Figure 14 Added-variable plots	39

List of Tables

Table 1 Summary of relevant basic network measures	5
Table 2 Edge list and network of phonological neighbourhood of "cat"	16
Table 3 An overview of different edge types in semantic networks, adopted from Engelthaler and Hills in Network Science in Cognitive Psychology, Vitevitch (2019), p. 167.	19
Table 4 A part of the table containing 6664 word vectors with 768 dimension computed by BERT	34
Table 5 Relevant macro measures of the semantic network	37
Table 6 Summary of the regression model	39

1. Introduction

In recent years, the intersection of network science and linguistics has unveiled new ways of understanding the cognitive processes underlying language comprehension and production. Network science, a branch of mathematics dedicated to the study of networks consisting of nodes and their interconnections, offers rigorous methodologies for modelling and analysing complex systems, including the intricate structures of language. This thesis investigates the application of network science to explore the semantic network structure of English and its implications for word processing. By treating words as nodes and their semantic relationships as links, this study seeks to elucidate how the configuration of such network representing meaning influences how speakers process words in mind.

The conceptualization of language as a network is not a novel idea. Early foundations were laid by structuralist linguistics, particularly by Ferdinand de Saussure, who suggested that elements of language are interconnected within a system (Saussure, 1959). Contemporary theories, such as Goldberg's Construction Grammar, Croft and Cruse's Cognitive Linguistics and various other linguistic frameworks, have increasingly employed the implicit idea about the network structure of language (Goldberg, 2009; Croft & Cruse, 2004). Even more recently, researchers such as Michael Vitevitch and his colleagues have pursued the study of language through the lens of network science that explicitly and formally defines language structures as a network. The atheoretical stance of network science provides a neutral platform, focusing on empirical data rather than theoretical biases, thereby allowing for a rigorous exploration of language networks through metrics like degree centrality, clustering coefficient, and average path length that measure different features of networks.

This thesis specifically addresses the structure and dynamics of semantic networks—networks where nodes represent words and link denote semantic similarities. The research aims to understand how these network structures impact word processing, particularly in tasks requiring lexical retrieval. The use of large language models (LLMs), such as BERT (Bidirectional Encoder Representations from Transformers), has potential in linguistic research as it opens new doors to quantitative analysis of semantics from text. These models enable the extraction of high-dimensional word vectors from large corpora, which can then be transformed into a network and analysed to identify patterns and semantic relationships within the mental lexicon.

The study builds upon previous research in both phonological and semantic networks, aiming to replicate and extend findings related to word processing. It employs computational

methods to create semantic network from a sample corpus. The resulting network is analysed using tools from network science, and its properties are correlated with empirical data from lexical decision tasks, sourced from the Massive Auditory Lexical Decision (MALD) database (Tucker et al., 2019), that are indicative of word processing efficiency. The goal is to explore the influence of the structure of the semantic network on the efficiency of word processing.

By examining the structural features of the semantic network and word processing, this research contributes to the growing body of knowledge on the cognitive representation of language. It also evaluates the relevance and potential of current LLMs in linguistic research and cognitive science, offering insights into their applications and limitations.

Ultimately, this work aims to explore how network science can be combined with the LLMs to introduce novel ways how semantics can be studied quantitatively in the light of psycholinguistic research of how people process words. The results of such work could potentially contribute to the growing body of studies that focus on the use of language networks in diagnosing and addressing language-related cognitive impairments that can benefit from sensitive quantitative approaches to modelling meaning.

The chapters that follow will delve into the theoretical background of network science with a focus on language networks and psycholinguistic research of word processing. The chapters will also introduce word embeddings, a method of representing words as numerical representations that capture their meanings and relationships based on how they are used in large collections of text. Finally, a detailed findings of this study based on the analysis of the semantic network and data from psycholinguistic word retrieval experiments will be introduced, setting the stage for a discussion about the implications of semantic network structures for word processing and practical applications in cognitive science.

2. Theoretical background

2.1. Network Science

Network science, originally a field within mathematics, focuses on the study of complex systems represented as networks composed of nodes and the connections between them. It allows researchers to model and analyse the structure of various phenomena, enabling the identification of previously unrecognized properties, structures, or behaviours. Many real-world phenomena can often be conceptualized as networks; for instance, the World Wide Web can be represented with webpages as nodes and hyperlinks as links, power grids can be modelled with power plants and transformers as nodes and cables as links, and academic

papers can be linked by citations, with the papers themselves serving as nodes (Barabási & Pósfai, 2016). Similarly, language can be represented as a network, where linguistic units, such as words, act as nodes, and the connections between them are defined by relationships such as phonological similarity (Chan & Vitevitch, 2009) or semantic similarity (Steyvers & Tenenbaum, 2005).

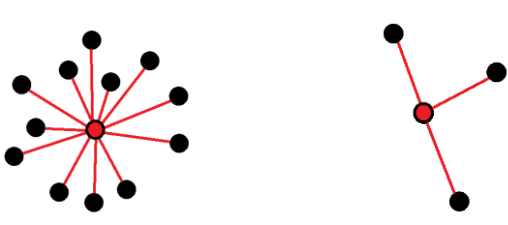
The idea of treating elements of language as nodes in a network is not novel, as it can be traced back at least to Saussure's structuralist perspective on language (Saussure, 1956). However, Saussure merely suggested the concept without formalizing it. Modern linguistic theories, such as Construction Grammar, or Cognitive Linguistics have also adopted the notion of a language network (Goldberg, 2009; Croft & Cruse, 2004), though these approaches often concentrate on the analysis of specific phrases instead on the broader picture. Furthermore, these approaches are typically embedded within comprehensive theoretical frameworks that include certain assumptions about the underlying mechanisms and principles of language. In contrast, the current study seeks to apply network science principles in a non-theoretical manner, focusing on empirical data and the statistical properties of language networks. This approach allows for the exploration of new insights into language without the necessity of engaging in theoretical debates or choosing an appropriate theoretical framework. Consequently, network science becomes a tool for the precise measurement and description of language data.

Network analysis provides a framework for understanding the structural characteristics of networks, which can be studied at different levels of detail. By analysing networks from various levels of detail, researchers can gain insights into both local and global properties. At the micro level, network analysis focuses on individual components, such as nodes and links, examining specific metrics pertaining to individual elements of the network. In the domain of social networks, an example of micro-level network analysis could involve examining individual users of social media. We can calculate network metrics that indicate the importance or influence of a particular node within a network. For instance, in a social media network, a metric known as "degree centrality" can be measured by assessing how many direct connections (or "friends" or "followers") a user has. Another micro-level metric could be "betweenness centrality," which measures the extent to which a particular user acts as a bridge between other users in the network. A user with high betweenness centrality would be crucial for connecting disparate parts of the network, as they lie on the shortest path between many other nodes. Analyzing these specific metrics for individual users helps to identify key

influencers or pivotal connectors within the social network. This granular approach allows for the exploration of the behavior and role of specific elements within the network.

Conversely, at the macro level, analysis shifts to a broader view, considering the overall structure and organization of the network. This involves calculating aggregate measures, such as average values, that characterize the network as a whole. These macro-level metrics provide a comprehensive understanding of the network’s global properties, including its connectivity, robustness, and overall topology. In the domain of language networks, macro-level analysis can be particularly useful in studying the overall connectivity and organization of a speaker’s mental lexicon, which is the network of words inside speakers mind acting like a mental dictionary (Aitchison, 2012). For instance, when examining individuals with aphasia, a language impairment often resulting from brain injury or as a neurodegenerative disorder, researchers can analyze the global properties of language networks constructed from a corpus of aphasic speech.

A specific example could involve comparing the language networks of individuals with aphasia to those of healthy speakers. By analyzing metrics such as network density (the ratio of actual connections to possible connections) and average path length (the average number of steps it takes to connect any two nodes in the network), researchers can assess the overall connectivity and structure of the language network. In individuals with aphasia, the network may show reduced connectivity, indicated by a lower density and longer average path lengths, reflecting the loss of access to certain words and weakened associations between them. Such macro-level analysis can reveal significant differences in the robustness and organization of the language networks between impaired and non-impaired speakers. Such differences in the overall topology of the network could potentially help distinguish between different types of aphasia or reveal minute changes to the mental lexicon that would otherwise be left unnoticed in raw speech or basic corpus analysis. By employing both micro and macro-level analyses, researchers can obtain a nuanced understanding of networks, capturing both detailed and overarching patterns within the data.

Measurement	Definition	Comparison between Higher and Lower Values
Degree centrality	The number of links of a node relative to the overall network size.	

Shortest path length	The shortest route between two nodes which equals the number of links it contains.	
Clustering coefficient	A measure of local clustering which signals how often the neighbours of a node tend to be neighbours of each other.	
Closeness centrality	A measure signalling how close a node is to all other nodes.	

Table 1 Summary of relevant basic network measures

Table 1 introduces some of the basic metrics relevant for the present work that network analysis employs to understand the structure and dynamics of networks, both at the level of individual nodes and across the entire network. One fundamental metric is degree centrality, which indicates the number of direct connections (links) a node has. This metric can highlight influential nodes within a network, such as central figures in social networks who are connected to many others. The concept extends to the network macroscopic level through the calculation of the average degree, which represents the average number of connections per node in the entire network, providing an overview of its general connectivity.

Another critical measure is the shortest path length, which refers to the minimum number of links required to connect two nodes. The network-wide macroscopic equivalent, the average shortest path length, is determined by calculating the shortest paths between all pairs of nodes and averaging these values. This metric offers insights into the overall efficiency of the network's structure, indicating how quickly information or resources can be transmitted throughout the network depending on what the network models.

The clustering coefficient further enriches the understanding of networks by quantifying the extent to which nodes in a network tend to form tightly-knit groups or communities. Defined as the degree to which neighbours of a node are interconnected, the clustering coefficient measures the presence of local clustering. A high average clustering coefficient, computed for the entire network, suggests that the network is composed of numerous small groups where most neighbours of a node are also neighbours of each other. Clustering coefficients range from 0 to 1, where 1 indicates that all neighbours of a node are fully interconnected, and 0 indicates no such interconnections among neighbours (Watts & Strogatz, 1998).

Closeness centrality is another relevant metric that helps identify key nodes, particularly in the so-called small-world networks that are introduced further below. It is defined as the reciprocal of the sum of the shortest-path lengths from one node to all other nodes in the network. In practice this means that nodes with high closeness centrality can be reached quickly from all other nodes, making them crucial for efficient information flow. In small-world networks, nodes that serve as hubs often exhibit high closeness centrality, in addition to a high degree centrality (Barabási & Pósfai, 2016). These hubs play a critical role in maintaining the network's overall connectivity and functionality.

By employing these metrics, whether at the microscopic level of individual nodes or across the network as a whole, researchers can obtain a nuanced understanding of both the local and global connectivity patterns. This comprehensive analysis enables the identification of structural features, such as tightly-knit communities or critical hubs, and provides valuable insights into the resilience, efficiency, and overall topology of the network. This work's scope lies mainly in the macroscopic domain; therefore, the average values of the relevant metrics will be analysed in the practical part.

So-called small-world networks are a distinctive type of network characterized by high local clustering and short path lengths, allowing for efficient information transfer and robust connectivity. This network structure is commonly found across a variety of social and cognitive phenomena, including social networks and language. In these contexts, small-world networks facilitate efficient communication and quick access to information, making them well-suited for complex, interconnected systems (Watts & Strogatz, 1998). For instance, in social networks, this structure enables rapid dissemination of information and fosters strong community ties, while in cognitive systems, it supports efficient processing and retrieval of information. The emergence of small-world networks in these domains underscores their role in maintaining functional and adaptive systems.

Small-world networks are characterized by specific structural properties that can be identified through metrics such as the average degree, clustering coefficient, closeness centrality, and average shortest path length. A high clustering coefficient indicates significant local clustering, meaning that nodes are densely interconnected within localized areas of the network. This is complemented by lower average shortest path lengths relative to a random graph and relatively high closeness centrality, which suggest efficient global connectivity. In such networks, it is possible to traverse the entire network by passing through a relatively small number of links, facilitating quick access from any node to any other (Watts & Strogatz, 1998). Additionally, small-world networks typically exhibit a higher average degree, reflecting a generally well-connected and clustered structure (Barabási & Pósfai, 2016). These networks usually have efficient connectivity and are resilient to damage (Watts and Strogatz 1998). For example, a small-world network representing a power grid would be resilient to cable damages avoiding blackouts with electricity spreading efficiently to all subscribers. In the context of understanding language as a dynamic cognitive system that can be represented as a small-world network, being a resilient network with efficient connectivity translates into speakers who can efficiently communicate, retrieve words instantaneously in a conversation, and effectively form thoughts, as well as such speakers whose communicative abilities are resilient to language impairments involving changes in the brain such as aphasia.

Figure 1 below contrasts a predominantly small-world network with a predominantly scale-free network to illustrate how can other types of networks differ from a small-world network. Scale-free networks are characterized by a power-law distribution in their degree centrality. This means that while most nodes have relatively few connections, a few nodes—known as hubs—have a disproportionately high number of connections. The presence of these highly connected hubs is a prominent feature of both scale-free networks and small-world networks (Barabási & Albert, 1999). The networks are not mutually exclusive. In fact, many real-world networks exhibit properties of both. For instance, a network can have a small-world structure with a high clustering coefficient and short path lengths, while also having a power-law degree distribution indicative of a scale-free structure. For example, in the context of semantic networks, Steyvers and Tenenbaum discuss that both small-world and scale-free properties have been observed in networks based on word associations, Wordnet and Roget's Thesaurus (for details see chapter 2.4). The small-world structure is indicated by the combination of high local clustering and short average path lengths, while the scale-free nature is seen in the power-law distribution of node connections, where a few words or concepts serve as hubs with many connections. The difference between the two networks in

figure 1 lies in the fact that the second network has only the scale-free characteristics with no local clusters while the first one also has small-world characteristics.

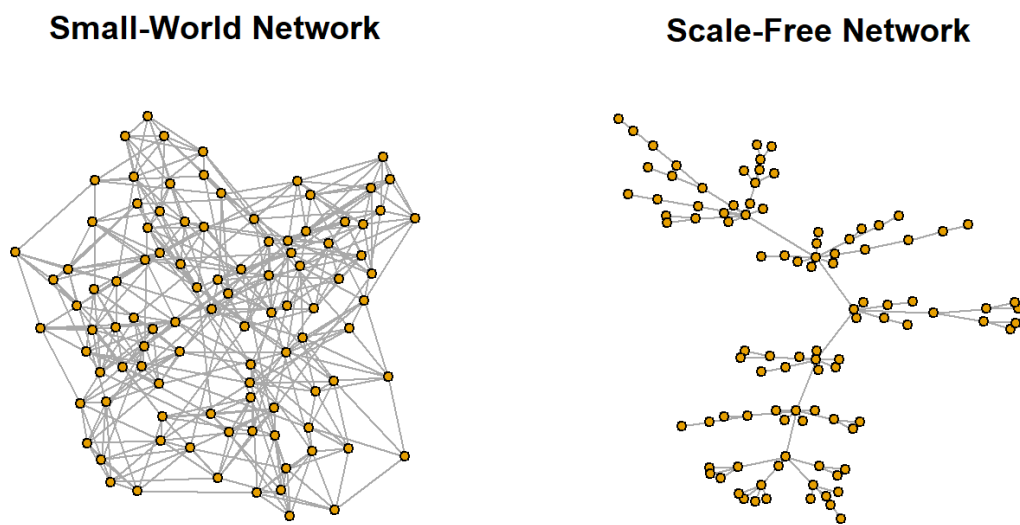


Figure 1 A comparison between small-world network and scale-free network

Small-world networks also contain hubs, but due to the high overall clustering of the network, they are less vulnerable to targeted attacks. Even if a hub is compromised, other well-connected nodes can compensate and handle the additional load, maintaining the network's stability. The uneven distribution of links results from preferential attachment, a process where new nodes are more likely to connect to already highly connected nodes, reinforcing the connectivity of hubs (Albert & Barabási, 2002). This concept can be related to children's vocabulary acquisition, where children are more likely to learn new words that are similar to or associated with words they already know. This phenomenon suggests that the acquisition of new vocabulary often builds upon existing knowledge, creating a network where frequently used or familiar words serve as hubs, facilitating the integration of new words.

Research in language acquisition supports this analogy. For example, Hills et al. (2009) found that children are more likely to acquire new words that are semantically related to their existing vocabulary. This process, often referred to as "semantic clustering," indicates that new words with connections to known words are learned more easily because they can be integrated into the child's existing mental lexicon. The existing words act as anchors, making it easier for children to relate new information to what they already understand.

It is crucial to note that small-world characteristics of a network are not categorical features; instead, they are represented by continuous values of metrics like average degree,

clustering coefficient, and closeness centrality; hence a network can have more or less of small-world characteristics. Both network types are pervasive in real-world systems but cater to different dynamics and resilience strategies within these systems with small-world networks being typically found in language networks.

Finally, weighted networks provide a more nuanced representation of relationships between nodes by assigning a value, or weight, to each link. This weight reflects the strength or intensity of the connection, adding depth and detail to the analysis of the network's structure. The use of weighted networks allows researchers to capture variations in the significance or influence of relationships, thereby enriching the dataset with more precise information (Barabási & Pósfai, 2016). The method for determining these weights varies depending on the type of relationship being represented. For example, in a social network of phone calls, the weight of a connection between two individuals can be represented by the total number of minutes they spend talking. This measure not only indicates that a relationship exists but also provides insight into the strength of the connection, as longer call durations suggest a stronger or more significant relationship. Similarly, in language networks, weights can be assigned to connections based on specific criteria relevant to the study's focus. For instance, in a semantic network where nodes represent words, the weight of an edge might reflect the degree of semantic similarity between the words, which can be quantified using various measures such as cosine similarity in vector space models. The inclusion of weights in network analysis enhances the ability to discern subtle differences and complexities within the network.

