More than the Sum of Their Words

Generating and Contrasting Large Linguistic Networks

Hanna Schmück, M.A.

Department of Linguistics and English Language

Lancaster University

A thesis submitted for the degree of Doctor of Philosophy in Linguistics

September 2024

More than the Sum of Their Words

Generating and Contrasting Large Linguistic Networks

Hanna Schmück

"You shall know a word by the company it keeps" (Firth, 1957), and, as this thesis attempts to demonstrate, also by how the company is kept. The primary motivation of this thesis is establishing a connection of current psycholinguistic evidence, i.e. experimental and theoretical findings regarding the structural design of the mental lexicon, to empirical findings from large-scale corpus-based collocation networks. One contribution of this work therefore lies in the triangulation (Noble & Heale, 2019, p. 67) of corpus linguistics, psycholinguistics and graph theory: bridging gaps between these approaches to language and developing new viewpoints on the data might help overcome or seriously limit fundamental biases and present a more well-founded manner of interpreting results from collocation analyses with regards to their capabilities of portraying mental processes and acting as a proxy for how readers/speakers perceive certain concepts. In order to address the existing research gap, a large-scale analysis of computationally generated corpus-based collocation networks based on the BNC 2014 and psycholinguistic word association networks based on the word association database SWOW-UK is carried out here. Word associations have been chosen as the basis for the psycholinguistic network since they portray the perceived relation between concepts via discrete linguistic units (Kang, 2018, p. 87), similarly to collocations. From the theoretical perspective, in addition to new insights into current open questions regarding the structure and organisation of collocational knowledge, this approach also provides new research prompts for investigating the internal structure of the mental lexicon (ML) further. Another key contribution of this thesis is the development of a full pipeline for large scale collocation network generation that can be used by other researchers, including a thorough explanation of graph theoretical concepts to a linguistic audience paired with an in-depth analysis of the suitability of existing approaches to Association Measure calculation to ascribe the identified collocations a perceptual reality. The findings reveal that combinations of association measures (corpus linguistic approach), particularly log Dice, LL, and $\chi 2$, provide the best approximation of word association networks (psycholinguistic evidence), though systematic discrepancies remain. Additionally, word association networks are more tightly knit and generally strongly connected when compared to the more specialised and fragmented nature of collocation networks.

Table of Contents

List of Tables

List of Figures

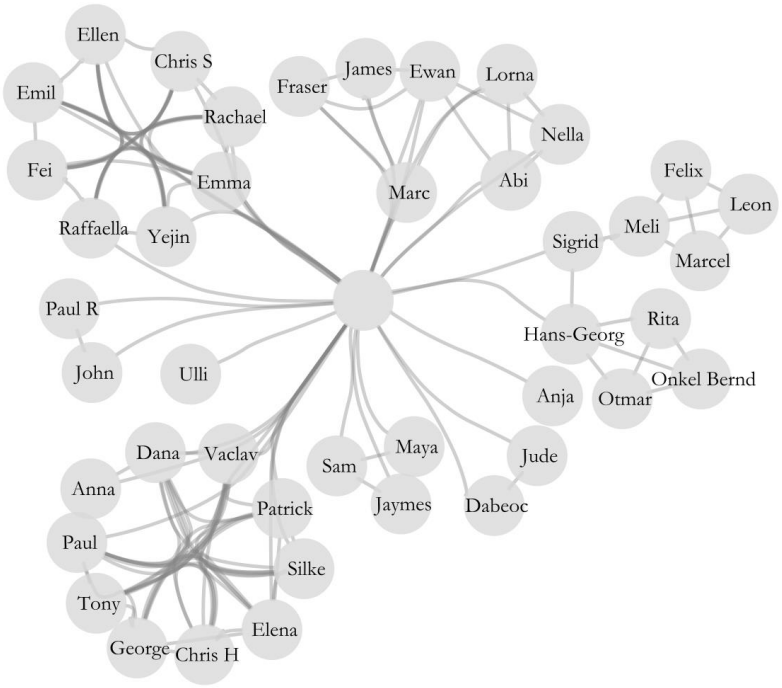
On a larger scale, the results of this research further show a prevalence of value judgements in clusters containing otherwise very concrete terms as illustrated by the *phone/touch* cluster (see Figure 37). This strengthens the argument that usage-based linguistic datasets are strongly impacted by emotional and affective relations which are universally present when communicating or during the

word association task (Kempe et al., 2013; Out et al., 2020; Sereno et al., 2015) and encourages research expanding research expanding Kousta et al. 's (2011) findings that systematic mental and linguistic processing differences emerge based on whether emotionally charged or neutral concepts are processed. The present data indicates that this could affect even seemingly value neutral lexical items such as *phone* as exemplified in Chapter 4.3.1. 201

List of Equations

Acknowledgements

As this thesis will go into great detail to explain, no word is an island – and the context is what gives meaning. I firmly believe that no human is an island either and our context is what gives us our meaning. I am incredibly grateful for having enjoyed the privilege of being surrounded by such a wonderful group of people who made it possible for me to be the person I am, and for this thesis to come into being.



Firstly, and most importantly, I'd like to thank my family: Onkel Bernd, my grandparents Rita and Otmar, my sister Melanie and her family, and especially my parents Sigrid and Hans-Georg, for always supporting my strange, wonky, and unpredictable life choices and for valuing the things I do. Without you two, I wouldn't be at all, and without you being the wise, hard-working, supportive, and kind parents you are I certainly could have never become the person I am now.

I would also like to thank Maya, for all the warmth, generosity, patience, and her infinite understanding for me vanishing into my books and code far too often over the past years. Sam, thank you for all the discussions, big and small, for sharing the good times, the bad times, and, most importantly, your life with me.

Jude, thank you for supporting me, wholeheartedly, no matter what life has in store for either of us. Thank you for always being dependable and trustworthy, and for putting moral integrity above all else. Ulli, thank you for believing in me from day one, and for being there for me no matter how far away from Münnerstadt my journeys take me. Dabeoc, thank you for so many hours of shared teas, recovering from PhD life together, and sharing your gift for storytelling with me. Anja, thank you for nine years of friendship, the sunshine you radiate wherever you go, for never missing a visit, and for always picking things up right where we left off. Jaymes, thank you for trusting me, and letting me be a part of your life – I'll always be amazed by your resilience and resourcefulness.

One of the greatest privileges of my academic life was being supervised by Professor Vaclav Brezina. I want to highlight not only the professional contributions he has made when discussing papers, organising fantastic research meetings, and reviewing my work, but also his personal and moral qualities. These qualities are, as is often the case in an academic setting, inextricably linked

to his work. Having hour-long discussions on topics less directly linked to the PhD project has enriched my life, and - in a climate where there is a dire lack of these figures - provided me with a role model. Vaclav, I admire your intellect, perseverance, and unwavering optimism, and I always walked away from our conversations a better person than I was going in.

I am also very grateful to my extended CASS family, who I have spent so many hours with since I was welcomed into the group in 2019. Special thanks to Chris Sanderson, Ellen Roberts, Emil Tangham, Rachael McCarthy, Luke Collins, Will Dance, Raffaella Bottini, Yejin Jung, Fei Zhu, Ruth Avon, and Emma Putland – it's been an absolute joy working with and alongside you, seeing you grow, succeed, show resilience, and celebrate wins. Chris, it has been an absolute joy to be your office neighbour, and I want to thank you for sharing your positive energy with all of us and for being the starlight of CASS.

I also want to thank the members of my (pre-/post-)confirmation panels Elena Semino, Tony McEnery, and Patrick Rebuschat. It's been a true privilege to share the earlier stages of my work with you. Thank you also to all the academics I have worked for and with alongside my PhD, especially Paul Rayson, Paul Baker, George Brown, Silke Brandt, Chris Hart, my internal examiner, Dana Gablasova, and my external examiner, Anna Siyanova-Chanturia, for being so open and generous with your time, and with sharing your expertise with me.

Lastly, I'd like to thank who I refer to as the *Glasgow bunch*: Marc Alexander, Lorna Hughes, James Balfour, Ewan Hannaford, Nella McNichol, Abi L Glen, and Fraser Dallachy. It has been an absolute joy working on DH topics and/or attending conferences with you throughout my PhD, and you have always offered me a fresh perspective on my research, and academia more generally.

1 Introduction

1.1 Introduction to Network Approaches in Linguistic Research

Perhaps the most common quote presented when dealing with research into collocations, "[y]ou shall know a word by the company it keeps !" (Firth, 1957, p. 11), might be slightly hackneyed, but it captures the essence of collocation research very effectively, even over half a century after the publication of Firth's seminal work. Two essential observations regarding collocations are conveyed here – firstly, meaning is contextual and thus emergent and dynamic (Ellis et al., 2009, pp. 108–109). Secondly, directly emerging from this, contextual relationships are circular since the contextual meaning of a word in use will set a precedent and affect its quality as a signifier for future events, see Figure 1. A language without collocations, a language without systematic recombination of individual elements to create larger meanings – is hardly imaginable.

Multi Word Expressions (MWEs) are defined as "(semi-)fixed, recurrent phrases" (Siyanova-Chanturia & Martinez, 2015, p. 549) or "lexical items which consist of more than one 'word' and have some kind of unitary meaning or pragmatic function" (Moon, 2015, p. 120) and cover concepts such as formulaic expressions, idioms, lexically determined combinations as well as compound nouns and prepositional verbs, and, most importantly for the present study, collocations (Evert, 2005, p. 337). Collocations are a fundamental concept in linguistic research as they represent building blocks of linguistic meaning both in terms of language production and perception as well as mental processing (Evert, 2005, p. 337). Existing research uses a plethora of different definitions depending on the respective research focus. Despite extensive efforts to examine the nature of processes underlying collocations both qualitatively and quantitatively over decades, sizeable research gaps still wait to be filled. This is particularly true for questions along the lines of which structures underlie collocational relationships, what mental representations exist thereof, if/how collocational associations can be modelled based on finite datasets, and if/how adding or removing elements from these representations impacts said structure. Considerations like these make collocational studies an ideal starting point for advancing linguistic theory on a larger scale (Barnbrook et al., 2013, p. 4). In more practical terms, collocations are also relevant since a variety of linguistic theories indicate that they, in their broadest definition as a commonly co-occurring group or set of words (Barnbrook et al., 2013, p. 3; Stulpinaitė et al., 2016, p. 31), are the basis for high language proficiency and fluency as they form essential units representing language through conventionalised and entrenched form-meaning mappings (Croft & Cruse, 2004, p. 292; Simpson-Vlach & Ellis, 2010, p. 488).

This thesis aims to address some of the key research gaps using a triangulation of corpus linguistics, psycholinguistics and graph theory. In order to do so, a large-scale analysis of computationally generated corpus-based collocation networks and psycholinguistic word association networks is carried out in this thesis. Word associations have been chosen as the basis for the psycholinguistic network since they portray the perceived relation between concepts via discrete linguistic units (Kang, 2018, p. 87), similarly to collocations. One special quality of the networks generated here is that they are, in contrast to much previous work e.g. on small-scale collocation networks à la GraphColl (Brezina et al., 2015, p. 139), based on the *entire* corpus or word association database. This means that they, quite literally, allow for an exploration of more than just the sum of the words they contain - they capture and can be used to analyse the rich contextual interlinking between words, concepts, and clusters, and provide insights into the structure of the corpus and the underlying language as a whole. Aside from new insights into current open questions regarding the structure and organisation of collocational knowledge, this approach could also provide new research prompts for investigating the internal structure of the mental lexicon (ML) further. This term is here defined as the way in which words are interlinked and stored in the human mind to facilitate efficient linguistic comprehension and production (Dasgupta et al., 2016, pp. 833–834). One prominent link between psycholinguistic networks and the structure of the ML and corpus-based approaches is Statistical Learning (SL). This important phenomenon in cognitive science can be broadly defined as "learning from the distributional properties of sensory input across time and space" (Frost et al., 2019, p. 1128). When applied to a linguistic context, learning individual patterns of linguistic elements is then based on factors such as frequency of co-occurrence, recency, distinctiveness, reliability etc. ((Ellis, 2006, 1,5f; Ellis & O'Donnell, 2014, p. 78), see Chapter 2.5.1 for a comprehensive description of this phenomenon). This is the very mechanism that drives the circular relationship between language perception and language production as depicted in Figure 1.

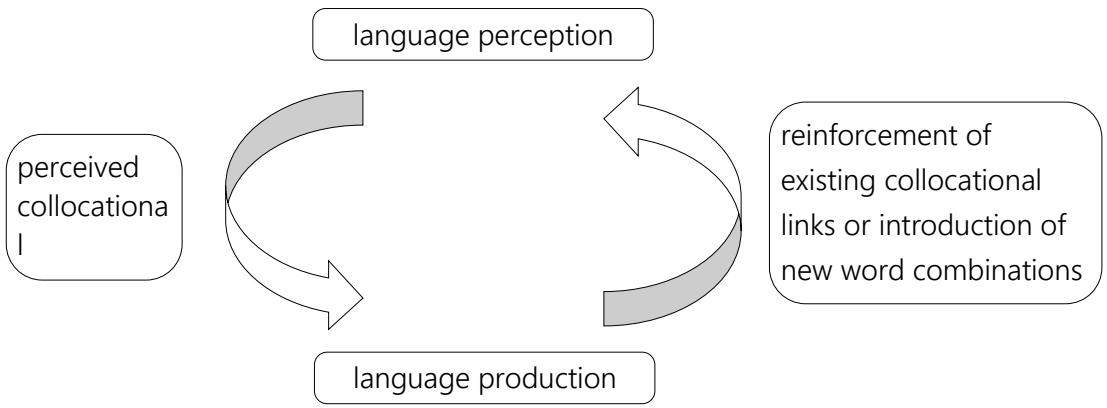


Figure 1: Diagram depicting the circular relationship of language production and language perception.

SL has been shown to facilitate complex learning tasks not only with respect to different linguistic skills that influence one another (Thiessen & Saffran, 2007, p. 97), but also across other domains such as visual learning (Rebuschat & Williams, 2012, p. 2). This facilitation is a result of multilevel learning, where one sub-process that is being acquired as part of learning process A is also relevant to learning process B. It is therefore assumed that learning linguistic skills impacts the construction and shape of the mental processes and is therefore highly likely to be a main factor for building and re-structuring the mental lexicon. One important point that represents an overlap in the fields of corpus linguistics and psycholinguistics is the fact that, just like the abovementioned learning processes, association scores obtained from corpora are also contingency-based (McConnell & Blumenthal-Dramé, 2019, p. 4); this allows for comparisons between these structures.