As we have explored, network science provides a powerful framework for understanding the structure and dynamics of complex systems, including language. The next chapter will delve into the psycholinguistic research of word processing and mental lexicon and follow up with the introduction of relevant phonological network research where network science meets psycholinguistics. By combining these disciplines, we can uncover meaningful insights into the cognitive architecture of language, offering a deeper understanding of how words are processed and interconnected in the human mind.

2.2. Psycholinguistics and Word Processing

Network science enables us to formalize and describe patterns in language use that reveal its internal organization. However, meaningful insights into the cognitive architecture of language emerge when these network models are combined with psycholinguistic research. Network science is particularly well-suited for modelling the mental lexicon, which is

generally understood in psycholinguistic research to have a structure akin to an interconnected network (Aitchison, 2012; Vitevitch, 2008). For instance, by analysing the mental lexicon as a network, researchers can explore how words are stored and retrieved, how phonological or semantic relationships influence word recognition, and how frequency and context affect word processing (Steyvers & Tenenbaum, 2005).

Word processing in psycholinguistics refers to the understanding and production of words by speakers. This inherently involves the mental lexicon, which is a theoretical repository of a speaker's units of language. These units can encompass a wide range of elements, including concepts, morphemes, words, phrases, or even longer segments. The precise nature of the mental lexicon remains a topic of ongoing debate within linguistics, with questions surrounding the size and nature of these items, as well as whether the lexicon includes the rules or patterns of use and combination of these units (Aitchison, 2012). Despite the importance of this debate, it is beyond the scope of this work. In accord with the tradition of network research in language, this study conceptualizes the units of language within networks as words. This choice is primarily motivated by convenience and practicality, and the study remains neutral regarding the exact nature of the mental lexicon's units. However, there is substantial evidence to support the significant role that words play in language processing.

This work focuses on the processing of words. In relation to the mental lexicon, this encompasses the mechanisms by which speakers retrieve words from their lexicon for production, as well as how they identify and process words for understanding. In psycholinguistics, word processing is typically studied through various experimental methods. Some of the relevant designs include word identification tasks, where participants are asked to recognize and identify words presented to them; lexical decision tasks, which involve determining whether a string of letters is a real word or a non-word; and naming tasks, where participants are asked to produce the name of an object or read aloud a word (Chan & Vitevitch, 2009; Siew & Vitevitch, 2016). Additionally, more general psychological experimental designs, such as free recall and cued recall tasks, are employed to study memory and retrieval processes. In free recall tasks, participants are asked to recall as many items as possible from a previously presented list, while in cued recall tasks, they are provided with cues to aid their recall. These experimental tasks can be administered in different modalities, typically involving either visual or auditory language stimuli. For instance, in visual word recognition tasks, participants might read words presented on a screen, while in auditory tasks, they would listen to spoken words and respond accordingly.

In a word identification task, participants are instructed to identify words played to them with various degrees of noise distortion added. Typically, the words belong to different groups that share a common feature, the independent variable, such as high frequency or low frequency. Researchers examine the influence of this independent variable on the participants' ability to correctly identify the words, which serves as the dependent variable. This dependent variable reflects the ease of word processing. For example, words of higher frequency are generally processed more quickly and accurately than words of lower frequency, which provides insights into the cognitive mechanisms underlying word recognition (Goldinger, 1996). Naming tasks, on the other hand, can be either auditory or visual. In these tasks, a word is presented to participants either on a screen (visual modality) or through auditory means (auditory modality). Participants are then required to repeat the word as quickly and accurately as possible. Researchers measure two primary dependent variables in this task: the reaction time from the presentation of the word to the participant's production of the word, and the accuracy of the participant's response. Similar to word identification tasks, the words used in naming tasks are categorized into different groups based on a common feature, which serves as the independent variable. The influence of this feature on reaction time and accuracy is analysed to understand its impact on the ease of word processing. For instance, researchers might compare reaction times and accuracy rates for high-frequency words versus low-frequency words. High-frequency words are generally recognized and named more quickly and accurately, reflecting their easier and more efficient processing within the mental lexicon. This helps in understanding how different factors, such as word frequency, influence the cognitive processes involved in word retrieval and production (Jescheniak & Levelt, 1994). Moreover, the independent variable can also be one of the network measures, such as degree centrality or clustering coefficient, which link language network measurements such as degree centrality or clustering coefficient with psycholinguistic experiment results. These network measures can influence word processing. For example, words with higher degree centrality may be retrieved more quickly due to their extensive connections, while words with a higher clustering coefficient may benefit from being part of tightly knit semantic or phonological clusters (Kenett et al., 2014). By incorporating network measures as independent variables, researchers can bridge language network measurements with psycholinguistic experiment results

The lexical decision task can be administered in either auditory or visual modalities. In this task, participants are presented with pairs of stimuli – one being a real word and the other a non-word. The non-word is crafted to resemble a real word but contains distortions or

alterations that prevent it from being a legitimate word. Participants must quickly and accurately determine which of the two stimuli is the real word. The word and non-word pairs are categorized into different groups based on a common feature, which serves as the independent variable. This feature could be something like word frequency, orthographic neighbourhood size, or network measures such as degree centrality or clustering coefficient. Researchers measure two primary dependent variables in this task: reaction time, which is the time it takes for participants to make their decision, and accuracy, which is the correctness of their response. These dependent variables provide insights into the ease and efficiency of word processing. For example, higher frequency words are generally recognized faster and more accurately, reflecting more efficient processing within the mental lexicon. By manipulating the independent variable, researchers can examine its influence on reaction time and accuracy. This allows them to test hypotheses about how different linguistic and network properties affect word processing. For example, if words with higher degree centrality are processed more quickly and accurately, this would suggest that network connectivity plays a significant role in lexical access and retrieval.

Finally, cued recall is a psychological experimental design used to examine how people remember information. Unlike tasks that are directly linked to linguistic knowledge, cued recall can be applied to a broad range of memory contexts. In a cued recall experiment, participants are presented with a cue or hint that helps them recall a previously learned item or association. This method contrasts with free recall, where participants must remember information without any cues, and recognition tasks, where they must identify previously learned information from a list of options. In cued recall, the cue can be a word, phrase, or any form of hint related to the target memory. For example, if participants were previously shown a list of paired associates like “dog-bone” and later given the cue “dog”, they are expected to recall the associated word “bone”. The primary measure in cued recall experiments is the percentage of accurately recalled items. This measure reflects the effectiveness of the cue in aiding memory retrieval and provides insights into the cognitive processes underlying associative memory. Cued recall experiments are particularly valuable for uncovering the associative cognitive networks that influence word retrieval. By analysing how different types of cues (semantic, phonological, or contextual) affect recall performance, researchers can infer the structure and strength of associations within the mental lexicon. For instance, semantic cues (words related in meaning) might be more effective in aiding recall than phonological cues (words that sound similar), suggesting stronger semantic associations in the mental lexicon (Anderson, 2013). Furthermore, cued recall tasks can reveal how different

variables, such as the type of cue or the length of the retention interval, impact memory retrieval. These experiments could potentially be designed to explore the role of network measures like degree centrality or clustering coefficient in word retrieval. For instance, words with higher centrality in a semantic network might be recalled more easily when given appropriate cues, indicating the importance of network connectivity in memory processes (Kenett et al., 2017).

One way to combine these psycholinguistic experimental designs with network science is to use network measurements as the independent variable and examine its influence on word processing as was suggested throughout this chapter. For example, researchers can conduct a lexical decision task with two groups of words distinguished by their clustering coefficients and examine how this network measure affects reaction time and accuracy in word recognition. Experiments that integrate network science and psycholinguistics have already been conducted with phonological networks. These studies utilize network metrics such as degree centrality and clustering coefficient to explore network structure impacts word retrieval and recognition (Vitevitch & Goldstein, 2014). By using these rigorous network science formalisms together with psycholinguistic experiments, researchers can uncover valuable insights into the cognitive underpinnings of language.

The following chapter delves into this intersection by focusing on phonological networks. It argues that the combination of network science methodologies with psycholinguistic research on lexical retrieval can be useful to investigate the mental lexicon's organization and functionality. The chapter 2.3 presents relevant studies that combine phonological network analysis with psycholinguistic experiments, demonstrating how these interdisciplinary approaches can advance our understanding of language processing.

2.3. Phonological networks

Network science has proven fruitful in numerous domains of scientific inquiry, but it remains a relatively young field within linguistics, despite the implicit recognition of the “networkness” of the language system dating back to at least Saussure's structuralist theories (Saussure, 1956). The idea that language can be conceptualized as a network is not new, but its formal application using network science methodologies has only recently gained traction.

There are various ways to operationalize the nodes and edges within a linguistic network for meaningful analysis. In phonological networks, which have garnered considerable attention, nodes typically represent individual words, while edges denote some form of phonological similarity between these words. A significant body of work has been

conducted on phonological networks, particularly by Michael S. Vitevitch and his colleagues, who defined the relationship between words in these networks by the substitution, addition, or deletion of a single phoneme to form what is known as a “phonological neighbor.” For example, the words “hat,” “cut,” “cap,” “scat,” and “_at” are considered phonologically similar to the word “cat” (Vitevitch, 2008: 3). In the network, all of these words would be linked based on this defined phonological relationship. Phonological networks constructed using such criteria display small-world characteristics, a property observed in many naturally occurring networks. Small-world networks are characterized by a high clustering coefficient and short average path lengths, meaning that most nodes (words) are not neighbours of one another but can be reached from one another by a small number of steps.

An example of such a network with small-world properties is illustrated in Figure 2, which depicts the phonological neighbourhood of the words “peach” and “speak.” In this example, we can visually observe at least three distinct clusters around “peach,” with “peach” itself acting as a hub—a node with a high degree centrality. This network’s clustering coefficient and closeness centrality are higher than those of a random graph, indicating the presence of significant local clustering and efficient overall connectivity (Watts & Strogatz, 1998). These characteristics suggest that phonological networks, like other small-world networks in other social and cognitive domains, may be a relevant component in the cognitive processes underlying language.

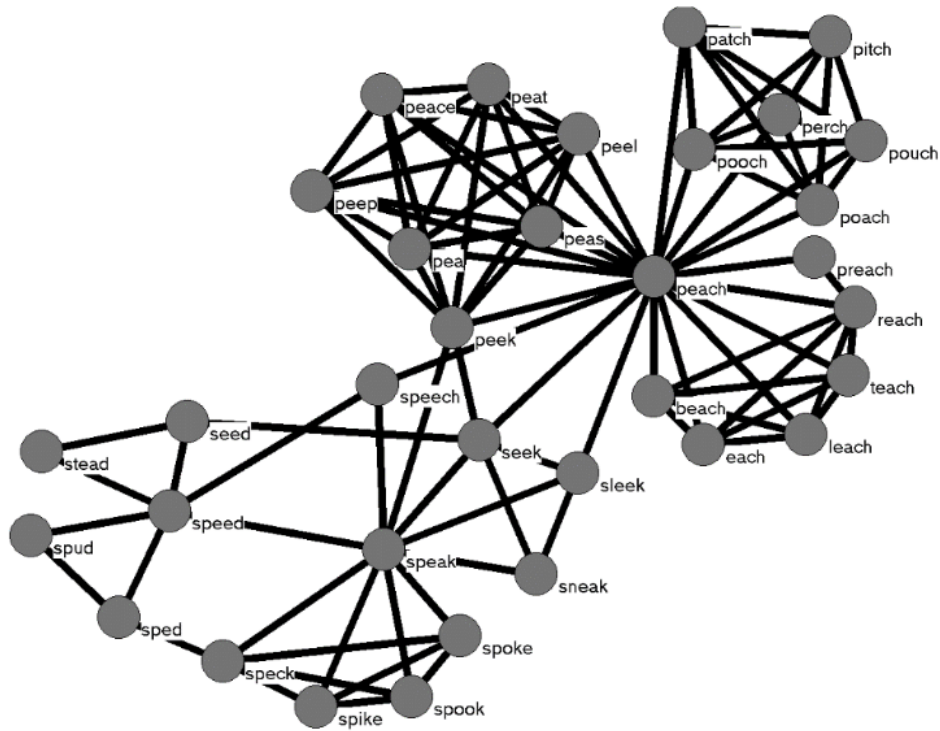


Figure 2 Example phonological network from Vitevitch et al. (2023)

While it is interesting and aesthetically pleasing to observe a structure emerge within a network, it is essential to question how such a language network contributes to our understanding of language. A network is more than just a visual representation; it is a robust data structure that enables the representation and analysis of complex relationships within a dataset. In the context of language, networks provide a rigorous framework for modelling the connections between words, phrases, or concepts, allowing researchers to quantify and analyse these relationships through various network metrics. These network-based approaches reveal patterns and structures that are often obscured in traditional linear representations of language, making networks not only valuable for visualization but also for deeper analytical insights. For instance, centrality measures can identify key words or concepts that serve as hubs within the mental lexicon, while clustering coefficients can reveal the degree to which words group together based on phonological or semantic similarities (Newman, 2018), providing insights into how language is organized and processed.

As a data structure, a network can also be represented as an edge list. In an edge list, the network is defined by listing all the connections (or edges) between the nodes. Each row in an edge list typically contains two elements – the two nodes that are connected. Optionally, a third element can be included to represent the weight of the connection, indicating the strength or importance of the relationship between the nodes. Table 2 compares a simple edge

list and network for the phonological neighbourhood of the word “cat” mentioned above. In this example, the edge list provides a clear and concise way to enumerate all phonological connections between “cat” and its neighbours. When visualized as a network, these connections reveal the structure of the phonological neighbourhood, with “cat” acting as a central node.

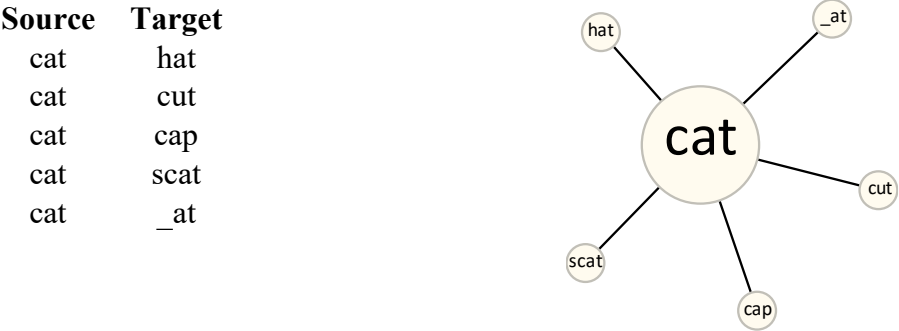


Table 2 Edge list and network of phonological neighbourhood of “cat”

Vitevitch and his colleagues have extensively researched phonological networks from a cognitive perspective, demonstrating how network science can illuminate various aspects of language processing. For instance, in his 2008 work, Vitevitch argues that graph theory—a mathematical framework used to study networks—can significantly enhance our understanding of word learning by mapping words and their phonological similarities into a network. Vitevitch highlights that the connectivity within these phonological networks plays a crucial role in language acquisition. Specifically, he suggests that words with a higher degree centrality, meaning they are phonologically similar to many other words, may be learned earlier during acquisition. These highly connected words, hubs in the network, can facilitate the acquisition of new words by providing multiple phonological pathways for learners to explore and reinforce (Vitevitch, 2008). Such network growth mechanism is called preferential attachment and was identified in other language acquisition studies (in children but also adults learning new words). For example, Storkel & Morrisette (2002) found that preschool children learn words with common sound sequences (i.e., words in dense phonological neighbourhoods with high degree centrality and clustering coefficient) faster than those with more rare sound sequences. Similarly, Storkel, Armbrüster, and Hogan (2006) found that university students learn new words from dense phonological neighbourhoods more quickly too.

In their 2009 study, Chan and Vitevitch conducted two lexical retrieval experiments to investigate whether clustering coefficients affect the accuracy and speed of word retrieval. In the first experiment, a perceptual identification task, participants were asked to identify words

that were played to them with added white noise distortion. The words were divided into two groups: one with a high clustering coefficient and one with a low clustering coefficient. To ensure the validity of the results, the words were controlled for various confounding factors, such as word frequency, subjective familiarity, and phonotactic probability. The results showed that the identification accuracy for words with a low clustering coefficient was 58% (sd = 8.4), while the accuracy for words with a high clustering coefficient was 72% (sd = 8.2). An ANOVA was performed on the accuracy rates between the two groups, revealing that the difference was statistically significant.

The other experiment conducted by Chan and Vitevitch was a lexical decision task, where participants were presented with words (without white noise) from two groups: real English words and non-words. The task was to correctly identify which items were real words and which were non-words. This experiment used the same set of real words as the previous experiment, along with additional non-words. The main variable of interest was the reaction time with which participants correctly identified the words. The results showed that participants responded more slowly to words with a high clustering coefficient (mean = 900 ms, sd = 86.6) compared to words with a low clustering coefficient (mean = 888 ms, sd = 82.1). Although the difference was small, it was still statistically significant.

The overall results from both experiments suggest that the clustering coefficient affects not only the accuracy of word retrieval but also its speed, reflecting overall word processing. These findings are open to interpretation as they may seem surprising or counterintuitive. On one hand, high-clustering words were identified more accurately in the first experiment, which might suggest that they are easier to process. However, in the second experiment, participants responded to high-clustering words more slowly, which appears to contradict the first result. My interpretation is that a high clustering coefficient might enhance accuracy under noisy conditions by leveraging the strong interconnections among neighbours, but it might slow down retrieval in tasks requiring quick, unambiguous decision-making due to increased competition among similar word forms. Conversely, a low clustering coefficient appears to support faster and more efficient retrieval by reducing this competition, particularly in situations where speed is crucial.

Another study specifically focused on the influence of clustering coefficient on word processing managed to replicate the finding that words with lower clustering coefficients are processed more easily (Vitevitch et al., 2011). Vitevitch et al. (2011) conducted a computational simulation of spreading activation within a phonological network, showing that word nodes with lower clustering coefficients were more activated than those with higher

clustering coefficients. Greater activation in word processing models suggests faster retrieval from mental lexicon supporting the interpretation mentioned above. Overall, the studies show that listeners are sensitive to clustering coefficients of words, and that the role clustering coefficient and other network metrics in word processing is multifaceted.

Recent studies have increasingly focused on how phonological networks can be applied to the study of language acquisition and impairments. For instance, Siew and Vitevitch (2020) explored network growth principles, offering insights into how language networks evolve and adapt over time. Their work highlights the dynamic nature of phonological networks, demonstrating how connections between words develop as individuals acquire new vocabulary and how these networks might reorganize in response to learning or cognitive changes. Building on this, Vitevitch et al. (2023) examined the resilience of phonological networks, illustrating how these networks maintain language functionality despite potential disruptions, such as damage to the brain or cognitive decline. Their findings emphasize the robustness of phonological networks, suggesting that even when some connections are lost or weakened, the overall structure remains functional, which has important implications for understanding language recovery following impairments like aphasia.

Furthermore, other recent approaches have suggested the potential of using language networks as supplementary diagnostic tools for language and other psychological impairments. Kennett and Faust (2019), in their contribution to Vitevitch's edited volume, propose that analysing the structure and connectivity of language networks could help identify early signs of cognitive decline or other neurological issues. These approaches represent a promising intersection between theoretical research and clinical application, indicating that network science could play a role in both understanding and diagnosing language-related disorders. Collectively, these studies underscore how network science provides a robust framework for modelling language and its cognitive underpinnings. By applying network principles to the study of language, researchers can gain deeper insights into how language is acquired, maintained, and sometimes impaired, offering new avenues for both theoretical research and practical applications in clinical settings.

Up to this point, I have primarily focused on phonological networks, as this work builds on their application in the study of word processing. Previous research has provided substantial evidence that individuals are sensitive to various network metrics, such as degree, clustering coefficient, and potentially others, during word processing. Notably, the role of clustering coefficient in phonological networks appears to be complex and multifaceted. The

next chapter shifts the focus to semantic networks, which are the central topic of this work. It will provide a similar overview of how semantic networks have been constructed and explore whether the network metrics that have proven relevant in phonological networks—such as clustering coefficient and degree—are also significant in the context of word processing within semantic networks.

2.4. Semantic networks

In phonological networks, the edges between nodes represent phonological similarities between words. However, edges can also represent other types of relationships, such as semantic ones. Semantic relationships between words or phrases have been quantified in different ways in language network science. Engelthaler and Hills (2019) distinguish four general approaches to quantifying these semantic relationships. I have adopted and presented the approaches from Engelthaler and Hills (seen in Vitevitch (2019: 167)) in table 3.

Table 3 An overview of different edge types in semantic networks, adopted from Engelthaler and Hills in Network Science in Cognitive Psychology, Vitevitch (2019), p. 167.

Edge type basis	Description	Reference
Perceptual and functional features	Edges are based on shared features.	(McRae et al., 2005, Preininger, Brand & Kříž, 2022, Vinson & Vigliocco, 2008)
Free associations	Edges are based on cue-target relationships in the free association task.	(Nelson, McEvoy, & Schreiber 2004, Lakhzoum et al. (2021))
Semantic and conceptual categorization	Edges are based on theoretically derived semantic and conceptual categorization of words.	Miller, G. A. (1995), Jarmasz, M., & Szpakowicz, S. (2003)
Natural language corpora	Edges are based on word co-occurrence.	(Church, K. W., & Hanks, P. 1990, Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. 2013, Pennington, J., Socher, R., & Manning, C. D. 2014)

One approach to constructing semantic networks involves creating edges based on the perceptual or functional features of words. In these networks, nodes typically represent individual lexemes, and the edges reflect the semantic proximity between these lexemes based on shared perceptual attributes (e.g., “is large,” “has fur”) or functional roles (e.g., “used in cooking,” “is a vehicle”). This approach is informed by semantic feature norms, such as those developed by McRae et al. (2005), Vinson & Vigliocco (2008), and more recently by Preininger, Brand, & Kříž (2022). In these studies, adult participants generate or evaluate lists of features along specific semantic dimensions for a given word.

For example, “dog” and “cat” might be connected within the network due to their commonalities in perceptual features (e.g., “having four legs”) or functional attributes, or through an overall statistic that computes their similarity across a set of these features and attributes. This methodological approach allows for the creation of a rich semantic landscape where the strength of associations between words is not merely based on direct lexical co-occurrence or syntactic proximity, but rather reflects conceptual schemas that inform human categorization and cognitive processing.

Unlike edges based on free associations or natural language corpora, the semantic dimensions used to define edges based on perceptual and functional features are clearly defined and grounded in a broader theoretical framework, which makes them easier to interpret. For instance, in the work of Preininger, Brand, and Kříž (2022) on quantifying socio-semantic features for Czech, participants evaluated words along dimensions such as gender, location, politics, valence, and age. Participants might be presented with a word like “dog” or “beard” and asked to rate it on a Likert scale for attributes like “urban”, “feminine”, or “positive”. The edges between words are then computed by calculating the similarity between pairs of words based on their overall scores within these dimensions.

The second approach to constructing semantic networks is basing the edges between words on the results from the free association task, which capture the spontaneous and often unpredictable links between words that can be elicited from people. In these networks, nodes represent individual words or concepts, while edges are drawn based on the immediate, unmediated responses elicited by cue words in participants. This approach is grounded in the methodology of free association norms, where participants are prompted with a cue word (e.g., “angry”) and respond with the first word that comes to mind (e.g., “furious”, “red”), as seen in studies by Nelson, McEvoy, & Schreiber (2004).

Despite the seemingly random and spontaneous nature of participants’ responses, free association tasks have proven to be a reliable experimental method with consistent results. A

key metric derived from these responses is the free association response probability for pairs of words, which indicates the likelihood that presenting one word will trigger the response of another. The strength and directionality of these connections reflect the associative depth between the words, revealing the context-free priming paths within the mental lexicon. However, it is important to note that free associations measure relative associative strength between words rather than absolute strength. For example, knowing that the cue word “book” primes the response “read” from 43% of participants tells us that this response is more strongly associated than “study,” which was produced by 5.5% of participants (Nelson, McEvoy, & Schreiber, 2004: 406). However, it does not provide absolute insight into their overall association within the mental lexicon. This data can be used to create a network with directed edges to reflect the direction of priming.

The study by Lakhzoum et al. (2021) applied semantic network analyses to explore norms of French word associations for concrete and abstract concepts. Through network analyses and metrics, they found that both concrete and abstract networks exhibited characteristics of a small-world structure, such as high clustering coefficients, sparse density, and small average shortest path lengths. Moreover, they observed differences in overall structural organization between abstract and concrete concepts, with concrete concepts showing denser connectivity. This work highlights the utility of semantic network analyses and visualization in understanding the organization of word associations in the mental lexicon. Interestingly, in figure 3, a node with the highest degree was the word for love *amour* and one with one of the fewest was the word for boredom *ennui* (Lakhzoum et al. 2021: 10).

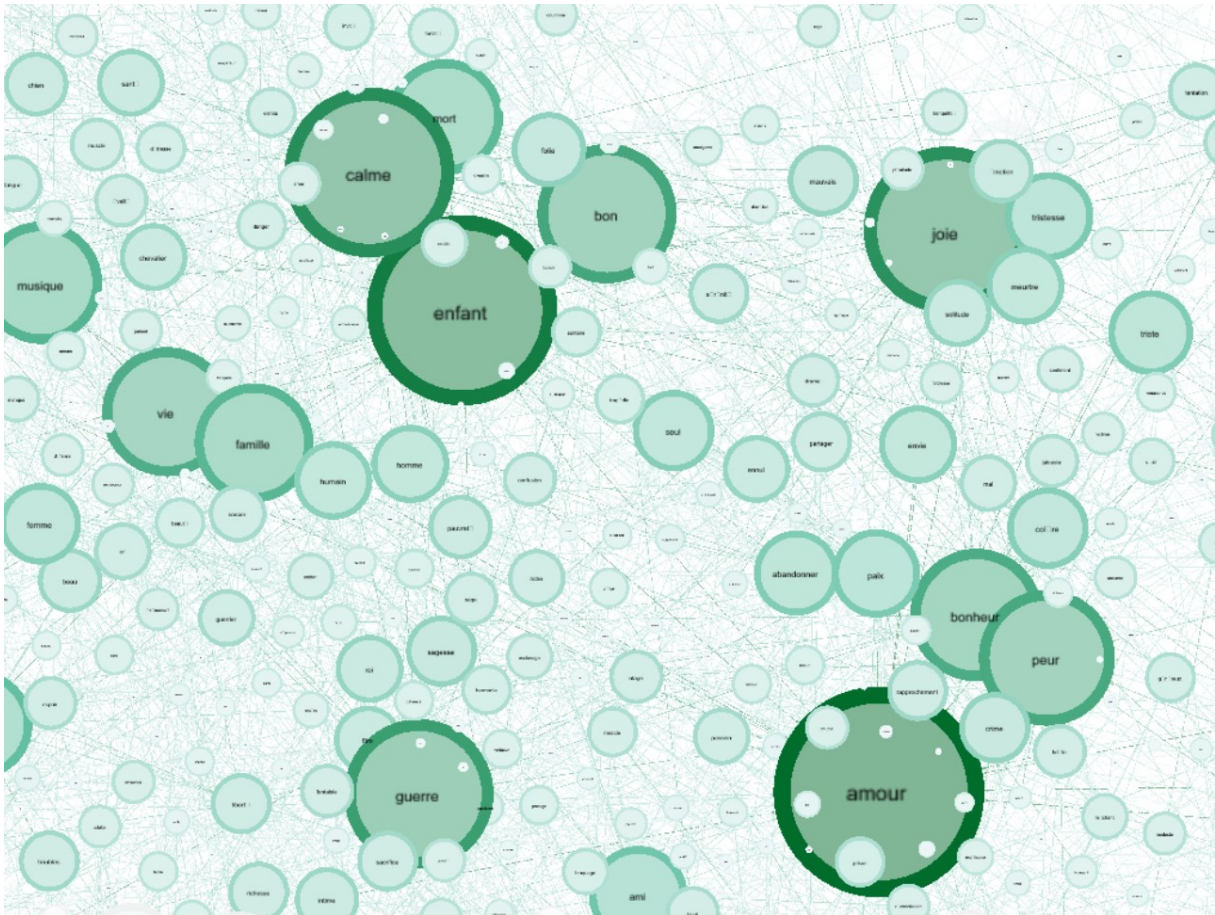


Figure 3 Visualisation of a portion of the network made by Lakhzoum et al. (2021) where the node size and colour intensity reflect its degree

Another approach to the construction of semantic networks Engelthaler and Hills labelled as Semantic-Conceptual Networks, where edges are based on conceptual and semantic relationships between words. In such networks, nodes still represent individual lexemes; however, the links between these nodes are established based on conceptual similarities, synonymy, antonymy, and hierarchical super-subordinate categorizations. These relationships are typically drawn from structured lexical databases like WordNet or Roget's Thesaurus. The information and structure of these meticulously constructed databases are informed by various linguistic theories and analyses. The types of relationships represented in these databases depend on the specific database and the kind of linguistic analysis and terminology it employs. Generally, researchers constructing networks from these databases compute a measure of semantic similarity between pairs of words. For example, in a network derived from WordNet, words grouped into synsets share a common concept, thereby establishing a connection between them based on conceptual similarity. The edges in such a network, therefore, reflect the similarity between words and phrases along these conceptual

Finally, semantic networks predicated on similar patterns of usage in natural language embody an empirical model for delineating semantic relationships, anchoring the connections between nodes not on explicit participant responses, perceptual similarities, or theoretically driven linguistic analysis but on the contextual co-occurrence and usage patterns observed within vast corpora of text. This makes the approach primarily data-driven constructing the semantic relationships between words in a bottom-up fashion. In this framework, nodes again signify individual words or concepts, while links are established based on the proximity, frequency, but usually other more intricate statistical means that operationalize how closely different words appear within natural language texts, capturing the implicit semantic connections inferred from linguistic context. This approach leverages the power of either word co-occurrence measures, semantic space models or the combination of both, which avoid the need for words to be directly associated or share explicit features, focusing instead on their emergent semantic similarity through common patterns of use across diverse linguistic contexts.

Semantic space models are a well-established technique in computational linguistics and natural language processing (NLP). These models use algorithms to transform words or phrases from text into vectors within a high-dimensional vector space, where the semantic similarity between words is represented by the spatial relationships between their corresponding vectors. The basic idea behind this approach is that words appearing in similar contexts tend to have similar meanings. By analysing large corpora of text, NLP researchers use algorithms to generate vector representations that capture semantic properties of words, typically interpreted as semantic similarity and associations (Lund & Burgess, 1996; Gastaldi, 2021). To create an edge list for a semantic network, researchers can calculate the Euclidean distance or cosine similarity between the vectors. Both measures reflect the closeness between two vectors, which can be interpreted as semantic similarity. Semantic space models and their applications are further elaborated in chapter 2.5.

Although there is some overlap between representing and analysing meaning through word co-occurrence measures and semantic space models, they differ in how word meaning is modelled, and the computational techniques and algorithms used to operationalize semantic relationships. Word co-occurrence measures focus on the frequency with which words appear together within a specified context or window in a text corpus. These measures are in accord with the premise formulated by Firth (1957) that words that frequently co-occur in similar contexts tend to have related meanings. Co-occurrence can be represented in matrices where rows and columns represent words, and each cell contains the frequency or a derived statistic,

e.g., Pointwise Mutual Information (PMI), indicating how often two words co-occur. These high-dimensional matrices reflect the directly observed co-occurrences within a relatively narrow context window, usually just a few words. However, this method may miss some semantic or pragmatic relationships that manifest over longer text spans, such as sentences, paragraphs, or larger discourse units. Despite these limitations, Church and Hanks (1990) proposed the creation of a measure *association ratio* that estimates word association norms directly from language corpora using information theoretic PMI and other statistic approaches. The motivation was to avoid the then prevailing method of deriving those norms from participants in psycholinguistic experiments which was laborious, costly and unreliable. It helped to pave the way for the development of further corpus linguistic methods and quantitative computational processing of language for linguistic research.

These developments eventually led to the creation of more complex semantic space models, such as Latent Semantic Analysis (LSA), Word2Vec, GloVe, and more recently, BERT. These models utilize machine learning to represent word meanings in continuous vector spaces, among other capabilities. While these models often begin with word co-occurrence data, they go beyond simple co-occurrence by applying sophisticated algorithms to compute multi-dimensional word vector representations, known as embeddings, for each word. These embeddings capture more context-sensitive semantic information, allowing the models to represent not just immediate co-occurrence relationships, but also implicit semantic connections that span broader contexts. For example, words that are used in similar contexts will have vectors that are close together in the semantic space, even if they do not frequently co-occur. The underlying basis for these embeddings can indeed start with co-occurrence data, but the transformation into a semantic space typically involves advanced computational techniques including the recent neural network-based models. Neural network-based models, such as Word2Vec, GloVe, and BERT, abstract away from raw co-occurrences to capture more nuanced semantic similarities and relationships. In Word2Vec, the training process involves predicting a word from its context (the Continuous Bag of Words model) or predicting the context from a word (the Skip-gram model). This process effectively learns vector representations that capture a wide array of semantic and syntactic relationships (Mikolov et al., 2013). GloVe, on the other hand, starts with the co-occurrence matrix but aims to learn vector representations by modelling the ratios of co-occurrence probabilities, which are believed to encode important semantic information (Pennington, Socher, & Manning, 2014). BERT (Bidirectional Encoder Representations from Transformers) represents a more recent and sophisticated approach, using deep learning techniques based on transformers. BERT

differs from previous models by considering the context of a word bidirectionally, meaning it takes into account both the preceding and following context when learning word embeddings. This results in a richer and more accurate representation of word meanings in various contexts, significantly improving performance on a wide range of NLP tasks (Gastaldi 2021).

Constructing a network from word co-occurrence data or semantic space models based on embeddings involves selecting and calculating a value for all pairs of words from the corpus that reflects their semantic similarity based on usage patterns in the text. Word nodes are then connected by edges, with the weight of each edge corresponding to this similarity value, thus forming a semantic language network. This methodology, which leverages large-scale textual data to represent semantics based on authentic, unmediated language output, provides a robust model for understanding meaning in language. It reflects how human cognition might be sensitive to usage patterns that are not immediately apparent.

It seems that research in semantic networks does not place as much emphasis on the role of specific network metrics, such as clustering coefficient, in word processing as Vitevitch's work (2009, 2011) did in the study of phonological networks. This work; therefore, aims to contribute to this research gap in semantic networks. Most examples of the semantic networks predicated on similar patterns of usage in natural language do not explicitly engage with the network science tools or visualize networks. These approaches mainly focus on computing matrices of co-occurrence values or, in the case of semantic spaces, compute and plot word vectors. Both can compute some measure representing similarity between two words such as Church and Hanks' (1990) *association ratio*, but measures of this kind have generally not been used to construct a proper semantic network.

Large language models like Word2Vec or BERT, which plot word vectors based on similarity, are not formal networks in the strict sense used in network science. While they are powerful tools developed for commercial applications like machine translation or chatbot assistants, a proper network requires a defined set of nodes and their explicit connections, which these models do not inherently provide. However, we can transform embeddings into a network by calculating the distance between every pair of embeddings in the semantic space and defining a link between them based on a cutoff value for that distance. Given the widespread success and unprecedented language capabilities of these large language models, this work aims to leverage the way they represent language for the purposes of language network research. Specifically, the goal is to construct a semantic network based on the BERT language model and then apply the tools of network science to identify meaningful correlations between network measurements from this constructed network and results from

lexical retrieval experiments obtained from The Massive Auditory Lexical Decision (MALD) database. This research is relevant for two main reasons. First, it builds on previous research in phonological networks by attempting to replicate the significant correlations found between certain network measures (e.g., clustering coefficient) and word processing within semantic networks. Second, depending on the results, it could either support or challenge the relevance of current large language models for linguistic and cognitive science.