In terms of its theoretical positioning, this thesis rests on two main pillars: Cognitive Linguistics with a particular focus on usage-based principles and construction grammar, and – to a lesser degree – also Functional Linguistics. In short, this means the following: this thesis aims to follow the two key commitments of Cognitive Linguistics, the cognitive commitment and the generalization commitment (Lakoff, 1991, pp. 53–55). This will be achieved by aiming to ground and cross-verify all methodological decisions using current findings from psycholinguistics and neurolinguistics and by aiming to advance knowledge of underlying processes rather than small isolated linguistic phenomena respectively. At the same time this thesis is also rooted in usage-based theory since it builds on the idea that usage events shape linguistic knowledge (Kang, 2018, p. 85; Langacker, 1987) and in cognitive grammar (Langacker, 1986, 1999). This entails that conventionalised form-meaning patterns such as cue-associations or collocations are studied in their respective contextual use and with their communicative purpose in mind. Predictions and inferences, and thus co-occurrence frequencies via Statistical Learning, are crucial for communication since this is necessary for effective communication (Kapatsinski, 2014, p. 29). The theoretical foundation is motivated and explored at length in Chapter 2.5.1.

The primary motivation of this thesis is establishing a connection of current psycholinguistic evidence, i.e. experimental and theoretical findings regarding the structural design of the mental lexicon, to data-based findings from large-scale collocation networks. One contribution of this work therefore lies in the triangulation (Noble & Heale, 2019, p. 67) of three main theories presented above. Bridging gaps between these approaches to language and developing new viewpoints on the data might help overcome or seriously limit fundamental biases and present well-founded, balanced explanations on the basis of the present exploration. For this reason, Chapter 2.5 of this thesis is dedicated to exploring the status quo of psycholinguistic research with a particular focus on processes and parameters that have been shown to impact language learning,

the mental lexicon, and language production. These factors are then used as a baseline of assumptions that outline the edges of what is here termed *psycholinguistic plausibility*. Any approach to collocation extraction that can be aligned with the empirical findings outlined in the abovementioned Chapters is then considered *psycholinguistically plausible*, and thus considered a valid option for network comparisons. It is, of course, essential to note here that the status quo of research in this area is likely to change over the years, and the framework at hand, presented in the shape of a decision flowchart for selecting *psycholinguistically plausible* Association Measures (AMs; defined and discussed extensively in Chapter 3.2) has thus been developed with compositionality and flexibility in mind to allow for future amendments.

Another matter that needs explaining in order to set the scene for this thesis is the similarities and differences between the compared concepts at hand, namely word associations and collocations. Word associations are here seen as primarily representing the mental representation of language as opposed to collocations that are, in the case of the BNC 2014 used in this thesis, extracted from a corpus consisting of ultimately communicative language. It is crucial to acknowledge that neither word association data nor corpus data purely captures linguistic relations. Both data types are also influenced by emotional and affective relations (Kempe et al., 2013; Out et al., 2020; Sereno et al., 2015) since non-linguistic experiences are expected to affect participants during the task or when communicating. Despite the fact that this noise is present in both datasets, it cannot reasonably be assumed that the effects are similar and would thus relativise each other rather than causing additive, more disruptive interference. Theoretically, the two concepts are therefore similar, but in practice far from being identical. This issue is, naturally, not the only limitation; a range of further limitations applicable to this project are explored in Chapter 5.6. Addressing most of the limitations would require a number of multi-lingual datasets, a significant amount of compute power, and full control of experimental setups and emotion states, sociolinguistic metadata, EEG-data, as well as longitudinal data for thousands of participants; a feat that cannot be achieved by any individual thesis. Therefore, this work should be read as an initial exploration into the field of large-scale linguistic networks and comparative linguistic network analysis, and act as a guide providing resources and starting points for future research rather than a finished, static, and all-encompassing research output.

1.2 Research Aims

After a brief examination of the research landscape surrounding collocations and word associations, specific research aims are presented here. There are three main aims this thesis is working toward: the first aim, which lies at the heart of the project, is methodological innovation. More specifically, the thesis aims to develop a novel approach to generating large linguistic

networks. The second aim, which partially underpins the first, is a critical evaluation of current practices surrounding collocation extraction and the generalisability of findings from collocation studies. Lastly, the third aim is examining the generalisability of collocations to mental association via a contrastive analysis of a large word association network with holistic collocation networks.

In the context of furthering methodological advancement, the following aspects are key: firstly, this study integrates three disciplines: corpus linguistics, psycholinguistics, and graph theory[1]. A triangulation of these approaches allows for the creation of a novel, integrated framework for analysing linguistic patterns and structures and aims to narrow the existing gap between psycholinguistics and corpus linguistics. While corpus linguistic research incorporating psycholinguistic research is generally sparse with Deshors and Gries (2022), Durrant and Siyanova-Chanturia (2015), and Gries (2012, p. 47) being notable exceptions, there is significant scope for synergies arising from combining corpus approaches to psycholinguistic areas of interest. Exemplary for this are Statistical Learning (see Chapter 2.5.1) and retrieval processes in the Mental Lexicon (see Chapter 2.5.3).

Additionally, this project entails the development of an open source, fully interpretable and adaptable pipeline to generate and contrast large linguistic networks. This pipeline streamlines data processing and analysis and allows researchers to fully customise the selection of collocation extraction parameters and other methodological considerations that service the basis of network generation. This is essential to allow for tuning the methodological approach to the respective research question the evaluation of different methodological approaches as well as ensuring reproducibility.

It is essential to note that the methodological development undertaken as part of this thesis is in no way intended to make existing methodologies redundant but rather constitutes a further addition to the methodological toolkit of corpus linguistics. Providing a fully customizable open-source pipeline further emphasises research sustainability by providing a new avenue to leverage these existing datasets. This is particularly relevant given the background of the high cost involved in generating large-scale balanced corpora such as the BNC 2014 (Brezina et al., 2021; Love et al., 2017) or collecting psycholinguistic data e.g. for the SWOW (Deyne et al., 2019, p. 987) word association database.

Given the widespread use of different association measures, it is essential to examine their reliability and validity. The second objective therefore is a critical evaluation of current practices surrounding

[1] While graph theory is often referred to as a theory or framework, it is also widely recognised as a discipline in its own right within the field of mathematics, see Bondy and Murty (2010); Gross et al. (2019) for further context and terminology.

the employment and interpretation of association measure. This is necessary since the generated networks are supposed to represent human linguistic knowledge in a psycholinguistically plausible manner (in line with the cognitive commitment (Lakoff, 1991, pp. 53–55), see Chapter 2.3). The concept of psycholinguistic plausibility (further explored in Chapter 2.5) is defined as follows: an approach is psycholinguistically plausible if it can be aligned with current theories and experimental findings from psycholinguistics, such as reading times, cue responses, Statistical Learning, the mental lexicon etc. A critical evaluation taking into consideration the psycholinguistic plausibility of individual association measures serves as the natural starting point for developing a new methodology that builds on collocation extraction. Psycholinguistic plausibility is generally not considered in previous corpus linguistics literature; properly constructing a psycholinguistically plausible collocation network bottom-up therefore requires re-thinking and re-evaluating a large number of standard practices in the field. Examples for methodological steps that need to be thoroughly assessed are the selection of a suitable unit of analysis, directionality, window spans, as well as the selection of specific AMs due to the underlying assumptions they are built on (e.g. MI scores assuming that language is random as a base for identifying collocations, an assumption which is known to be false). All of these factors greatly influence the shape of collocations, but the underlying assumptions as well as their use in combination are seldom meticulously discussed and motivated in corpus linguistic research (Gries, 2012, pp. 47–48).

Putting the methodological innovation and evaluation of existing approaches to collocation extraction into practice, the third aim is to systematically evaluate the differences between a large word association network, here SWOW-EN and subsections thereof, with holistic collocation networks based on the BNC 2014. This evaluation is necessary since collocation identification is often used with the ultimate aim of generalising to stance and attitude, e.g. in corpus-assisted discourse analysis (S. Chen, 2013; Galasinski & Marley, 1998) without a thorough evaluation as to how directly repeated textual co-occurrence influences listeners'/readers' mental processing.