The following chapter introduces word vectors, also known as embeddings, which are how large language models represent word meanings. Understanding these large language models and their word vectors is crucial for interpreting the types of relationships encoded between the word nodes in the semantic network constructed in this work.

2.5. Large Language Models and Word Vectors

The rapid development of machine learning techniques in the 2010s and 2020s has introduced new methods for quantitatively representing meaning and language using vector representations of words, embeddings, based on broad patterns in which the words appear in text. One of the foundational algorithms in this field is Word2Vec, introduced by Mikolov et al. (2013a). Word2Vec transforms words into high-dimensional vectors, assigning each word a set of numerical values that capture its meaning based on its contextual usage in large corpora of text.

Word2Vec operates by representing words as numerical vectors in a multi-dimensional space, where each dimension corresponds to a certain aspect of the word's meaning as abstracted from the text (Mikolov et al., 2013a). Multidimensional word vectors are arrays of numbers assigned to words, representing their positions in a high-dimensional semantic space. A common way to visualize these vectors is by applying dimension-reducing statistical methods such as t-distributed stochastic neighbor embedding (t-SNE) or principal component analysis (PCA), which allow for the plotting of these vectors in a 2D space.² These vectors are learned through neural network models trained on vast amounts of text data. The central idea behind Word2Vec is the distributional hypothesis, proposed by Harris (1954), which posits that words appearing in similar contexts tend to have similar meanings. As a result, words with patterns of use in text have vectors that are close to each other in the vector space.

² An on-line tool for visualizing 2D and 3D word vectors can be found at <https://projector.tensorflow.org>.

Word2Vec has several practical applications, such as computing semantic similarity between pairs of words, understanding word analogies (often referred to as ‘word mathematics’), and text classification. For example, we can quantify the semantic similarity between words by measuring the cosine similarity or Euclidean distance between their word vectors (Mikolov et al., 2013b). Beyond analogies, these vectors are also used in machine translation, text classification (e.g., sorting texts into genres), and sentiment analysis, which evaluates whether a text is positive, negative, or neutral.

For linguistics, the development of word vectors represents an opportunity to explore semantic analysis in new, quantitative, empirical, and data-driven ways, complementing traditional theoretical frameworks. Models like Word2Vec and its successors provide researchers with tools to analyze the semantic nuances of language at a level of detail that was previously unattainable. By examining word vectors and their relationships, linguists can uncover subtle shades of meaning, polysemy, and semantic shifts over time. Moreover, by analysing historical texts and corpora, researchers can track the evolution of language, identifying semantic changes, lexical innovations, and shifts in usage patterns, offering insights into the dynamics of language change. These computational models also offer valuable insights into how language might be processed and represented in the human mind. By correlating the representations of language in these models with human data, researchers can test theories of language processing and the organization of the mental lexicon.

Since the development of Word2Vec in 2013, more advanced language models have emerged. This work specifically utilizes BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art natural language processing model introduced by Devlin et al. (2018). Unlike Word2Vec, BERT’s word vectors are highly sensitive to context. BERT employs a deep learning architecture known as Transformers, which captures long-range dependencies and contextual relationships within text.

One of the key distinctions between BERT and Word2Vec lies in their approach to contextual understanding. While Word2Vec generates fixed-length vectors for individual words based on a linearly parsed context, BERT processes text bidirectionally, taking into account the entire context of a sentence by analysing both the words that precede and follow a given word. This bidirectional processing allows BERT to capture more nuanced meanings that could be overlooked by models that analyse text in a linear fashion, either from left to right or vice versa (Gastaldi, 2021).

Despite the advancements of models like BERT, there are still limitations to their current capabilities, and they should not be assumed to provide complete and accurate

that they could take a word vector for “man”, subtract it from the vector for “king”, add the vector for “woman” and the resulting closest vector in the vector space would be “queen”.

$$1) \text{ Man} - \text{King} + \text{Woman} = \text{Queen}$$

Geometrically speaking, this means that the same distance and direction between pairs of word vectors can be interpreted as similar kind of semantic relationship, e.g. gender in example 1. This logic is illustrated in figure 6 where we see various pairs of words distinguished by their grammatical gender. The word vector pairs are plotted so that relatively similar distance and direction is between them. In this sentence, the distance in the semantic space is able to encode semantic similarity in general and other linguistically relevant distinctions in particular as well.

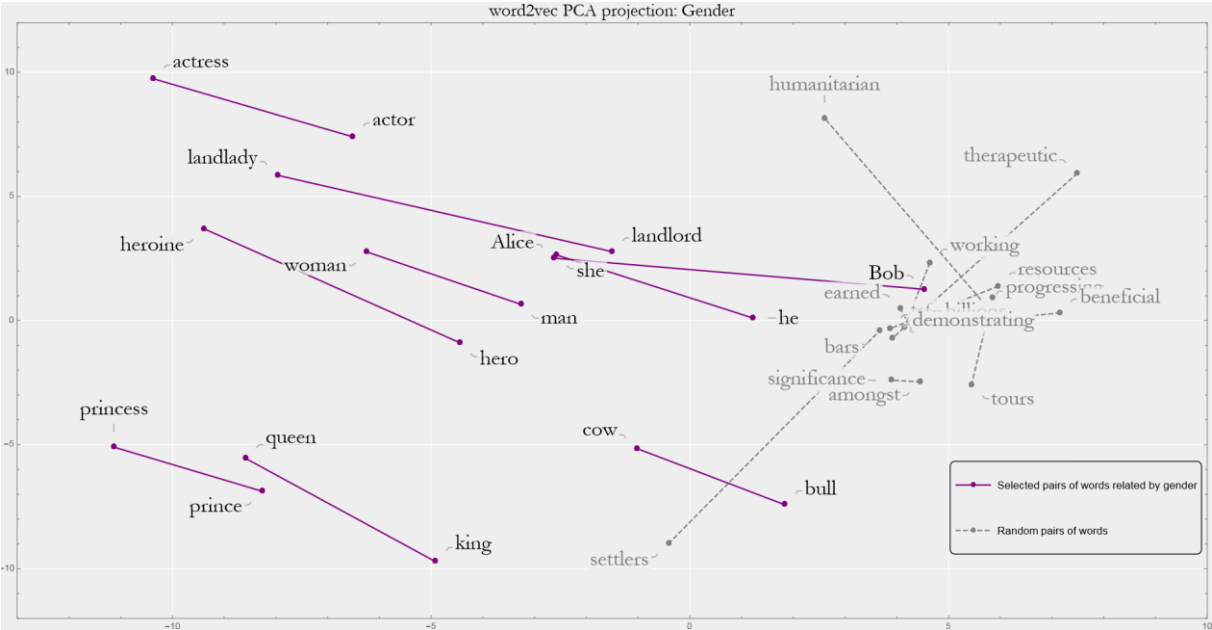


Figure 6 Offset representing the gender relation, revealed in the embedding space by a PCA projection. Credit: Gastaldi (2021)

Different kinds of mainly semantic analogies have been discovered within semantic spaces made with Word2Vec but also syntactic patterns such as verb tenses or adjectival comparatives plotted in figure 7 and 8. Figure 7 shows how the Word2Vec identifies the pattern between adjectives and their comparative forms, grouping words like “strong”, “stronger” and “strongest” together in the vector space. Figure 8 illustrates how Word2Vec recognizes relationships between different forms of irregular verbs, such as “give”, “gave” and “given”. The consistent patterns in these selected triads demonstrates the ways in which special representation can be used to encode linguistic relationships and how distance in space can serve as a basis for determining links between pairs of words in a network.

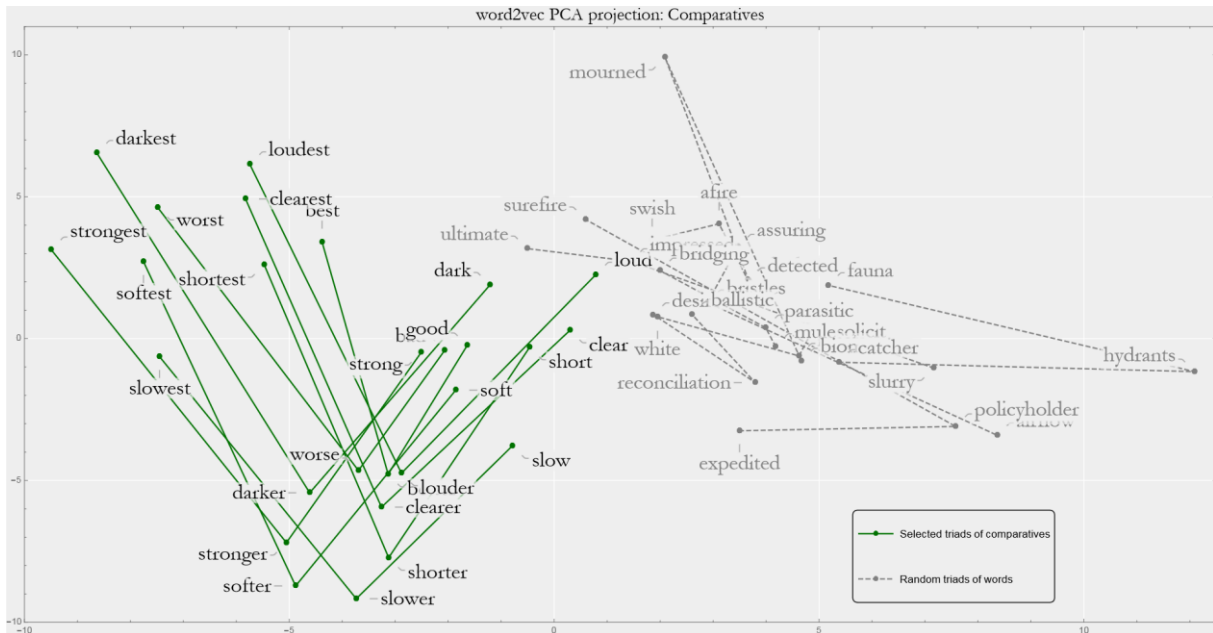


Figure 7 Pattern in the embedding space (word2vec) corresponding to the comparative category (base, comparative and superlative forms). Credit: Gastaldi (2021)

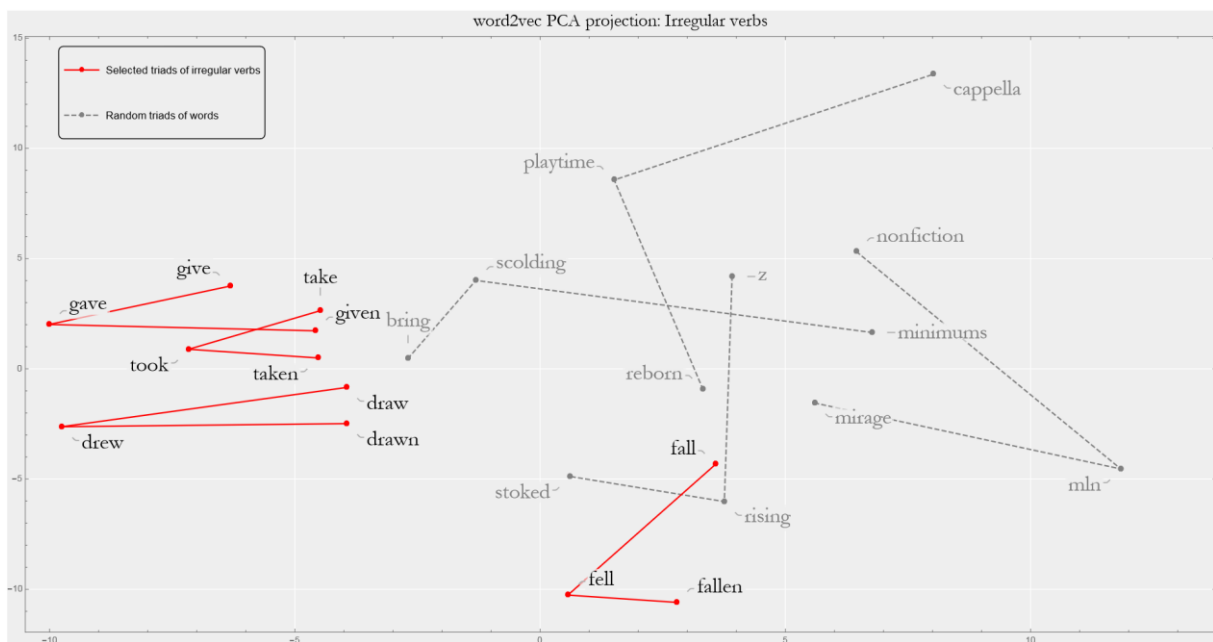


Figure 8 Pattern in the embedding space (word2vec) corresponding to conjugation of irregular verbs. Credit: Gastaldi (2021)

Finally, word vectors are also able to track diachronic language change when computed on texts from different times. Figure 9 shows a plot from Hamilton et al. (2018) in which the development of the words “gay”, “broadcast”, and “awful” was studied. For instance, “gay” transitioned from meaning “cheerful” in the 1900s to being associated with sexual orientation by the 1990s. Similarly, “broadcast” shifted from an agricultural term to one related to media, and “awful” changed from meaning “awe-inspiring” to “terrible”. This demonstrates the capability of word vectors to capture and represent changes in word meanings over time.

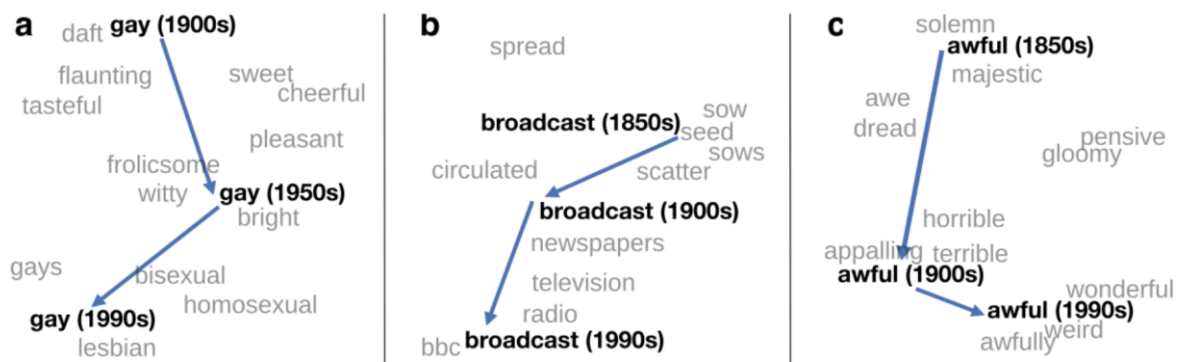


Figure 9 Visualization of semantic change based on word2vec. Credit: Hamilton et al. (2016).

These examples provide evidence that word vectors can capture various kinds of semantic relationships from raw text. Figures 5 – 9 present visualizations of word vectors using statistical techniques that reduce the dimensions of the original vectors so that they could be plotted which also leads to loss of some of the information contained in them. Chapter 3 describes the material and method of the current work that was used to train BERT to acquire the word vectors which were finally transformed into a network and analysed in combination with psycholinguistic data from word processing. The construction of a network prevents the reduction of dimensionality of the vectors, conserving as much of the semantic information as possible.

3. Material and Method

This work integrates machine learning techniques to compute word vectors and transform them into a semantic network of English with network science to analyse the network in the light of psycholinguistic research of word processing. There are two general research questions:

1. Are word vectors a relevant source for a cognitively insightful semantic network?
2. If so, are there any significant correlations between the structure of such network and word processing efficiency?

Therefore, one goal of this work is to test the relevance of creating semantic network based on word vectors for psycholinguistic and cognitive research. Meaningful semantic

network would have small-world characteristics found in other social and cognitive phenomena modelled by networks. Another goal would be to find significant correlations between measures of such network and results of word processing experiments for individual words. We can ‘measure’ or describe the structure of a semantic network by computing the network measures for all the words contained in the network. We can also measure the efficiency of word processing by measuring the reaction times of words in lexical decision task, for example. By correlating the network measures and reaction times for the same words, we can investigate whether there is a relationship between the two. If so, this would answer the second research question and suggest that the network is able to capture a cognitively meaningful representation of semantics. There are two sources of data for this work – a text sample from the TV Corpus that served as the training data for BERT and reaction times for the words contained in the resulting semantic network from The Massive Auditory Lexical Decision (MALD) database. Both sources of data and their analysis is described in detail in the following sections. As an overview of the practical part of this work, main steps are presented below:

1. Training BERT on a sample from The TV Corpus
2. Outputting word vectors from BERT
3. Creating edge list by computing cosine similarity between every pair of word vectors
4. Visualizing the network with GEPHI
5. Obtaining reaction times from lexical retrieval experiments in MALD database
6. Correlating reaction times with network measurements

3.1. Computation of the Semantic Network

I trained the open-source language model BERT in the python programming language from which the model is accessible through the ‘transformers’ library. BERT can be trained on tokenized text data to output word vectors. I trained it on a sample from The TV Corpus which is freely available for downloading in its raw-text form at english-corpora.org/tv/. The TV Corpus contains subtitles of tv shows from the 1950’s to the present time. I chose this corpus for its greater proximity to authentic spoken language which I wanted to model with the resulting network. I downloaded the raw text and tokenized it on sentences which is the necessary input form for BERT. The training of BERT itself is computationally demanding so I used the computational infrastructure of MetaCentrum VO, a catch-all virtual organization

of the Czech National Grid Organization, which operates and manages distributed computing infrastructure which is available to use for academic purposes at metavo.metacentrum.cz/.