After having outlined the aims of this thesis, namely the evaluation of existing collocation extraction approaches, the development of a new method which enables the generation and comparison of large linguistic networks, and the application of this method, it is also essential to consider a key limitation of this approach. The evaluative component rests on the foundation of experimental knowledge since this constitutes the benchmark for psycholinguistic plausibility. All experimental knowledge, particularly in fields adjacent to neuroscience and psychology, can only ever be regarded as the best evidence currently available rather than an absolute, irrevocable truth. In consequence, existing theories and psycholinguistic findings that inform the present approach also simply hold the status of being not currently disproved, and adaptations to the model are

possible and necessary should the evidence shift in favour of alternative theories of mind and linguistic processing (cf. McEnery and Brezina (2022) for epistemological grounding of empirical statements in language analysis). Despite the uncertainties associated with relying on any empirical knowledge, the pursuit of this research remains worthwhile since the methodological decisions underpinning the proposed network generation pipeline can be changed dynamically given the emergence of new psycholinguistic findings and alternative theories.

The specific operationalised questions used to achieve the aims of this thesis are described in Chapter 2.9, the final chapter of the Literature Review which outlines the current state-of-the-art approach to collocation extraction as well as psycholinguistic findings in greater detail.

1.3 Organisation of the Thesis

This Chapter provides a brief roadmap detailing the organisation of this thesis and motivating the chosen structure. First, Chapter 1 introduces the linguistic grounding of Large Linguistic Networks and sets the stage for investigating collocations and psycholinguistic data as the basis for these networks. It further formulates the aims of this thesis: methodological advancement in the shape of introducing a new pipeline to generate and analyse Large Linguistic Networks, the necessary underlying evaluation of current practices in the field, and a triangulation of corpus linguistics, psycholinguistics, and graph theory. Chapter 2 reviews the existing literature with a particular focus on network approaches and explores the theoretical underpinning of this thesis which informs the research questions and methodological decisions. In order to achieve this, construction grammar, usage-based approaches, and functional linguistics are introduced in Chapter 2.3. Following this, Chapter 2.4 outlines the diverging definitions of collocation and provides an overview of the types of collocation which serves as a partial theoretical basis for later association measure evaluation. Chapter 2.5 introduces experimental research which on which the later defined concept of psycholinguistic plausibility is based via three major sections: Research on language learning processes, especially Statistical Learning, research on linguistic memory and the Mental Lexicon, as well as research on retrieval processes for language production. Lastly, the final component required to introduce a new methodology to display Large Linguistic Networks, graph theory, is introduced in Chapter 2.7. Special emphasis lies on the presentation of graph theoretical concepts which can be harnessed for linguistic purposes and their discussion in a manner accessible to a linguistic audience. Chapter 3 entails the evaluation of existing approaches to collocation extraction in terms of their capacity to be aligned with the findings presented in Chapter 2.5 including an evaluation of fundamental differences and inconsistencies relating to the generation of contingency tables across the field in Chapter 3.2.1 and a discussion of twenty different pre-existing collocation statistics in Chapter 3.2.3. Due to the identified inconsistencies, the pipeline for network generation

developed for this thesis does not rely on the output from any pre-existing corpus software and has been written from scratch. Chapter 3.4 concludes this section of the thesis with a flowchart for the AM selection process detailing which approaches to collocation extraction can be considered psycholinguistically plausible. The last major component of this thesis is the empirical evaluation of the proposed pipeline and a discussion of similarities and differences emerging from corpus-based collocation networks and psycholinguistic word association networks in Chapter 4. Chapter 4.2.1 outlines how the corpus, here the BNC 2014 (Brezina et al., 2021; Love et al., 2017), was chosen, pre-processed, tagged, and ingested into the LLN pipeline. Chapter 4.2.2 mirrors this approach for the English component of the word association dataset SWOW (Deyne et al., 2019, p. 987). The results of the comparative evaluation of word associations and collocations carried out via LLN are presented in 4.3 and structured into micro-, meso-, and macro level analyses.

Chapter 5 entails an interpretation of the findings as well as the acknowledgments of the limitations of this approach. Taking this into account, Chapter 5.7 then discusses practical applications of the LLN method in a number of subdomains of linguistics and related disciplines. Lastly, Chapter 6 presents a reflection on the journey towards generating large linguistic networks and provides an outlook for future research and application of LLN in the evolving wider field of corpus statistics.

2 Literature Review

Foundations: Linguistic Theory, Collocations, Psycholinguistic Evidence, and Graph Theory

2.1 Research Foundation

The first part of the thesis is a critical literature review spanning all topic areas that are relevant to research into large-scale corpus-based and psycholinguistic collocation networks. The aim of this Chapter is not only to provide an insight into the linguistic theory that the methodology is built on alongside the base definitions of key concepts like collocations, the mental lexicon and network approaches, but also to result in a methodological unification between the wider areas of psycholinguistic research, corpus linguistics and graph theory. A closer methodological connection between these specialist areas holds considerable potential to improve and substantiate existing knowledge about a variety of language-related mental processes, such as language learning and information storage and retrieval in the mental lexicon.

After a look at network representations in Chapter 2.2 their general utility is explored and two subtypes which are integral to this thesis: collocation networks and word association networks are discussed. In order to position this research in the wider field and to explain the foundations and some of the underlying assumptions of research questions raised here, the two main theoretical frameworks relevant for this thesis are then presented and discussed in Chapter 2.3. Firstly, construction grammar and the tradition of usage-based (Gries & Ellis, 2015, p. 229) approaches to cognitive linguistics (Lakoff, 1991, pp. 53–55; Langacker, 2008) are introduced. Core principles of functional linguistics (Halliday & Matthiessen, 2004, p. 46; Hasan, 2009, pp. 309–310) are then explored and aligned with the aims of this thesis. In Chapter 2.4 the different linguistic definitions of collocation are presented along with a classification of these phenomena into different sub-types. In the next step, brief descriptions of different applications of collocation-based research are provided for two main reasons: Firstly, this thesis is in its essence methodologically driven and grounded in theory. Secondly, gaining an understanding of what the challenges and merits of each of the applied research areas is essential as the foundation for building a useable tool that has the potential to enrich these fields. The developed methods include a range of tools for sentence-wide collocation extraction and are specifically designed for replication by other researchers; Appendix A provides adaptable interactive code to facilitate this. After this, several relevant psycholinguistic concepts are explored in greater detail as part of Chapter 2.5. The mental lexicon as defined in psycholinguistic research is then formally introduced, followed by a detailed account of research relating to language learning processes with a focus on Statistical Learning (Ellis, 2006, p. 1; Ellis & O'Donnell, 2012, p. 265), linguistic memory, and retrieval processes in the mental lexicon. The empirical evidence on which the statements made in Chapter 2.5 are built represent one of the key avenues for linking corpus linguistic and psycholinguistic research. Limitations and the extent of generalisability based on the status quo of current research are also outlined in this context. The

last part of the literature review is Chapter 2.7 which introduces graph theory with a special focus on its utility as a basis for network explorations and presents three layers of linguistically relevant measures. On the micro-level, measures regarding individual linguistic units are explored, the meso-level focuses on clusters of interconnected linguistic units, and the macro-level allows for exploring the holistic structure of the dataset at hand. Finally, Chapter 2.8 acts as a summary and conclusion before Chapter 2.9 explains how the identified gaps in the literature lead to the research questions aimed to be answered by this thesis.