The output of training BERT was a list of word vectors for each token. In total, there was 9683 tokens each of which was represented by 768 dimensional vectors. To ensure a meaningful representation of semantic similarity between word vectors, I excluded vectors containing non- alphabetic characters, thus eliminating punctuation, special symbols, and numbers. This reduced the number of word vectors to 6664. Table 4 contains a snapshot of the table with the filtered word vectors.

Table 4 A part of the table containing 6664 word vectors with 768 dimension computed by BERT

	0	0.1	1	2	3	
1	i	0.4131588	-0.5164368	-0.070106514	0.24670188	0.1
2	used	0.6167732	-0.935629	0.5481038	0.20566644	1.
3	to	0.33126172	-0.43050405	0.4024777	-0.4328461	0.7
4	date	0.6792053	-0.105997495	0.5854818	-0.86545646	0.2
5	a	-0.25574097	-0.69546574	0.51769054	-0.17579953	0.7
6	girl	0.3497564	-1.0572797	0.2570811	-0.24540454	0.0
7	who	-0.43124136	-0.5059501	0.3511485	-0.17387336	0.3
8	was	-0.036158808	-1.056186	0.3266946	-0.4658766	0.3
9	the	0.22171763	-1.2748561	0.6104466	-0.22556755	0.2
10	captain	0.6563785	-0.41294754	0.09312493	-0.5945709	-0.2
11	of	-0.8298847	-0.3123703	0.19391257	-0.1716561	-0.02
12	the	-0.20958453	-0.3160304	-0.034724604	-0.3607289	-0.02
13	netball	0.5075988	0.30638695	0.93784165	-0.5134297	-0.1
14	team	-0.2594932	-0.09756132	0.19902855	-0.76846117	-0.5
15	is	0.38111046	-0.4069852	0.18678449	-0.48942044	0.6
16	he	0.11544245	-0.497555	0.71277905	0.4922282	0.
17	okay	0.0743646	-0.3843242	0.7288899	-0.2971244	1.
18	but	0.3264922	0.12335851	-0.17650524	0.09791387	-0.

The first column contains word labels, the rest are values of the multidimensional vectors with 768 dimensions. The rows are the individual words.

Next, I calculated cosine similarity between every pair of word vectors which reflects how similar the vectors are based on the angle between them and the origin point of the vector space. The cosine similarity value ranges from -1 to 1. 1 indicates that the two vectors are identical in orientation. In the context of word vectors, this would mean that the two tokens are semantically very similar or nearly identical. 0 implies orthogonality of vectors, suggesting no semantic similarity between the two tokens. -1 is less common in word vectors but theoretically would indicate completely opposite meanings (Sidorov et al., 2014). However, interpreting these values can be somewhat opaque and context-dependent as there are no universally agreed-upon thresholds. Generally, higher values (closer to 1) indicate greater

similarity. Often-used values for considering two vectors similar include 0.5 and 0.7 (Zhou et al., 2021). Figure 10 displays a histogram of the distribution of the cosine similarity values calculated between word vectors computed by BERT for the sample from the TV Corpus.

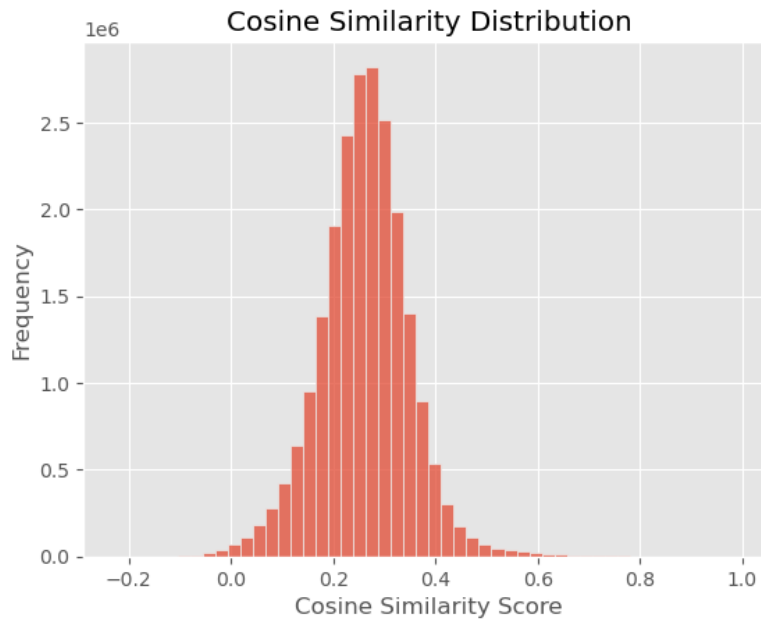


Figure 10 Cosine similarity distribution

The distribution follows a normal pattern, with most cosine similarity values clustering around 0.3. Since values above 0.7 are rare, a threshold of 0.5 is used to establish connections between nodes in the network. In the resulting semantic network, each node represents a token, and a weighted link is created between any pair of nodes where the cosine similarity exceeds 0.5. The weight of each link corresponds to the cosine similarity value between the respective word vectors. This process generates the final edge list for the network construction, which is further refined by retaining only nouns (excluding proper names) and verbs (excluding abbreviated verb forms, like ‘re’) to ensure a semantic network that can be interpreted as conveniently as possible.

3.2. Word Processing Data

The final step was to include the data from lexical decision task from The Massive Auditory Lexical Decision (MALD) database. MALD database is a freely available auditory and production data set for speech and psycholinguistic research that contains time-aligned stimulus recordings for 26,793 words and 9592 pseudowords, and response data for 227,179 auditory lexical decisions from 231 unique monolingual English listeners. It is a valuable source of reaction times for different words from the auditory lexical decision task. The word reaction times from MALD database were extracted to match the words in the semantic

network. Reaction time values under 200ms and over 4000ms were excluded to avoid erroneous data and outliers. The rest of the reaction time values were averaged for each unique word (as a single unique word may have multiple reaction times assigned to it from multiple experimental trials). The final step of this study is to do a regression analysis between the word reaction times and network measures of the words to test if there are any statistically significant correlations. In the end, three basic network measures were chosen for the analysis: closeness centrality, clustering coefficient, and degree centrality. While the whole network contained 869 unique words, 238 were not present in the MALD database so the regression analysis was conducted on 631 unique words for which there is both the mean reaction time and the three network measures. Appendix contains a full list of those words with their values of average reaction time from MALD and their values of degree centrality, closeness centrality and clustering coefficient. This table was the bases for the regression analysis. Next section presents the results.

4. Research

Figure 11 shows a visualization of the semantic network produced in the network visualization software GEPHI. It contains a total of 869 nodes (i.e. unique words) connected by 3196 edges. Edge thickness reflects its weight – the thicker, the higher the cosine similarity signaling semantic similarity. Furthermore, node size reflects its degree, therefore; the bigger the node, the higher its degree is, meaning the number of links. The layout of nodes does not reflect anything. Visual inspection of the semantic network reveals small-world characteristics with clusters of high-degree hub nodes and more scarcely connected low-degree nodes. Another small-world characteristic is that despite the fact that the network contains almost a thousand of nodes, its diameter, i.e. the longest path, is 11 which means that the semantic network is relatively compact. Table 5 displays basic macroscopic measures of the network.

Box Plots of Mean Word Reaction Time and Network Measures

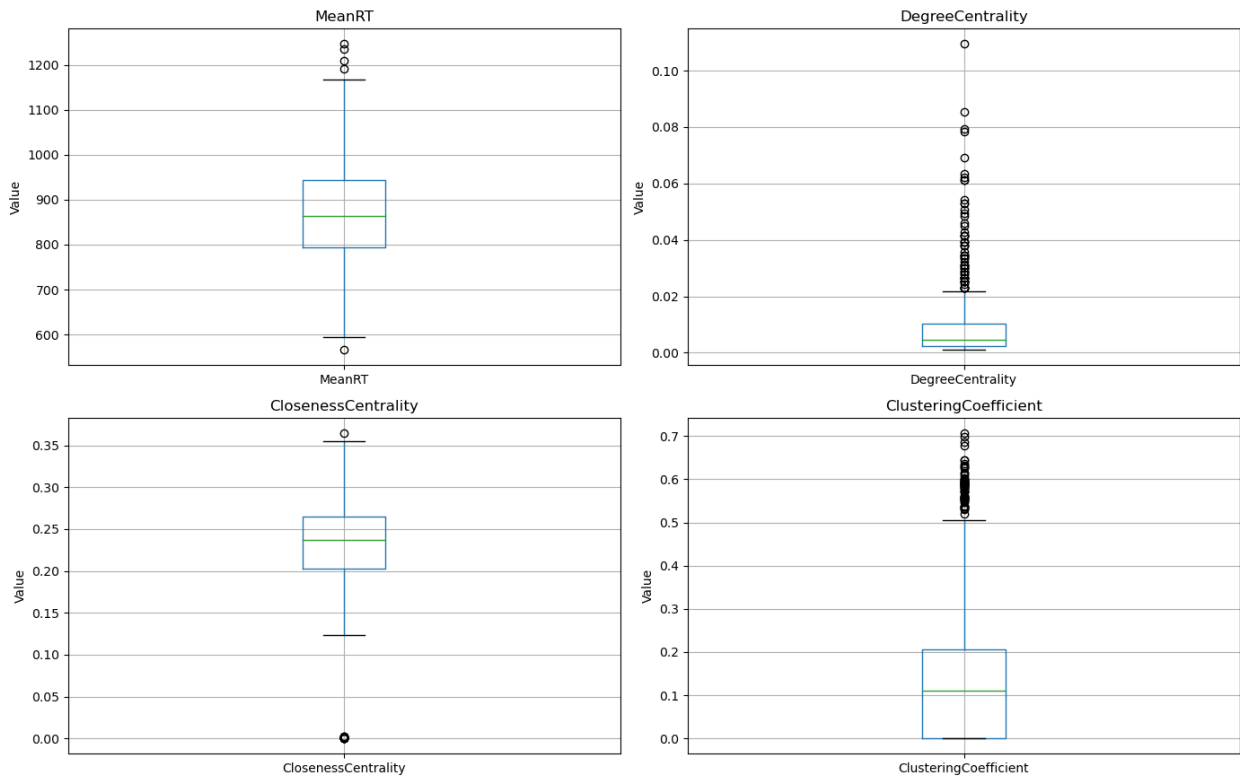


Figure 12 Box plots of mean word reaction times and network measures

A multiple linear regression model was computed to analyse how the three network measures relate to the mean reaction time (MeanRT), with all variables being transformed using the natural logarithm. The model predicts the natural logarithm of MeanRT as a function of the natural logarithms of *degree centrality + 1*, *closeness centrality + 1*, and *clustering coefficient + 1*. The “+1” in the formula ensures that there are no issues with taking logarithms of zero. The model suggests there may be a negative relationship between degree centrality and MeanRT, meaning that nodes with higher degree centrality scores tend to have lower reaction times, possibly indicating more efficiency or priority in word processing, but it is not statistically significant. The overall fit of the model is quite weak, as indicated by the low R-squared value, meaning that other variables not included in the model might be influencing MeanRT. Table 6 shows summary of the regression model.

```
Call:
lm(formula = log(MeanRT) ~ log(DegreeCentrality + 1) + log(ClosenessCentrality +
  1) + log(ClusteringCoefficient + 1), data = stats_reg)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-0.42548 -0.08355 -0.00469  0.08331  0.36327
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.75980   0.02222  304.214 <2e-16 ***
log(DegreeCentrality + 1) -0.96550   0.51105  -1.889  0.0593 .
log(ClosenessCentrality + 1)  0.02516   0.11829   0.213  0.8316
log(ClusteringCoefficient + 1)  0.06403   0.04050   1.581  0.1144
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.1251 on 624 degrees of freedom
Multiple R-squared:  0.01156, Adjusted R-squared:  0.006803
F-statistic: 2.432 on 3 and 624 DF, p-value: 0.0641
```

Table 6 Summary of the regression model

Added-variable plots are used to show the relationship between a given independent variable (each network measure) and the dependent variable (average reaction time), while accounting for the presence of other independent variables in the model. From these plots in figure 13, we can conclude that degree centrality has a slight negative impact on MeanRT, even after accounting for the other variables in the model. Meanwhile, closeness centrality and clustering coefficient do not seem to have a strong independent effect.

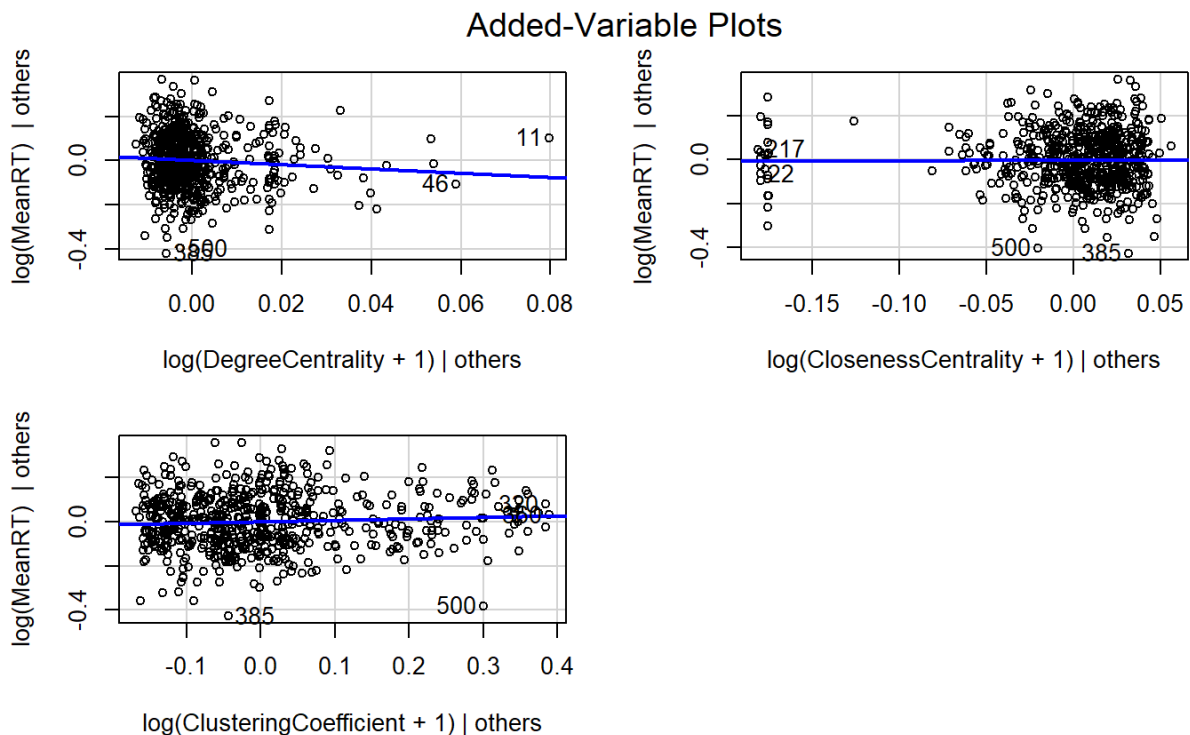


Figure 13 Added-variable plots

5. Conclusion

The results provide cautious optimism for using word vectors combined with network science to build semantic networks for psycholinguistic research. The semantic network created using BERT word vectors exhibited a small-world structure, similar to other social and cognitive phenomena modelled with networks. This finding, related to the first research question, suggests that word vectors can capture cognitively relevant information when transformed into a network. Regarding the second research question, degree centrality was found to influence reaction times in a lexical retrieval experiment from the MALD database, with results approaching statistical significance. Specifically, the data suggest that as the number of word's neighbours increases, its reaction time decreases. It could be understood that more neighbours tend to speed up word processing, suggesting that mental lexicon leverages semantic proximity in a sense that it is able to recognize words faster when they are semantically similar to many others. However, clustering coefficient and closeness centrality did not show a clear relationship with reaction time. Overall, this means that there is not enough evidence in the present work to conclusively answer the second research question. Although the overall model fit was weak, future research could focus on improving it by exploring additional network measurements of which there are many or other control variables such as word frequency. Another potential direction for studying semantic networks based on word vectors could involve using different training data sources or alternative language models. While this study used a sample from the TV Corpus reflecting spoken language, other sources like spoken language corpora or experimentally elicited narration from individual speakers should be considered. Future research might also explore more advanced models beyond BERT.

Research in both phonological (Kennet & Faust 2019) and semantic networks (Colunga & Sims 2017, de Boer et al. 2018, Hadley et al. 2019) is increasingly focusing on practical applications, such as studying language acquisition and language impairments. By accurately capturing and analysing the structure and dynamics of semantic networks, this research could aid in developing supplementary tools for diagnosing cognitive impairments. These tools could help detect subtle, incremental changes in language that occur in the early stages of conditions like dementia or delayed language development. Therefore, I believe that semantic networks based on word vectors specifically and language network in general promise a new direction in linguistic research that can not only bring new insights about the structure and dynamics of language but also practical application in various domains in which

language as a cognitive faculty of the human mind is relevant.