2.2 Network Representations of Linguistic Data

In this Chapter, network representations of linguistic data are discussed, and their subtypes are explored alongside their potential for research applications. It is firstly important to mention that this thesis foregrounds specifically *linguistic* networks as opposed to somewhat more well-researched *social* networks generated on the basis of textual data (e.g. Moretti (2011) and Stiller et al. (2003) who carry out social network analyses on the works of Shakespeare) – on a superficial level the visualisation of these two types of networks might seem somewhat similar (see Figure 2). However, the key difference here is that the nodes in a linguistic network represent *linguistic elements* (i.e. words, syllables, phonemes etc.) and the edges *their linguistic relations* to one another (i.e. collocational relations, syntactic, semantic, translational etc.), whereas nodes in social networks represent *actors* (such as individuals on twitter, authors, characters in a play etc.) and the associated edges the *social relations* between them (i.e. twitter replies/shared threads, citations, occurrence in the same chapter/act etc.).

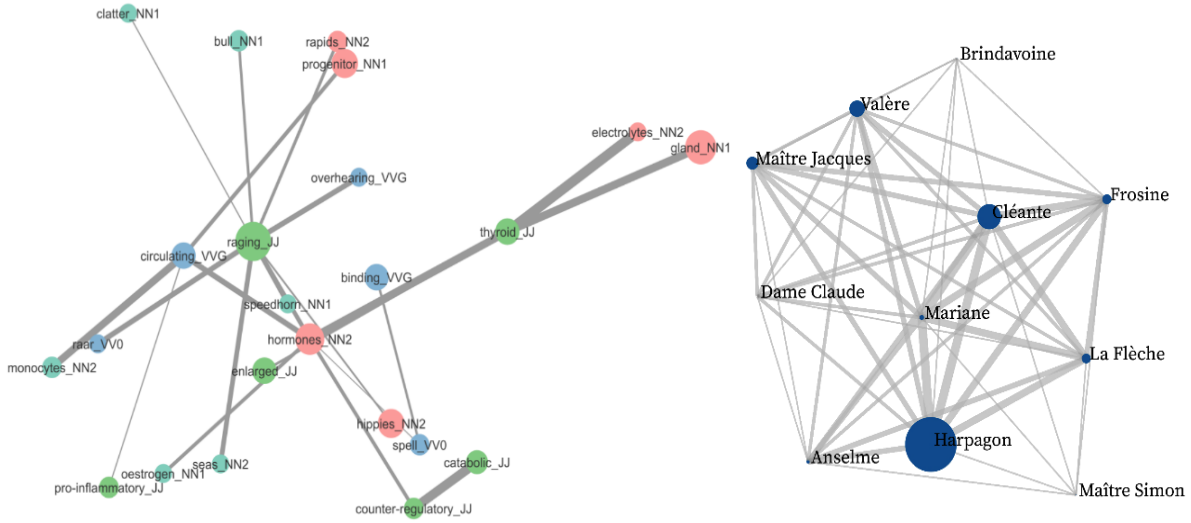


Figure 2: Linguistic network (left, here collocations surrounding the term *hormones* in the BNC 2014), and social network based on textual data (right, here representing co-occurring characters in the same act in Molière's *L'Avare*).

Linguistic networks, especially large-scale or corpus-wide networks, are less common in existing literature, but the utility of this approach to both a range of academic and non-academic fields is illustrated in the following chapter. Initially, general motivations for choosing a network angle on researching language is presented. These are then elucidated with more specific motivations for using collocation networks and word-association networks in particular. The three disciplines that are drawn on in this thesis are corpus linguistics, psycholinguistics, and graph theory. Similarly to corpus linguistics, where arguments can be made towards this field being methodological or theoretical in nature, network science is also seen to hold both theoretical and methodological potential (Vitevitch & Goldstein, 2014, p. 131).

2.2.1 General Utility of Network Approaches to Language

Firstly, exploring linguistic data using networks and graph theoretical methods allows for approaching research questions, particularly questions regarding underlying cognitive processes, from a new angle. The first immediate difference between network approaches and traditional analyses of concordance lines or collocation tables consists of the way a researcher interacts with the data – network approaches allow for looking at the whole dataset as one unit of interest and evaluating structures and patterns that emerge from this without a mandatory focus on an intuition-based starting point for the analysis (Castro & Siew, 2020; Jihua Dong & Buckingham, 2018, p. 120; Sinclair & Coulthard, 1975) – see the worked example in Chapter 2.2.2. Where conventional analyses will rely on a researcher's intuition as to which phenomenon or search term will be of interest, a network analysis can provide a data-driven starting point that helps minimise researcher bias. Secondly, network approaches allow for testing and refine existing hypotheses using a previously unexplored approach. Several studies have successfully proposed refined theories on the basis of network features; exemplary for this is Griffiths et al.'s (2007, p. 1073) study using data from fluency task experiments. Their work shows that fluency predictors based on measures calculated from network properties (in this case PageRank) outperform simple word frequency (Beckage & Colunga, 2016, pp. 15–16; Griffiths et al., 2007, p. 1073) – the property that has previously been considered most influential in determining fluency related to individual words and thus underlines the utility of network approaches. Studies such as Arbesman et al. (2010, p. 332) further demonstrate not only the utility of network approaches, but also their applications in contrastive and cross-language research specifically. In their work, the authors investigate semantic and phonological networks of English and Spanish and demonstrate powerfully how contrastive network analyses can provide immediate starting points for testable predictions, such as differences in retrieval speeds for high-degree nodes in languages with correlative semantic and phonological

networks (e.g., Spanish) when compared to high-degree nodes in languages without said correlation (e.g., English).

The application of graph theoretical methods has furthermore been demonstrated to be relatively robust even when applied to data of poor quality (non-normalised online data containing emoji, misspellings etc.), see Veremyev et al. (2019) for an extended discussion. In addition to this, another strength of network representations lies in their ability to represent and quantify complex and asymmetric linguistic relationships in different ways, a property that is essential to linguistic processes. This is particularly important since it captures the linear nature of language and helps highlight differences between bidirectional and one-sided word co-occurrences. Large-scale network approaches also have the potential to – at least partially – capture broader contextual effects through accounting for all connections exhibited by individual words with every other word in the dataset. In this sense, networks can be seen as a natural choice for analysing and representing linguistic elements in a more cognitively plausible way than as individual, separate elements.

2.2.2 Collocation Networks

In a corpus linguistic context, collocation networks are of particular interest. Collocation networks are henceforth defined as networks generated on the basis of association measures where the nodes represent collocates and the edges represent their relations to each other as retrieved from one or more AMs (Brezina, 2018, p. 60). Collocation networks allow for dense information access in CDA contexts (Brezina, 2016, p. 106; Brezina et al., 2015, p. 164); their primary purpose is to enable an insight into the design of underlying structures of a greater number of connected linguistic units in use as opposed to looking at pure statistics for relatively isolated nodes and collocates (Baker, 2016, p. 161).

Network approaches thus generally allow for expanding the viewpoint of a researcher beyond the numerical values gained from AMs and the frequencies of collocations in tables alone – which are the standard in CL (Gries & Ellis, 2015, p. 231) – to a perspective that also incorporates other collocates of a node word and allows for exploring shared collocates. All of these features already enable a richer analyses than an exploration of the numbers alone and they can help to identify concepts for example in a quantitative exploration of the aboutness of a given word (Baker, 2016, p. 161; Brezina, 2016, p. 90).