6. References

- Aitchison, Jean. *Words in the Mind: An Introduction to the Mental Lexicon, 4th Edition*. 4th edition. Chichester, West Sussex; Malden, MA: Wiley-Blackwell, 2012.
- Albert, Reka, and Albert-Laszlo Barabasi. “Statistical Mechanics of Complex Networks.” *Reviews of Modern Physics* 74, no. 1 (January 30, 2002): 47–97. <https://doi.org/10.1103/RevModPhys.74.47>.
- Anderson, John R. *The Architecture of Cognition*. 0 ed. Psychology Press, 2013. <https://doi.org/10.4324/9781315799438>.
- Balota, David A., Melvin J. Yap, Keith A. Hutchison, Michael J. Cortese, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and Rebecca Treiman. “The English Lexicon Project.” *Behavior Research Methods* 39, no. 3 (August 1, 2007): 445–59. <https://doi.org/10.3758/BF03193014>.
- Barabasi, Albert-Laszlo, and Reka Albert. “Emergence of Scaling in Random Networks.” *Science* 286, no. 5439 (October 15, 1999): 509–12. <https://doi.org/10.1126/science.286.5439.509>.
- Barabási, Albert-László, and Márton Pósfai. *Network Science*. Cambridge: Cambridge university press, 2016.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. ‘On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜’. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. Virtual Event Canada: ACM, 2021. <https://doi.org/10.1145/3442188.3445922>.
- Boer, J. N. de, A. E. Voppel, M. J. H. Begemann, H. G. Schnack, F. Wijnen, and I. E. C. Sommer. ‘Clinical Use of Semantic Space Models in Psychiatry and Neurology: A Systematic Review and Meta-Analysis’. *Neuroscience & Biobehavioral Reviews* 93 (1 October 2018): 85–92. <https://doi.org/10.1016/j.neubiorev.2018.06.008>.
- Chan, Kit Ying, and Michael S. Vitevitch. ‘The Influence of the Phonological Neighborhood Clustering Coefficient on Spoken Word Recognition.’ *Journal of Experimental Psychology: Human Perception and Performance* 35, no. 6 (2009): 1934–49. <https://doi.org/10.1037/a0016902>.
- Church, Kenneth Ward, and Patrick Hanks. ‘Word Association Norms, Mutual Information, and Lexicography’. *Computational Linguistics* 16, no. 1 (1990): 22–29.
- Colunga, Eliana, and Clare E. Sims. ‘Not Only Size Matters: Early-Talker and Late-Talker Vocabularies Support Different Word-Learning Biases in Babies and Networks’. *Cognitive Science* 41, no. S1 (2017): 73–95. <https://doi.org/10.1111/cogs.12409>.
- Croft, William, and D. Alan Cruse. *Cognitive Linguistics*. 1st ed. Cambridge University Press, 2004. <https://doi.org/10.1017/CBO9780511803864>.

- Denny, Matthew. 'Social Network Analysis'. Massachusetts: University of Massachusetts, 2014.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding'. arXiv, 24 May 2019. <https://doi.org/10.48550/arXiv.1810.04805>.
- Engelthaler, Tomas, and Thomas T. Hills. "Modeling Early Word Learning Through Network Graphs." In *Network Science in Cognitive Psychology*, edited by Michael S. Vitevitch, p. 166-183. 1st ed. Routledge, 2019. <https://doi.org/10.4324/9780367853259>.
- Ettinger, Allyson. 'What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models'. *Transactions of the Association for Computational Linguistics* 8 (December 2020): 34–48. https://doi.org/10.1162/tacl_a_00298.
- Firth, John Rupert. *A Synopsis of Linguistic Theory, 1930-1955*, 1957.
- Gastaldi, Juan Luis. 'Why Can Computers Understand Natural Language?: The Structuralist Image of Language Behind Word Embeddings'. *Philosophy & Technology* 34, no. 1 (March 2021): 149–214. <https://doi.org/10.1007/s13347-020-00393-9>.
- Goldberg, Adele E. "The Nature of Generalization in Language." *Cognitive Linguistics* 20, no. 1 (January 2009). <https://doi.org/10.1515/COGL.2009.005>.
- Goldinger, Stephen D. "Words and Voices: Episodic Traces in Spoken Word Identification and Recognition Memory." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22, no. 5 (1996): 1166–83. <https://doi.org/10.1037/0278-7393.22.5.1166>.
- Hadley, Elizabeth B., David K. Dickinson, Kathy Hirsh-Pasek, and Roberta Michnick Golinkoff. 'Building Semantic Networks: The Impact of a Vocabulary Intervention on Preschoolers' Depth of Word Knowledge'. *Reading Research Quarterly* 54, no. 1 (January 2019): 41–61. <https://doi.org/10.1002/rrq.225>.
- Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 'Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change'. arXiv, 25 October 2018. <https://doi.org/10.48550/arXiv.1605.09096>.
- Hills, Thomas T., Mounir Maouene, Josita Maouene, Adam Sheya, and Linda Smith. "Longitudinal Analysis of Early Semantic Networks: Preferential Attachment or Preferential Acquisition?" *Psychological Science* 20, no. 6 (June 2009): 729–39. <https://doi.org/10.1111/j.1467-9280.2009.02365.x>.
- Hilpert, Martin. *Ten Lectures on Diachronic Construction Grammar*. Distinguished Lectures in Cognitive Linguistics, vol. 26. Leiden ; Boston: Brill, 2021.
- Harris, Zellig S. 'Distributional Structure'. *Word* 10 (1954): 146–62. <https://doi.org/10.1080/00437956.1954.11659520>.
- Humphries, Mark D., and Kevin Gurney. 'Network "Small-World-Ness": A Quantitative Method for Determining Canonical Network Equivalence'. *PLOS ONE* 3, no. 4 (Winter 2008): e0002051. <https://doi.org/10.1371/journal.pone.0002051>.

- Jarmasz, Mario, and Stan Szpakowicz. 'Roget's Thesaurus and Semantic Similarity'. In *Current Issues in Linguistic Theory*, edited by Nicolas Nicolov, Kalina Bontcheva, Galia Angelova, and Ruslan Mitkov, 260:111. Amsterdam: John Benjamins Publishing Company, 2004. <https://doi.org/10.1075/cilt.260.12jar>.
- Jescheniak, Jörg D., and Willem J. M. Levelt. "Word Frequency Effects in Speech Production: Retrieval of Syntactic Information and of Phonological Form." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20, no. 4 (July 1994): 824–43. <https://doi.org/10.1037/0278-7393.20.4.824>.
- Kenett, Yoed, David Anaki, and Miriam Faust. "Investigating the Structure of Semantic Networks in Low and High Creative Persons." *Frontiers in Human Neuroscience* 8 (June 10, 2014). <https://doi.org/10.3389/fnhum.2014.00407>.
- Kenett, Yoed N., Faust, Miriam. 'Clinical Cognitive Networks A Graph Theory Approach'. In *Network Science in Cognitive Psychology*, edited by Michael S. Vitevitch, p. 136-165. 1st ed. Routledge, 2019. <https://doi.org/10.4324/9780367853259>.
- Kipfer, Barbara Ann. *Roget's International Thesaurus*. HarperCollins Publishers, 2022.
- Lakhzoum, Dounia, Marie Izaute, and Ludovic Ferrand. 'Semantic Network Analysis of Abstract and Concrete Word Associations'. arXiv, 18 October 2021. <https://doi.org/10.48550/arXiv.2110.09096>.
- Lund, Kevin, and Curt Burgess. 'Producing High-Dimensional Semantic Spaces from Lexical Co-Occurrence'. *Behavior Research Methods, Instruments, & Computers* 28, no. 2 (1 June 1996): 203–8. <https://doi.org/10.3758/BF03204766>.
- McRae, Ken, George S. Cree, Mark S. Seidenberg, and Chris Mcnorgan. 'Semantic Feature Production Norms for a Large Set of Living and Nonliving Things'. *Behavior Research Methods* 37, no. 4 (November 2005): 547–59. <https://doi.org/10.3758/BF03192726>.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 'Efficient Estimation of Word Representations in Vector Space'. arXiv, 6 September 2013a. <https://doi.org/10.48550/arXiv.1301.3781>.
- Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. 'Exploiting Similarities among Languages for Machine Translation'. arXiv, 16 September 2013b. <https://doi.org/10.48550/arXiv.1309.4168>.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 'Distributed Representations of Words and Phrases and Their Compositionality'. arXiv, 16 October 2013c. <https://doi.org/10.48550/arXiv.1310.4546>.
- Miller, George A. 'WordNet: A Lexical Database for English'. *Communications of the ACM* 38, no. 11 (autumn 1995): 39–41. <https://doi.org/10.1145/219717.219748>.
- Nelson, Douglas L., 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, no. 3 (1 August 2004): 402–7. <https://doi.org/10.3758/BF03195588>.

- Newman, Mark. *Networks*. Vol. 1. Oxford University Press, 2018.
<https://doi.org/10.1093/oso/9780198805090.001.0001>.
- Pennington, Jeffrey, Richard Socher, and Christopher Manning. ‘GloVe: Global Vectors for Word Representation’. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by Alessandro Moschitti, Bo Pang, and Walter Daelemans, 1532–43. Doha, Qatar: Association for Computational Linguistics, 2014.
<https://doi.org/10.3115/v1/D14-1162>.
- Preininger, M., Brand, J., & Kříž, A. (2022). Quantifying the Socio-semantic Representations of Words. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44).
- Saussure, Ferdinand de. *Course in General Linguistics*. New York : Philosophical Library, 1959.
<http://archive.org/details/courseingenerall00saus>.
- Sidorov, Grigori, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. ‘Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model’. *Computación y Sistemas* 18, no. 3 (29 September 2014): 491–504. <https://doi.org/10.13053/cys-18-3-2043>.
- Singhal, Amit. ‘Modern Information Retrieval: A Brief Overview’. *IEEE Data Engineering Bulletin* 24 (1 January 2001).
- Siew, Cynthia S. Q., and Michael S. Vitevitch. ‘Spoken Word Recognition and Serial Recall of Words from Components in the Phonological Network.’ *Journal of Experimental Psychology: Learning, Memory, and Cognition* 42, no. 3 (2016): 394–410.
<https://doi.org/10.1037/xlm0000139>.
- Siew, Cynthia S. Q., and Michael S. Vitevitch. ‘An Investigation of Network Growth Principles in the Phonological Language Network.’ *Journal of Experimental Psychology: General* 149, no. 12 (December 2020): 2376–94. <https://doi.org/10.1037/xge0000876>.
- Steyvers, Mark, and Joshua B. Tenenbaum. ‘The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth’. *Cognitive Science* 29, no. 1 (2 January 2005): 41–78. https://doi.org/10.1207/s15516709cog2901_3.
- Storkel, Holly L., and Michele L. Morrisette. “The Lexicon and Phonology: Interactions in Language Acquisition.” *Language, Speech, and Hearing Services in Schools* 33, no. 1 (January 2002): 24–37. [https://doi.org/10.1044/0161-1461\(2002/003\)](https://doi.org/10.1044/0161-1461(2002/003)).
- Storkel, Holly L., Jonna Armbrüster, and Tiffany P. Hogan. “Differentiating Phonotactic Probability and Neighborhood Density in Adult Word Learning.” *Journal of Speech, Language, and Hearing Research* 49, no. 6 (December 2006): 1175–92.
[https://doi.org/10.1044/1092-4388\(2006/085\)](https://doi.org/10.1044/1092-4388(2006/085)).
- Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods*, 51(3), 1187–1204. <https://doi.org/10.3758/s13428-018-1056-1>

- Vinson, David P., and Gabriella Vigliocco. ‘Semantic Feature Production Norms for a Large Set of Objects and Events’. *Behavior Research Methods* 40, no. 1 (1 February 2008): 183–90. <https://doi.org/10.3758/BRM.40.1.183>.
- Vitevitch, Michael S., and Rutherford Goldstein. “Keywords in the Mental Lexicon.” *Journal of Memory and Language* 73 (May 2014): 131–47. <https://doi.org/10.1016/j.jml.2014.03.005>.
- Vitevitch, Michael S., ed. *Network Science in Cognitive Psychology*. 1st ed. Routledge, 2019. <https://doi.org/10.4324/9780367853259>.
- Vitevitch, Michael S., Gunes Ercal, and Bhargav Adagarla. ‘Simulating Retrieval from a Highly Clustered Network: Implications for Spoken Word Recognition’. *Frontiers in Psychology* 2 (2011). <https://doi.org/10.3389/fpsyg.2011.00369>.
- Vitevitch, Michael S. ‘What Can Graph Theory Tell Us About Word Learning and Lexical Retrieval?’ *Journal of Speech, Language, and Hearing Research* 51, no. 2 (April 2008): 408–22. [https://doi.org/10.1044/1092-4388\(2008/030\)](https://doi.org/10.1044/1092-4388(2008/030)).
- Vitevitch, Michael S., Nichol Castro, Gavin J. D. Mullin, and Zoe Kulphongpatana. ‘The Resilience of the Phonological Network May Have Implications for Developmental and Acquired Disorders’. *Brain Sciences* 13, no. 2 (23 January 2023): 188. <https://doi.org/10.3390/brainsci13020188>.
- Watts, Duncan J., and Steven H. Strogatz. ‘Collective Dynamics of “Small-World” Networks’. *Nature* 393, no. 6684 (June 1998): 440–42. <https://doi.org/10.1038/30918>.
- Zhou, Kaitlyn, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. ‘Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words’. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 401–23. Dublin, Ireland: Association for Computational Linguistics, 2022. <https://doi.org/10.18653/v1/2022.acl-short.45>.

7. Résumé

Cíl této diplomové práce bylo vytvořit sémantickou síť angličtiny založenou na slovních vektorech a prozkoumat zdali struktura takto vytvořené sítě odráží některé kognitivně relevantní vztahy mezi slovy. Taková práce na jedné straně zkoumá možnosti využití současných metod strojového učení, které vytváření vektory slov, pro lingvistický výzkum. Na straně druhé se angažuje v relativně nové a dynamické disciplíně jazykových sítí, která vytváří síťové modely různých aspektů jazyka a analyzuje je s pomocí formálních nástrojů vědy o sítích. Celá práce má ambici přispět k poznání o tom, jak mluvčí zpracovávají slova ve své mysli, a proto je výsledná sémantická síť konfrontovaná s výsledky psycholingvistických experimentů na zpracování slov s cílem otestovat, zdali struktura takové sémantické sítě, měřená skrze vybrané síťové metriky, má vliv na efektivitu zpracování slov mluvčími, měřenou skrze reakční časy pro daná slova ve výše zmíněných experimentech. Jde tedy o interdisciplinární práci, která kombinuje počítačové postupy pro vytvoření slovních

vektorů a jejich transformaci do sítě, formální nástroje vědy o sítích pro analýzu struktury výsledné sítě a výsledky psycholingvistických experimentů pro porovnání se síťovými metrikami.

Kapitola 2 postupně představuje nezbytnou teorii od vědy o sítích, psycholingvistický výzkum zpracování slov, konkrétní aplikace vědy o sítích pro lingvistický výzkum v podobě fonologických a sémantických sítí a nakonec velké jazykové modely a vektory slov. V kapitole 2.1 o vědě o sítích jsou stručně nastíněny její aplikace v jiných výzkumných odvětvích jako sociální sítě, které zkoumají struktury společenských vazeb. Je zde také představen koncept struktury malých světů (*small-world structure*), která je typická pro různé sociální a kognitivní fenomény reprezentované jako síť, např. sociální sítě mapující síť známostí mezi lidmi. Hlavní část je věnována představení relevantních síťových metrik jako *degree centrality*, *closeness centrality* a *clustering coefficient*, které byly potom využité pro porovnání struktury sémantické sítě a psycholingvistických výsledků. Kapitola 2.3 představila konkrétní aplikaci vědy o sítích v lingvistickém výzkumu fonologických sítí. Propojení mezi slovy ve fonologické síti je typicky založeno na vztahu fonologické podobnosti. V takové síti je pár slov propojen právě tehdy, liší-li se právě jedním fonémem. Jde převážně o výzkum Michaela S. Vitevitcha a jeho kolegů, na který má diplomová práce převážně navazuje. Vitevitch a kol. ve svém výzkumu často propojovali výzkum fonologických sítí s psycholingvistickými experimenty, ve kterých se ukázala nejasná a komplikovaná role *clustering coefficient* na rychlost zpracování slov (Vitevitch 2009, 2011). To byla hlavní inspirace pro to, zkusit prozkoumat roli různých síťových metrik v sémantických sítích a využít přitom současného rozvoje metod strojového učení pro zpracování jazyka. Kapitola 2.4 představila současný výzkum sémantických sítí. Propojení využívaná ve studiích sémantických sítí jsou založena na širším spektru sémantických vztahů. Takový vztah je typicky kvantifikován a od určité hranice blízkosti dvou slov jsou propojeny v rámci sémantické sítě. Tyto sémantické vztahy v sítích mohou být rozděleny do různých skupin na základě dat, ze kterých je takový sémantický vztah určen. Tato diplomová práce využívá vektory slov pro výpočet sémantické podobnosti na základě které jsou slova propojena do sémantické sítě. Tímto postupem spadá sémantická síť v mé diplomové práci do kategorie sémantických sítí založených na jazykových korpusech a kookurenci slov, na základě které algoritmy strojového učení využité v této práci vytváří vektory slov. Vektory slov a velké jazykové modely jsou tak poslední teoretickou částí, kterou je pro správnou interpretaci mé sémantické sítě nutné představit v kapitole 2.5. Jsou zde uvedeny některé současné jazykové modely jako BERT, které mohou být natrénovány textem na základě kterého vytvoří vektory

slov. Kapitola se snaží nezabíhat do příliš technických detailů a postupně představuje hlavně příklady relevantních využití slovních vektorů a tzv. sémantických prostorů, ve kterých slovní vektory zachycují různé hlavně sémantické vztahy mezi slovy. Hlavním cílem této kapitoly je na konkrétních příkladech ukázat, proč je vzdálenost mezi slovními vektory definičním kritériem pro vytvoření propojení v mé sémantické síti.

Kapitola 3 a 4 potom představují praktickou část mé diplomové práce. Kapitola 3 explicitně formuluje dvě základní výzkumné otázky, které jsem chtěl ve své práci zodpovědět. První je, zdali mohou být vektory slov relevantním základem pro vytvoření sémantické sítě, která poskytne vhled do kognitivních procesů jazyka. Tato otázka může být kladně zodpovězena, pokud bude mít výsledná síť strukturu malých světů. Navazující výzkumnou otázkou je, že pokud taková sémantická síť bude mít strukturu malých světů, zdali bude možné identifikovat konkrétní strukturní vlastnosti, které budou ovlivňovat efektivitu zpracování slov. Pro zodpovězení takové otázky je třeba statisticky otestovat vztah mezi síťovými proměnnými pro slova v síti a reakčními časy pro stejná slova z psycholingvistických experimentů. Kapitola 3 také zmiňuje dva zdroje dat mé práce, kterými je vzorek 1000 vět z TV Corpus, který jsem využil pro natrénování jazykového modelu BERT, abych získal vektory slov pro sémantickou síť. Druhým zdrojem jsou reakční časy pro jednotlivá slova v mé síti z psycholingvistických experimentů *lexical decision task*, které jsem získal z MALD databáze. Reakční časy odrážejí efektivitu zpracování slov a mohou tak být porovnána se síťovými metrikami pro stejná slova. Kapitola 3.1 popisuje detailněji postup a okolnosti vytváření sémantické sítě z výše uvedených zdrojů. Kapitola 3.2 potom podobně popisuje získání a vlastnosti psycholingvistických dat z databáze MALD. Kapitola 4 prezentuje výsledky mé práce, kterými jsou vizualizace sémantické sítě a lineární regrese. Ta zkoumala tři vybrané síťové metriky vypočítané pro každé slovo v síti - *degree centrality*, *clustering coefficient* a *closeness centrality* - jako nezávislé proměnné a jejich vliv na průměrný reakční čas pro stejná slova z *lexical decision task*. Regresní analýza ukázala potenciálně negativní vztah mezi *clustering coefficient* a průměrným reakčním časem. U zbylých dvou proměnných se neukázal jasný vztah. Nicméně výsledné vztahy se neukázaly jako statisticky signifikantní, jen negativní vztah mezi *clustering coefficient* a průměrným reakčním časem se blížil statistické signifikanci.

Kapitola 5 interpretuje výsledky a uvádí je do kontextu. Výsledná sémantická síť má strukturu malých světů, takže první výzkumná otázka může být zodpovězena kladně – slovní vektory mohou zachytit kognitivně relevantní sémantické vztahy mezi slovy. Druhá otázka je zodpovězena spíše negativně tím, že mezi vybrannými síťovými proměnnými a reakčním

časem, který reflektuje efektivitu zpracování slov, se neukázal statisticky signifikantní vztah. Závěr je i přesto optimistický, protože se nabízí různé možnosti, jak na výzkum v mé diplomové práci přímo navázat a statisticky signifikantní vztah mezi strukturou takové sémantické sítě a zpracováním slov prokázat. Budoucí výzkum může zkoumat možnosti využití jiných textů pro trénink jazykového modelu – nabízí se třeba korpusy mluveného jazyka, nebo experimentální elicitované vyprávění jednotlivých mluvčích. Je také možné se zaměřit na jiné, nebo komplexnější síťové metriky, než ty vybrané pro mou práci. Posledně je také možné prozkoumat možnosti využití jiných jazykových modelů, než je BERT. Celkově jsou jazykové sítě rostoucím odvětvím lingvistického výzkumu s potenciálním využitím například pro diagnostické účely jazykových poruch, které vyžadují zachycení a popis často subtilních změn řeči. Takovým požadavkům nahrává současný rozvoj metod strojového učení, který třeba skrz slovní vektory nově zvládne kvantitativně zachycovat sémantické vztahy mezi slovy. Má práce snad zvládla demonstrovat užitečnost jak sémantických sítí, tak slovních vektorů pro lingvistický výzkum.

8. Apendix

Item	meanRT	DegreeCentrality	ClosenessCentrality	ClusteringCoefficient
abuse	910.9091	0.003128	0.001043	0.166667
academy	931.6667	0.002086	0	0
advanced	1019.2	0.005214	0.054697	0.15
affects	879.1111	0.002086	0	0.5
agents	792	0.005214	0.002781	0.15
aggression	1040.25	0.003128	0.026561	0.5
agreement	824.4444	0.003128	0.014175	0
air	799.875	0.014599	0.034306	0.090909
animals	984.875	0.006257	0.024453	0.1
answer	716	0.016684	0.049623	0.126374
ape	857	0.003128	0	0.166667
approach	737.1111	0.004171	0.019123	0.166667
area	847.875	0.004171	0	0.083333
army	693.1111	0.006257	0.002781	0.066667
arts	828.3333	0.002086	0.001854	0
ask	836.3333	0.017727	0.047786	0.17619
asked	874.4	0.020855	0.004965	0.111111
assassins	883	0.007299	0.022942	0.02381
assignment	1010.273	0.004171	0.007764	0.25
auction	910.625	0.004171	0.030269	0
audience	986.3333	0.003128	0.02881	0.333333
aunt	1032.875	0.002086	0.018142	0
authorities	958	0.003128	0.010991	0.5

authorized	1080.556	0.006257	0.042137	0.166667
baby	751.5	0.013556	0.009675	0.118182
bag	918	0.006257	0.029155	0.166667
barriers	890.8889	0.002086	0.019287	0
based	941.3333	0.002086	0	0.5
bastard	814.1111	0.004171	0.032392	0.166667
bathroom	809.5	0.004171	0.011606	0
beam	813	0.004171	0.02523	0.166667
bear	736.1818	0.009385	0.02887	0.119048
bed	805	0.001043	0	0
behave	794.7778	0.006257	0.027876	0.366667
belief	847.5556	0.005214	0.012976	0.1
believe	1037.5	0.031283	0.016224	0.137566
belong	785	0.003128	0.001043	0
bidding	1024.125	0.002086	0	0
birthday	797.875	0.002086	0.017716	0
bit	843.6667	0.010428	0.069121	0.142857
body	864	0.01147	0.00139	0.013889
book	606.8889	0.005214	0.024962	0.05
boss	816.5556	0.03024	0.022092	0.432266
bother	772.7143	0.004171	0.003128	0
bottom	909.5556	0.002086	0	0.5
brake	810.8	0.008342	0.024522	0.142857
bride	806.625	0.001043	0	0
broken	839.25	0.002086	0	0
brother	729.6667	0.014599	0.033456	0.181818
brought	773.6667	0.003128	0.018394	0.333333
bubble	798.6667	0.003128	0.003926	0.166667
building	856.3	0.009385	0.00237	0.02381
built	895	0.003128	0.002208	0.166667
bunch	896.625	0.005214	0.014251	0.1
business	773.125	0.003128	0.002897	0.333333
buy	861.625	0.012513	0.020691	0.188889
call	783.25	0.039625	0.00393	0.096825
called	963.875	0.014599	0.024199	0.106061
calls	1058.25	0.007299	0.008854	0.261905
came	1057.5	0.025026	0.06233	0.151515
camp	905	0.007299	0.039379	0.190476
captain	924.875	0.003128	0.001564	0.166667
car	694.5	0.009385	0.01604	0.238095
care	767.3333	0.009385	0.037857	0.166667
careless	810.5556	0.002086	0	0
cartoon	919.5556	0.002086	0.001043	0.5
case	766.2222	0.009385	0.032086	0.111111
catching	822.2222	0.005214	0.014523	0.25
cell	779.5	0.007299	0.009511	0.3
century	947.75	0.007299	0.0275	0.2
challenge	924.75	0.001043	0	0

chance	1048.667	0.003128	0	0
charge	904.875	0.017727	0.044726	0.25
charity	828.125	0.002086	0	0
check	769.5	0.005214	0.020311	0.15
checked	808.625	0.003128	0.015541	0.166667
chest	774.3333	0.002086	0.017727	0
chief	870	0.002086	0.00139	0.5
child	823.25	0.007299	0.02562	0.35
children	879.25	0.006257	0.023693	0.033333
chocolate	784.1429	0.006257	0.019728	0.366667
choking	831.3636	0.007299	0.02213	0.404762
chose	966.2222	0.003128	0.011014	0.5
chuck	982.2	0.035454	0.066298	0.369875
church	744.25	0.001043	0	0
cinema	831	0.003128	0.015851	0
civilians	1053.556	0.005214	0.028376	0
claims	760.4444	0.003128	0.025547	0.333333
cleaning	1011.875	0.006257	0.017529	0.166667
client	931	0.008342	0.056758	0.053571
close	801.5	0.004171	0.024193	0.5
coconut	833.7273	0.003128	0	0.5
code	640	0.001043	0	0
coins	776.875	0.002086	0.00139	0
collection	1000.444	0.002086	0.011924	0
collections	841.2857	0.003128	0.010757	0
come	768	0.092805	0.015448	0.105319
coming	790	0.045881	0.085166	0.317653
commitment	940.5	0.010428	0.041713	0.1
company	949.25	0.047967	0.045315	0.289372
computer	850.1111	0.01147	0.016526	0.027778
concerned	921.625	0.006257	0.010707	0
conditions	839.1	0.001043	0	0
consider	777.875	0.004171	0.002503	0
consultant	1022.667	0.002086	0	0
continues	941	0.008342	0.022803	0.339286
control	871.875	0.012513	0.04836	0.033333
cook	910.25	0.01147	0.025936	0.136364
cooler	835.5714	0.001043	0	0
cop	859.125	0.006257	0.011068	0.233333
cops	909.5455	0.012513	0.013106	0.144444
corner	874.375	0.001043	0	0
costume	840.7778	0.004171	0.034322	0
coughing	859.3	0.007299	0	0.404762
council	895.625	0.001043	0	0
country	779.75	0.003128	0.019789	0.333333
couple	815.75	0.007299	0.018269	0.142857
course	789.8889	0.005214	0.005178	0
court	858.1	0.005214	0.001043	0.166667

cousin	806.5	0.012513	0.005035	0.144444
cracking	782.25	0.005214	0.021491	0.2
crew	981.625	0.008342	0.04072	0.133333
cross	902.125	0.006257	0.023212	0.2
crying	862.9	0.003128	0.006763	0.333333
curse	896.5455	0.001043	0	0
cute	800	0.002086	0.001043	0
dad	683.3333	0.013556	0.006703	0.181818
dale	1027.818	0.001043	0	0
date	707.75	0.006257	0.001043	0.066667
dawn	861.7778	0.002086	0	0.5
day	729	0.009385	0.024684	0.261905
days	649.125	0.017727	0.020627	0.138095
deal	788.625	0.007299	0.039951	0.142857
death	671.5556	0.013556	0	0.045455
decide	860	0.006257	0.025861	0.2
decided	832.1429	0.006257	0.01435	0.1
delivery	923.6667	0.004171	0.022864	0.083333
demonstrate	940.5	0.001043	0	0
department	964	0.005214	0.020944	0
depends	1057	0.006257	0.026822	0.066667
detective	1064.222	0.006257	0.001043	0.266667
determine	906.2	0.008342	0.017631	0.267857
did	825.5	0.055266	0.01437	0.101961
died	897.875	0.006257	0.003128	0.083333
dinner	687.875	0.014599	0.044425	0.098485
discount	854.5	0.004171	0.027069	0.333333
diversified	850.1111	0.002086	0.001043	0.5
do	945.7143	0.119917	0.004171	0.08573
doctor	774	0.01147	0.018863	0.097222
documents	928.7778	0.01147	0.04313	0.036364
does	782.0909	0.027112	0.027083	0.141304
doing	876.125	0.046924	0.001043	0.070321
done	733.875	0.017727	0.030192	0.138095
doubt	723.4444	0.003128	0.010548	0.166667
drag	823.9	0.013556	0.008009	0.072727
drained	935.625	0.008342	0.019968	0.125
draw	773.2222	0.012513	0.056194	0.128788
dressed	987.5556	0.010428	0.039667	0.107143
drink	812.3333	0.004171	0	0.333333
drive	780.375	0.009385	0.003193	0.138889
driving	912.25	0.013556	0.014271	0.102564
drop	822.125	0.007299	0.053466	0.333333
drug	677.3333	0.006257	0.032477	0.033333
dude	928.75	0.009385	0.06307	0.486111
dust	760.75	0.002086	0.019173	0.5
embrace	1209	0.01147	0.023381	0.181818
employment	805.4444	0.004171	0.009155	0.166667

end	826	0.004171	0	0
energy	791.5	0.014599	0.026816	0.045455
engine	945.5556	0.009385	0	0.138889
enjoy	806.6667	0.002086	0	0.5
events	919	0.003128	0.01155	0
exist	816	0.015641	0.050783	0.160256
expect	864.3333	0.007299	0.007861	0.2
experiment	793.25	0.005214	0.002681	0.05
explain	948.875	0.003128	0	0
expression	1001.375	0.006257	0.019668	0.066667
eyes	872	0.007299	0.016722	0.05
face	950.8333	0.005214	0.008112	0.1
faced	881.4444	0.007299	0.019051	0.071429
faith	794.2222	0.005214	0.013407	0.1
fall	905.2222	0.003128	0	0
family	766.1429	0.01147	0.008274	0.109091
feel	869.6	0.056309	0.079338	0.189291
feelings	987.875	0.007299	0.00237	0.15
fellow	1028.875	0.007299	0.033043	0.166667
felt	1059	0.007299	0.018853	0.2
fence	894.2	0.006257	0.033687	0.1
fiance	838.2222	0.027112	0.083399	0.332308
field	1050.667	0.006257	0.022988	0.133333
fig	878.3333	0.002086	0.001043	0
fight	779	0.028154	0.040275	0.058333
fighting	1075.375	0.003128	0.031595	0
figured	1082.875	0.012513	0.046895	0.136364
file	850	0.007299	0.040714	0.071429
find	754.5	0.046924	0.013968	0.084164
fingers	884.5	0.002086	0.001043	0
finish	918.5455	0.006257	0.03695	0.266667
fire	793.4444	0.015641	0.01086	0.032051
firing	1149.9	0.003128	0.009159	0.166667
fits	1001	0.002086	0	0.5
flap	997	0.006257	0.01484	0.133333
flowers	819.6364	0.003128	0.013609	0.166667
flows	1112.667	0.01147	0.028767	0.166667
food	768.125	0.015641	0.05075	0.166667
fought	832.5	0.002086	0	0
found	966	0.003128	0.018734	0.166667
friend	742.1111	0.015641	0.041864	0.147619
friends	978.5	0.016684	0.009385	0.071429
game	769	0.007299	0.008596	0
garage	799.5	0.010428	0.028351	0.122222
gather	956.5	0.006257	0.062945	0.1
gave	865.875	0.014599	0.03131	0.203297
gentlemen	776.4444	0.007299	0.010896	0.095238
get	844.25	0.090719	0.017794	0.112745

gets	892.5556	0.008342	0.001877	0.410714
getting	782.4444	0.012513	0.066821	0.122222
girl	830.125	0.020855	0.00139	0.088235
girls	815.5556	0.003128	0.007886	0.166667
give	743.6	0.070907	0.026994	0.163636
giving	805.625	0.002086	0.00237	0
glass	1122	0.008342	0.005735	0.233333
go	696.1	0.06048	0.01165	0.08539
gods	796.5556	0.002086	0.001043	0
goes	884.5556	0.016684	0.044112	0.25
goin	1116.6	0.006257	0.008874	0.4
going	718.375	0.035454	0.036283	0.104839
gone	991.3333	0.014599	0.013651	0.121212
got	791.625	0.062565	0.025777	0.086207
grabbed	1000.667	0.013556	0.049297	0.141026
grant	787.1429	0.006257	0.031595	0.166667
grew	1147.778	0.006257	0.022118	0.233333
group	892	0.003128	0.007108	0
growing	837.8889	0.004171	0.002781	0
guess	803.75	0.013556	0.029778	0.136364
gunfire	830.625	0.01877	0.02297	0.116667
guys	772	0.017727	0.007008	0.071429
had	858.7778	0.043796	0.036227	0.051923
hand	881.8889	0.006257	0	0
hands	735.25	0.001043	0	0
has	866	0.031283	0.021873	0.113757
hate	689.875	0.016684	0.014998	0.225275
hats	929.7778	0.005214	0.030096	0.15
have	834	0.070907	0.007678	0.074825
having	748.2	0.015641	0.002781	0.044872
head	834.375	0.005214	0.026802	0
health	857	0.006257	0	0.083333
hear	789.1111	0.019812	0.065814	0.231618
heard	790.125	0.016684	0.032518	0.093407
heart	893.1667	0.007299	0.022946	0
help	712.2222	0.032325	0.049015	0.124384
history	886.2222	0.002086	0.01602	0.5
hit	801.75	0.040667	0.04415	0.345479
hits	790.625	0.004171	0.001668	0.083333
hold	1009	0.019812	0.012101	0.055147
hole	882	0.001043	0	0
home	870.375	0.017727	0.026507	0.071429
homework	809.1	0.008342	0.027428	0.107143
honest	830.1111	0.003128	0.016803	0.166667
honey	893.875	0.002086	0	0
hope	821	0.032325	0.034347	0.386022
horn	831.6667	0.004171	0.018835	0.416667
hospital	825.3636	0.007299	0.020869	0.119048