Collocation networks can be broadly divided into two categories: The first type is small networks that are based on individual nodes of interest and their first to n-th order collocates (Brezina, 2016, p. 90). The second, less conventional type is large-scale networks that aim to represent a whole body of text and do not have a single focus point; these networks also allow for more extensive

graph-theoretical explorations and comparisons with networks based on phenomena observed in other disciplines. A part of this thesis is the development of a toolkit to generate such networks, as well as an extensive evaluation of capabilities and limitations.

While the first type of networks provides only very limited insights into larger-scale phenomena such as most grammatical patterns, multi word units or previously unknown conceptual metaphors, it also exhibits a series of benefits when compared to larger-scale networks. One of these strengths is the reduced computational cost when generating small scale networks. This is the case since a much smaller number of nodes and edges need to be computed due to the exclusive focus on direct collocates of a pre-identified search term. Another benefit lies in the greatly improved instant readability and often also interpretability due to the absence of surrounding noise caused by neighbouring but unrelated nodes that occur in larger scale networks. These factors make small-scale networks ideal for investigating specific lexical items that a researcher has already identified as relevant either by looking at the networks surrounding the respective nodes or by investigating overlaps of shared collocates (Brezina, 2018, p. 80). One publicly available and easily accessible tool to view small-scale collocational relationships already exists in the form of the GraphColl feature in #LancsBox / #LancsBox X (Brezina & Platt, 2024; Brezina et al., 2020). This software tool allows for the selection of different association measure parameters, window spans (based on word positions surrounding the node) and thresholds, and retains information on the type, lemma and POS level before plotting the first order collocates of a given search term. The main insight that is to be gained from using this methodology is what a 'rough sketch of the lexicogrammar of the word' might look like (McEnery & Brezina, 2019, p. 104). Figure 3 shows a model output generated using GraphColl in #LancsBox X.

A major contribution of this thesis is the systematic large-scale exploration of the second, non-localised type of networks. Despite larger-scale networks being sparse in previous literature, the concept of using graph theory to study aspects of language has been employed before in some restricted contexts. Network science as a methodology has been introduced to linguistics starting in the early 2000s; for a review of network approaches to language in the widest sense in that time span see Mehler (2008, pp. 349–350). These approaches were, however, mostly web-graphs displaying connections between different entries such as Wikipedia articles, newspaper articles, citations, thesaurus entries or hyperlinks – in short, relationships between linguistic meta-information rather than relationships on the textual level itself as visualised in Figure 2. The only early publication on collocation graphs to the knowledge of the author at time of writing, and a

more detailed, statistically founded description of the organisation of language using network approaches is Ferrer-i-Cancho and Solé (2001, p. 2263) and Dorogovtsev and Mendes (2001).

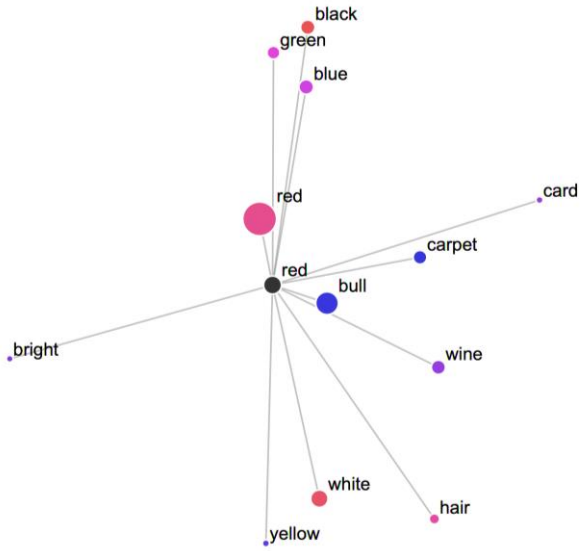


Figure 3: First-order collocates of *red* in the BNC 2014[2] as displayed using the GraphColl functionality of #LancsBox X.

Beyond this, some more comprehensive projects have used graph theory as a means to analyse other types of linguistic data, such as cue-response pairs (Deyne et al., 2019, pp. 998–999), orthographic networks (Trautwein & Schroeder, 2018, p. 12), and phonological networks (Vitevitch, 2008). In their work, Ferrer-i-Cancho and Solé (2001, p. 2263) have used a subset of the BNC 1994 as a basis for an investigation of graph-theoretical properties of language. They report that networks created on the basis of word co-occurrences in a window span of L0-R2 that fulfil the criterion of $p_1p_2 < p_{12}$ exhibit small world and scale free properties and therefore resemble networks found in non-linguistic domains (ibid.). In-depth explorations of smaller-scale properties such as network structures, clusters and nodes fulfilling a special role in the networks, as well as comparisons between networks based on different, more refined Association Measures than $p_1p_2 < p_{12}$ are presented as part of this thesis. A further contribution of the work undertaken here is the replicability and adaptability of the network generation resources provided here as opposed to rigid, one-off approaches in previous literature.

When comparing this approach to the existing GraphColl visualisation two main differences emerge: first and foremost, the methodology presented in this thesis aims to incorporate graph theoretical measures – not merely a new type of visualisation - into the corpus linguist's toolbox. This is essential because a wide variety of graph theoretical parameters can help analyse and explore

[2] eCPN: word(lowercase), log Dice, 9, L10-R10,C5-NC1 | node colour: word frequency (blue: max), size: collocational frequency, undirected, no filter, static, x-axis orientation based on word-order.

existing corpora data in new ways without the need for further data collection. The second core difference is the extent of the visualised material itself. While GraphColl focuses on predefined search terms and their first-, second-, etc. order collocates (and does not by default display the interrelations among the collocates in the graph, i.e. there are no edges between any two collocates of *red* plotted in Figure 3), large linguistic networks allow for visualising the entire collocational space in one dynamic graph. Since the distances are also, unlike in GraphColl, determined by the interrelations of all collocational pairs to one another (and not just the n-th order collocates), this visualisation holds a potential for exploring centrality, key nodes, clusters etc. A full overview of the features that can be used for analysis is provided in Chapter 2.7. Figure 4 displays the roughly equivalent output based on the LLN pipeline developed for this thesis, showing the impact interrelations between the selected nodes have on the overall information value of the network. It is essential to note here that the metrics used to obtain are not fully identical since #LancsBox X does not implement sentence-spans as a span option and the frequency cut-off for the LLN network is based on relative frequencies. The graphs illustrate how inter-collocate connections re-frame the collocational space.

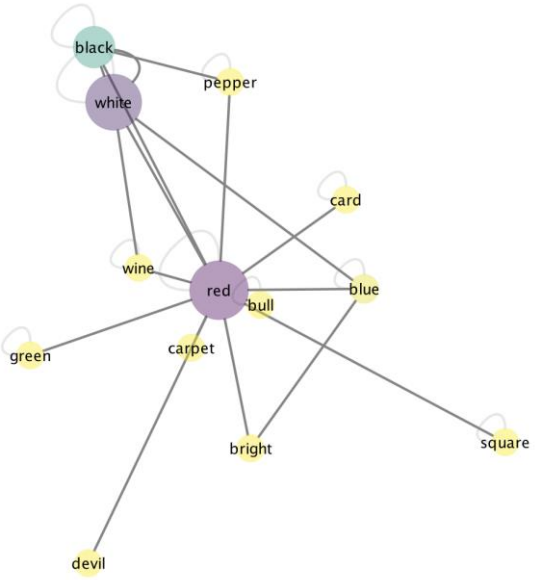


Figure 4: First-order collocates of *red* in the BNC 2014[3] as obtained from LLN.

In a similar vein, large-scale collocation networks facilitate dynamic explorations of a micro-, meso- and macro-level of the networks themselves. In practice, this means that word nodes with specific graph theoretical properties can be analysed on at least three levels: as the individual node with its centrality measures, clustering coefficients etc.; as the central node surrounded by its first order

[3] eCPN: lemma, log Dice, 5.42, sentence-span, 1 per million words, 1 | node colour & size: betweenness centrality (purple: max), directional, no filter, static, Kamada-Kawai layout.

collocations (whose positions, unlike in a small-scale network, are also determined by all other nodes in the network), and as part of the entire network, see Figure 5.

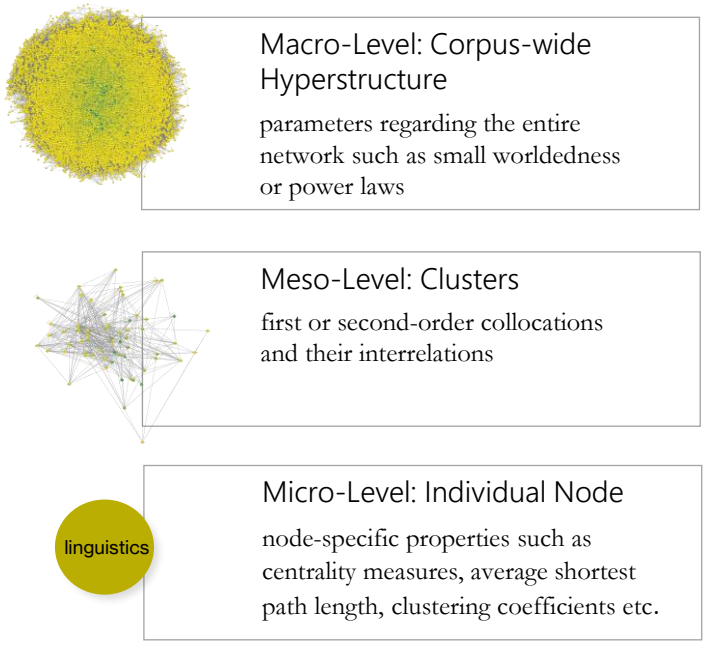


Figure 5: Illustration of possible levels of analysis for large-scale linguistic networks.

Having discussed how the visualisation options differ, this section briefly touches on the utility of graph theoretical features to linguistic research. Although a comprehensive discussion of graph theoretical features is reserved for Chapter 2.7, a cursory glance at the new opportunities provided by corpus-wide collocation networks reveals the following: It is now possible to assess the overall connectivity and size of the network containing all collocations. Since connections between *all* collocations (as identified through an AM of the researcher's choice) are plotted it becomes immediately obvious if high-scoring collocations or collocations of special interest belong to a large, connected component which anchors them in the wider context of the given corpus or if they only exist in an isolated, smaller component. Exemplary data from the Lone Wolf project (Schmück & Malone, 2023) for example, show that amongst the highest scoring collocations in the Lone Wolf Corpus (LL, peak, bigram, POSfiltered) are *am_N,gmt_N* (LL = 21,608) and *pm_N,gmt_N* (LL = 19,670) – these are timestamps and it becomes apparent that they are in fact isolated from all other collocational relationships in this subcorpus. Centrality information is particularly valuable in a CDA context as an aid to researchers for distinguishing discourse-central terms and topics of interest from disconnected outliers. Interestingly, beyond providing empirical evidence, graph theoretic analyses can also enrich linguistic study on a deeply theoretic level: non-trivial shared properties emerge from a large number of linguistic networks as Baronchelli et al. (2013, p. 352) have found when examining networks based on different languages. Due to the

nature of these observations, it can be speculated that not a universal grammar, but rather a limitation of human brainpower leads to systematic grammatical constraints (ibid.).

2.2.3 Psycholinguistic networks

In addition to co-occurrence networks and corpus-based approaches, network representations of psycholinguistic data are also highly relevant to this thesis. Benefits of applying graph theory and structural analyses to psycholinguistic data are largely identical to the benefits explored above and thus not reiterated at length. A specific benefit, however, is the fact that a vast subfield of psycholinguistics aims to explore, navigate, and provide insights into the structure of the Mental Lexicon – a task networks lend themselves to particularly easily.

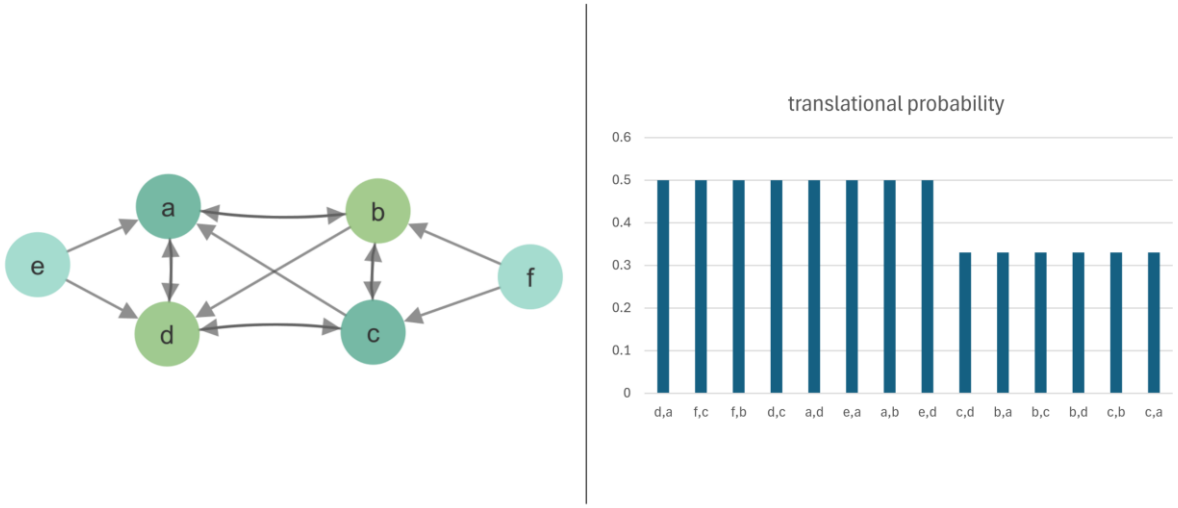


Figure 6: Illustration of the utility of network visualisation in linguistic research. Network visualisation (left) and pairwise statistical information regarding translational probabilities (right). Figure simplified and adapted from Karuza et al. (2016, p. 635).

Figure 6 clearly demonstrates the benefit of using these methods to study the structure of the Mental Lexicon. In this graph, the sequence presented at the top stems from a classical translational probability experiment where a participant is shown this sequence of syllables. The bar chart below this shows a visualisation of the bigram frequencies obtained from the syllable-sequence: While the chart appears symmetric, it fails to capture the complex underlying structure that only emerges from network representation.

On a general level, a large body of word association studies follows the distributional approach which builds on the hypothesis that co-occurring collexemes and constructions form a network in our mental grammar (A. C.-H. Chen, 2022, p. 212). Network approaches and graph theory enable the visualization of the entire network of associations, highlighting clusters of strongly associated words and identifying central or peripheral words within the network. The interrelations observed

in psycholinguistic networks can directly represent associative paths and form a part of the Mental Lexicon. Network approaches in this domain have therefore been presented as a remedy to the stagnation regarding advances in word association research (Fitzpatrick & Thwaites, 2020)., especially since networks have been shown to reveal structural properties of associative networks and the special role of individual lexical items (ibid.). For instance high degree, i.e., a large number of neighbours attached to a word, can be easily identified using graph theory and has been described as a marker for lexical availability (Ferrer-i-Cancho & Solé, 2001, p. 2264). Research further suggests that larger-scale processes such as search and retrieval are chiefly affected by the way information flows (Borge-Holthoefer & Arenas, 2010, pp. 1290–1291), a process which can be mapped and simulated using holistic word association networks. By leveraging these tools, researchers can gain a deeper understanding of the complex, multi-dimensional nature of word associations, thereby enriching psycholinguistic studies. In contrast, an alternative approach exploring associations in a word association database via search terms of individual, pre-defined words overlooks these observations does not offer any insights regarding information flow, overall connectivity, density of the whole association space, etc. and would only allow for a snapshot of a single associative domain.