hours	878.375	0.007299	0.022464	0.3
house	690.125	0.021898	0.019705	0.038012
humiliated	906.625	0.003128	0.030411	0.166667
hurt	909	0.03024	0.055639	0.111111
husband	722.375	0.015641	0.006757	0.108974
ice	954.6667	0.005214	0.008002	0.15
idea	838.7273	0.003128	0	0
idiot	727.25	0.002086	0	0
imagination	938.1111	0.002086	0.00139	0
invites	1082.2	0.005214	0.01723	0.2
job	740.6667	0.013556	0.00958	0.090909
join	787.1	0.002086	0.024962	0
joints	1168	0.003128	0.00139	0
joke	733.6364	0.005214	0.023944	0
keep	735.3636	0.009385	0	0.152778
kept	961.5556	0.002086	0.001043	0.5
kid	795	0.016684	0.026762	0.098901
kidding	784.3333	0.004171	0.006763	0
killed	796.625	0.021898	0.049739	0.104762
killer	806.25	0.01147	0.055249	0.127273
kind	930.1111	0.005214	0.026872	0.15
knew	754.6667	0.021898	0.057085	0.116959
know	904	0.051095	0.007422	0.094357
known	923.3	0.005214	0.018307	0.1
knows	1068.5	0.015641	0.027812	0.173077
lab	945.4444	0.006257	0.04045	0.066667
labyrinth	996.7778	0.001043	0	0
lack	821	0.001043	0	0
lake	840.2222	0.002086	0	0
laugh	872	0.002086	0.001043	0
laughs	732.875	0.008342	0	0.285714
lawyers	792.7	0.001043	0	0
leaked	1060.111	0.01147	0.01956	0.145455
learn	778.875	0.008342	0.04315	0.232143
leave	808.5	0.059437	0.00869	0.224579
left	840.3333	0.013556	0.004965	0.1
let	945.625	0.085506	0.004171	0.122468
letting	876.3636	0.004171	0.004965	0.083333
liar	786.4444	0.002086	0	0
lieutenant	802.2222	0.006257	0.012054	0.166667
life	840	0.028154	0.020341	0.026667
lift	938.2222	0.002086	0.018142	0
lifted	734.8889	0.001043	0	0
list	955.1	0.004171	0.053553	0
listen	717.375	0.025026	0.047438	0.186147
live	914.5	0.003128	0.017963	0.166667
lives	711.1111	0.006257	0.024347	0.2
living	736.875	0.009385	0.03673	0.055556

lobby	869.5714	0.005214	0.01573	0.15
looking	1246.833	0.008342	0.022637	0.160714
lose	808.4444	0.014599	0.029197	0.142857
loser	833.1111	0.002086	0.014573	0
lot	907.6364	0.014599	0.009067	0.068182
love	756.125	0.025026	0.024977	0.0671
lunch	745	0.004171	0.035838	0.083333
lying	957.3333	0.005214	0.03883	0
madame	884.4444	0.004171	0.008174	0.166667
made	936.3	0.034411	0.024025	0.107955
magazine	810.625	0.002086	0	0
maintaining	1106.5	0.003128	0.038123	0.166667
make	999.1	0.042753	0.033648	0.107287
making	664.6667	0.004171	0.036004	0.166667
male	756	0.003128	0.001043	0.333333
males	880.8	0.003128	0	0.333333
man	791	0.028154	0.041217	0.066667
manager	991.8889	0.005214	0.034997	0.166667
mark	928.6667	0.006257	0.00139	0
market	894.5556	0.006257	0.035861	0.083333
mean	994.625	0.01877	0.016918	0.195833
memory	839.625	0.003128	0.041081	0.166667
men	707.375	0.009385	0.034893	0.125
mention	859.5556	0.002086	0.018142	0.5
milk	693.5	0.010428	0.024767	0.2
mind	808.25	0.004171	0.001043	0
minute	873.375	0.002086	0	0
miss	967.5714	0.012513	0.01068	0.188889
mistake	929.875	0.005214	0.017379	0.1
mixed	1014.727	0.002086	0.001877	0
model	777.8889	0.004171	0.034152	0.25
mom	718.8571	0.015641	0.006257	0.102564
money	632.25	0.003128	0.001043	0
monitor	865.5714	0.019812	0.068188	0.040936
month	784.5	0.01147	0.018654	0.208333
months	768.625	0.003128	0.014469	0.333333
morning	768.5556	0.014599	0.024456	0.143939
mother	879.1111	0.016684	0.025297	0.10989
motor	595	0.003128	0.012574	0.5
move	790.875	0.020855	0.024219	0.104575
moving	815.6667	0.012513	0.032673	0.066667
music	748.5455	0.015641	0.001043	0.102564
name	862.2	0.019812	0.051442	0.069853
names	907.6	0.005214	0.047017	0.1
needs	850.3636	0.007299	0.023752	0.15
news	858.2222	0.015641	0.028491	0.108974
night	848	0.010428	0.021389	0.071429
object	763.625	0.007299	0.002897	0.1

ocean	953.5556	0.003128	0.006282	0
offer	902.8889	0.003128	0.019032	0
offering	848	0.006257	0.002902	0
office	851	0.003128	0.00721	0.166667
officer	911.6667	0.009385	0	0.095238
online	660.6667	0.002086	0	0
operation	967.25	0.004171	0.035129	0.083333
ounce	916.875	0.003128	0	0.166667
package	855.8889	0.004171	0.024621	0.166667
page	791.4444	0.004171	0.001043	0
paint	693.9	0.002086	0	0
palm	884.8889	0.001043	0	0
part	895.1111	0.002086	0.001877	0
party	1062.25	0.014599	0.028906	0.022727
passionate	900.5	0.002086	0.018668	0.5
past	795.6667	0.001043	0	0
path	803.1	0.009385	0.009584	0.069444
patient	877.5556	0.012513	0.01852	0.083333
patterns	1029.857	0.003128	0.038566	0
peace	697.8889	0.001043	0	0
people	688.3333	0.069864	0.006527	0.15024
perform	908.5	0.014599	0.058757	0.137363
person	779.125	0.009385	0.032456	0.095238
pets	807.3333	0.004171	0.020992	0.083333
phone	975.625	0.01147	0.019446	0.125
photo	759	0.01147	0.043644	0.097222
photon	970.25	0.001043	0	0
pick	630.6	0.002086	0.019614	0
picked	691.6	0.004171	0.020732	0.166667
picking	820.125	0.004171	0.024236	0.083333
picture	818	0.014599	0.017756	0.098485
pictures	788.5	0.006257	0.044007	0.066667
pie	748.4545	0.008342	0	0.321429
place	1069.833	0.015641	0.016455	0.038462
places	1040.333	0.006257	0.015082	0.066667
planned	960.2857	0.004171	0.011494	0
play	738.1111	0.01147	0.050151	0.081818
pole	949.5	0.037539	0.089843	0.42381
police	784.4	0.009385	0.043332	0.119048
pop	944.125	0.001043	0	0
practices	869.2222	0.010428	0.001877	0.188889
preparatory	1105.8	0.006257	0.055459	0.166667
press	909.625	0.009385	0.051586	0.180556
priority	609	0.003128	0.018142	0
prisoner	807.7778	0.003128	0.034282	0.166667
privileges	904.7778	0.005214	0.024684	0.15
procedures	986.1818	0.009385	0.053927	0.069444
program	797.5	0.005214	0.002086	0.1

programming	987.6667	0.004171	0.034959	0.083333
protect	799.625	0.021898	0.021352	0.102339
pull	952.8889	0.009385	0.011215	0.142857
purse	766.5	0.003128	0.035398	0
pursued	988.125	0.01147	0.054256	0.145455
put	777.2222	0.03024	0.045095	0.122507
puts	812.2857	0.010428	0.025719	0.155556
question	847	0.007299	0.036357	0.25
race	866.3333	0.006257	0.022775	0.2
radio	1044.2	0.004171	0.025004	0.25
rage	767.5556	0.003128	0.001043	0
raise	863.625	0.027112	0.090807	0.281538
ran	867.1429	0.003128	0.009346	0.333333
re	860.7778	0.043796	0	0.105128
realized	993	0.007299	0.01155	0.214286
rebuild	1105.667	0.002086	0.002208	0
received	831.7778	0.002086	0	0
recognize	1065.667	0.017727	0.023497	0.095238
record	993.6667	0.004171	0	0.166667
relationship	799	0.014599	0.017331	0.083333
relax	886.8333	0.003128	0.017963	0.166667
remains	978.3333	0.001043	0	0
remember	757.1111	0.007299	0.015704	0.2
researchers	868.8889	0.005214	0.032735	0.1
resembling	782.125	0.001043	0	0
restraining	858.375	0.002086	0	0
results	1044.5	0.01147	0.039634	0.490909
return	893.75	0.003128	0.028816	0
roast	872.7273	0.002086	0.019614	0
rock	741.3333	0.008342	0.028583	0.089286
room	806.1818	0.016684	0.019326	0.06044
rose	766	0.004171	0.01604	0.166667
route	763	0.003128	0.032794	0.166667
ruled	851.875	0.001043	0	0
run	658.6667	0.01147	0.027784	0.081818
sacks	1100.375	0.001043	0	0
safety	981.5	0.006257	0.063869	0.2
said	956.5556	0.025026	0.023554	0.127706
sale	1104.25	0.002086	0	0.5
saved	952.1429	0.019812	0.029661	0.121324
saw	822.5	0.009385	0.033797	0.190476
say	849.25	0.04171	0.056629	0.139403
saying	922.75	0.014599	0.00139	0.121212
says	973.6667	0.046924	0.033544	0.243939
scene	976.8	0.005214	0	0.05
school	945.125	0.012513	0.008833	0.05303
scramble	833.875	0.002086	0.00139	0
sea	811.625	0.005214	0.007499	0.1

secluded	934.1111	0.007299	0.037308	0.166667
seconds	798.4545	0.006257	0.022905	0.25
security	922	0.007299	0.03831	0.190476
see	973.875	0.036496	0.023325	0.08428
seed	891.5	0.001043	0	0
seemed	958	0.013556	0.029879	0.154545
seems	920.2857	0.026069	0.048217	0.13834
seen	1016.875	0.015641	0.050054	0.089744
sell	886.25	0.036496	0.045127	0.274621
sells	855.5455	0.01877	0.054223	0.160131
senator	799.8571	0.002086	0	0.5
send	1129.222	0.006257	0.032296	0.1
sending	863.1667	0.010428	0.037507	0.1
sent	1004.375	0.006257	0.021583	0.266667
sentence	932.7778	0.004171	0.027581	0.333333
service	886.8889	0.003128	0.015404	0.333333
serving	1047.7	0.002086	0.001877	0.5
set	861.8	0.015641	0.042955	0.147619
setting	657.5	0.013556	0.08294	0.070513
shape	795.375	0.002086	0.002503	0
shaped	821	0.004171	0.031491	0.416667
share	847.6667	0.008342	0.02615	0.428571
sharks	904.3333	0.001043	0	0
shock	833.5	0.005214	0.028989	0.05
shoot	1022.125	0.033368	0.031563	0.392137
shotgun	828.5	0.003128	0	0.166667
show	954.875	0.009385	0.011638	0.047619
shows	1041.25	0.002086	0.00958	0
shut	864.2	0.017727	0.022354	0.352381
side	1100.3	0.005214	0	0
sighs	1192	0.013556	0.003128	0.236364
sight	934.25	0.004171	0.024224	0.166667
singing	931.125	0.010428	0.03148	0.133333
sirens	996	0.006257	0	0.466667
sister	925.625	0.007299	0.005735	0.142857
sit	773.75	0.006257	0.021902	0.166667
slaves	1033.75	0.003128	0.004985	0.333333
sobs	944.3333	0.005214	0.003337	0.45
sold	1145.222	0.005214	0.035985	0.05
sort	1015.111	0.037539	0.034543	0.180036
soul	887.375	0.003128	0	0
soup	905.7778	0.012513	0.025992	0.189394
spaghetti	829.25	0.001043	0	0
speak	818.8889	0.007299	0.001043	0.05
speech	1007	0.009385	0.029712	0.111111
sperm	934.875	0.001043	0	0
spit	849.2222	0.009385	0.02157	0.222222
splendid	826.7778	0.002086	0.005456	0.5

spot	1115.25	0.010428	0.033887	0.1
spotted	871.5	0.008342	0.050161	0.089286
stabbed	931.8	0.002086	0.038688	0
staff	960.25	0.004171	0.049308	0
stain	1115.375	0.003128	0.002086	0
stand	892.6667	0.005214	0.005214	0.15
standing	954	0.01147	0.036483	0.055556
start	939.75	0.012513	0.025236	0.122222
started	1236.375	0.015641	0.001043	0.07619
statement	976.1	0.001043	0	0
stations	1101.909	0.006257	0.06675	0.133333
stay	1082.6	0.063608	0.069704	0.191993
stealing	887.7143	0.002086	0	0
step	933.7143	0.001043	0	0
stink	915.3	0.002086	0	0
stole	1103	0.008342	0.044071	0.178571
stop	801.875	0.047967	0.017772	0.250529
stopped	927.5	0.002086	0	0
stops	986.5	0.005214	0.019662	0.5
store	1120.6	0.002086	0	0
strained	1151	0.001043	0	0
students	916.4444	0.006257	0	0.033333
study	968.25	0.009385	0.024092	0.111111
stuff	873.6667	0.020855	0.021132	0.071895
summer	908.5	0.002086	0	0.5
supply	919.5	0.007299	0.023401	0.05
surface	1038	0.005214	0.004011	0.2
surge	969.6667	0.007299	0.001043	0.095238
surprises	972.3333	0.001043	0	0
surround	991.4444	0.005214	0.039273	0.35
surveillance	926.8333	0.005214	0.024709	0.05
sweep	1016.375	0.002086	0.025442	0
swept	943.4444	0.007299	0.022888	0.071429
swim	889.8889	0.003128	0.004965	0
tables	807.8889	0.003128	0.035791	0.166667
take	912.375	0.027112	0.02696	0.144928
taking	830.5	0.008342	0.024402	0.053571
talk	741.625	0.04171	0.031283	0.128734
talking	770.2222	0.027112	0.008874	0.072464
tape	566.7143	0.008342	0.054286	0.089286
tapes	956.5556	0.006257	0.014645	0.066667
teach	754	0.016684	0.051004	0.153846
team	935.4286	0.013556	0.00237	0.027273
teenagers	779.2222	0.005214	0.018242	0.15
tells	941.5	0.008342	0.038579	0.339286
terrorist	779.25	0.005214	0	0.2
thank	985.3333	0.040667	0.036725	0.361862
thanks	937	0.013556	0.023814	0.1

theory	828.2222	0.003128	0.002086	0
thing	929.25	0.015641	0.003259	0.083333
things	837.875	0.025026	0.052535	0.080087
think	909	0.044838	0.001043	0.119512
thinks	808	0.014599	0.054742	0.208791
thought	960.5	0.022941	0.01535	0.078947
thrust	886.75	0.003128	0.01155	0.5
thunder	952.5	0.003128	0.002383	0.166667
tickets	838.25	0.001043	0	0
times	1008.429	0.009385	0.041628	0.047619
tires	1112	0.010428	0.021063	0.222222
today	715.6667	0.042753	0.065916	0.283537
told	752.625	0.03024	0.017498	0.08547
tomorrow	826.8571	0.010428	0.015593	0.267857
ton	792.1667	0.001043	0	0
tonight	885.375	0.010428	0.03481	0.107143
tons	794.1429	0.007299	0.020752	0.190476
took	999.3333	0.01877	0.011237	0.183333
track	869.2222	0.005214	0.002208	0.05
train	877.3333	0.003128	0	0
travel	975.7	0.004171	0.036104	0
treat	731.25	0.020855	0.007508	0.110526
tricked	737.25	0.004171	0.001043	0.083333
trophy	931.2	0.005214	0	0
truck	745.7778	0.004171	0.031351	0.083333
trust	730.375	0.002086	0.013338	0.5
try	937.7	0.005214	0.022016	0.333333
trying	843.2222	0.013556	0.021216	0.118182
tumor	771.5	0.040667	0.011216	0.376518
tutor	918.1111	0.003128	0.007008	0.333333
uncle	756.875	0.007299	0.033975	0.190476
understand	936	0.001043	0	0
uniform	965.125	0.003128	0.013909	0.166667
union	783.1111	0.004171	0.020209	0.25
units	953.1667	0.002086	0.01637	0
used	738.7778	0.014599	0	0.071429
vacancy	1076.1	0.003128	0	0.166667
versions	990.4	0.002086	0.028007	0.5
victim	783.3333	0.017727	0.018004	0.07619
victims	892.5556	0.006257	0.04109	0.1
volunteer	1068.889	0.008342	0.051885	0.125
waist	724.4444	0.001043	0	0
wait	825.7778	0.028154	0.021583	0.386667
walk	786.1111	0.004171	0.015861	0.333333
walking	995.4	0.008342	0.069009	0.178571
want	906.8	0.03024	0.012066	0.176638
wanted	810	0.017727	0.040969	0.104762
wanting	875.3333	0.004171	0.031886	0.25

wants	970.375	0.010428	0.048494	0.311111
war	819.25	0.009385	0.001043	0
warrior	898.5	0.001043	0	0
watch	985.625	0.002086	0.016471	0
water	787	0.005214	0.024962	0.166667
way	834.3333	0.020855	0.018248	0.144737
weapon	757.25	0.008342	0.029299	0
wearing	1003.111	0.009385	0.024128	0.142857
week	711.3333	0.012513	0	0.177778
weeks	973.5	0.006257	0.017554	0.2
went	826.5	0.010428	0	0.232143
whip	935.375	0.003128	0.002897	0.166667
wife	814	0.01147	0.028282	0.097222
window	840.2	0.006257	0.01643	0
wine	982.4286	0.007299	0.024896	0.190476
winning	716.3	0.004171	0	0.083333
wishing	751.625	0.002086	0.002639	0
witness	835.5556	0.004171	0.039665	0.416667
woman	724.5556	0.012513	0.039082	0.188889
won	786.6364	0.007299	0.04771	0.047619
word	688.4545	0.003128	0.028926	0.166667
work	717.3333	0.029197	0.06275	0.096923
working	756.7	0.010428	0.008009	0.160714
world	938.5	0.017727	0.033171	0.071429
worry	805.5	0.002086	0.008258	0
wrap	728.375	0.003128	0.021514	0
wrapped	840	0.005214	0.025294	0.2
yards	933	0.002086	0.021861	0.5
year	752.375	0.013556	0.024199	0.109091
years	937	0.012513	0.025784	0.122222
yells	1055	0.014599	0.021984	0.214286
yesterday	943.625	0.006257	0.003406	0.25