2.3 Linguistic Framework and Underlying Theories: Reviewing Theoretical Approaches to Semantic Representation

This thesis is, whilst inter-disciplinary, first and foremost firmly rooted in linguistics and it rests on two theoretical pillars: Cognitive Linguistics with a particular focus on usage-based principles and construction grammar, and – to a lesser degree – also Functional Linguistics. These linguistic theories are often seen as discrete and conflicting, but they show significant overlap in certain applied areas of study such as corpus linguistics. It is perhaps surprising to find two large theoretical schools of thought, functionalist and cognitive approaches, listed side by side here and the following chapter therefore explores how these intersect and why the present thesis lies exactly in this space.

2.3.1 Cognitive Linguistics

The first theory to be explored here is a major, overarching one: Cognitive linguistics. Both corpus linguistics and the behavioural methods – to which the experimental setup used to obtain the cue-association dataset used in this thesis belongs – are part of the cognitive toolkit (Evans, 2019, 47–48). The field of cognitive linguistics emerged via the cognitive revolution in the past decades and is more influential now than most structuralist and functionalist approaches of studying language and the mind. Generally speaking, cognitive linguists hypothesise about the conceptual system

represented in language and the corresponding linguistic structures themselves (Evans, 2019, p. 15). The first individual papers recognising the potential of combining corpus methods and cognitive linguistics have been published over 20 years ago (see Schmid (2000) and Mukherjee (2004) and first advances towards a systematic use of corpus methodologies for cross-validating cognitive theories have been made in the early 2010s (Arppe et al., 2010). Nevertheless, and despite notable research such as that of Stefanowitsch and Gries (2005) and others in the Journal of Corpus Linguistics and Linguistic Theory which focus on the nexus between (cognitive) theory and corpus linguistics, the cognitive/corpus interdisciplinary field remains rather small in comparison to other fields such as discourse/corpus studies[4].

Researching in the tradition of cognitive linguistics means following its two key commitments: the cognitive commitment and the generalization commitment (Lakoff, 1991, pp. 53–55). In Lakoff's own words, the former binds researchers to "make one's account of human language accord with what's generally known about the mind and brain from disciplines other than linguistics" (ibid, p. 54). This key commitment is closely followed in this thesis since the aim is to explore properties of collocational structures not only from an evaluative and descriptive standpoint, but also to advance knowledge regarding underlying processes. Creating hypotheses on the basis of theories or mechanisms that do not conform with current experimental evidence such as findings emergent from the psycholinguistic dataset used in this study as well as a wide range of further psycholinguistic and neurolinguistic studies is therefore considered futile. This commitment also mandates continuously updating the models and methods presented here based on emergent knowledge whilst the project is ongoing.

The second tenet of Cognitive Linguistics, the generalisation commitment, advocates for a research focus on overarching principles that could potentially govern all facets of human language – this is also in line with the research goals present here for three main reasons:

Firstly, this thesis investigates Statistical Learning processes that have been found to influence all types of language learning and even non-linguistic learning processes in a very broad manner.

Secondly, this work posits that lexical and grammatical elements exist on a continuum, rather than being binary classifications. This perspective aligns seamlessly with the cognitive viewpoint on lexicogrammar (Berber Sardinha, 2020, p. 2; Herbst, 2018, p. 3; Langacker, 2008, p. 3). Lexicogrammar, in this context, is perceived as a spectrum that spans from closed (grammatical)

[4] This is, of course, a rough approximation here based on Google Scholar articles (1) and peer reviewed articles available via the Lancaster University library (2). The ratio appears to lie between 2:1 and 3:1 with 11,800 (1) | 33,719 (2) articles containing the terms "corpus linguistics" and "cognitive linguistics" published between 2000 and 2021 and 35,600 (1) | 60,384 (2) articles containing the terms "corpus linguistics" and "discourse".

systems to open (lexical) systems, with equal importance attributed to lexis and grammar. This interpretation, consistent with Sinclair and Carter's (2004, p. 164) critique of common interpretations of lexicogrammar in Systemic Functional Linguistics (SFL), does not merely represent an expanded grammar that includes lexical items in its most comprehensive form, but puts equal emphasis on lexical phenomena. In a similar vein, Construction Grammar also promotes the concept of morphosyntax rather than discrete manifestations of morphology and syntax (Haspelmath, 2011, p. 31). The prominence of this paradigm in CL can be attributed to the fact that analyses of texts often reveal a deep intertwining of lexis and grammar, making it impossible to cleanly separate them from semantics and usage patterns (Ellis et al., 2009, p. 90). Moreover, in practical terms CL by design requires clear and quantifiable rules regarding categories of different elements – this necessity commonly leads to critical discussions and evaluations of established rules and, in turn, to the establishment of more nuanced approaches that allow for regarding concepts as continuous rather than discrete.

Thirdly, network approaches allow the abstraction of findings based on corpus or association data such as the ones obtained in this thesis on a meta-level: Network-wide explorations make it possible to determine graph theoretical properties such as small-worldedness (see Chapter 2.7.3), density, centrality etc. or to model random walks over a network in order to simulate how individual components might be accessed. These processes can then be compared to other networks and possible generalisations can be found (e.g. between word association and semantic networks). This approach is therefore perfectly aligned with the generalisation commitment.

Lastly, the four E's of cognitive science (Embodiment, Embeddedness, Enactivism, and the Extended Mind (Rowlands, 2010; Ward & Stapleton, 2012)) play an important role in positioning the research carried out in this thesis. As can be seen in Figure 1, the present network exploration of communicative language and word associations aims to represent both language production and perception using large-scale datasets. It is, however, essential to consider that every utterance and every perception is shaped by the fact that language users are situated and limited by their own bodies, experience and interact with their surroundings, and inadvertently influence and are influenced by societal norms and rules. Practically speaking, attempts to model and quantitatively examine all of these complex dynamic processes using current resources and technology are futile. The four E's should - and will in this thesis - guide the interpretation of the results obtained and provide starting points for further research.

Usage-based theory

Within the wider field of cognitive linguistics, two more immediately tangible theories are of key importance to this thesis: Usage-based theory (Barlow & Kemmer, 2000; Kang, 2018, p. 85; Langacker, 1987) and cognitive grammar (Langacker, 1986, 1999), two concepts that are somewhat entangled since cognitive linguistics is usage-based by design (Tribushinina & Gillis, 2017, p. 14). Usage-based theory is key since the collocational and psycholinguistic elements to be examined here represent symbolic constructions (form-meaning mappings). According to usage-based theory, these mappings – which consist partly of collocations, but can also consist of other linguistic elements – are conventionalised in specific speech communities and ultimately serve as the basis for communication. In this sense, frequencies of word associations or collocations are assumed to represent mental semantic knowledge (Kang, 2018, p. 85), which makes them a good empirically measurable proxy for researching the mental lexicon itself. This theory furthermore foregrounds the role entrenchment plays in language practices, a notion that directly relates to collocational frequencies and their relationship to word combinations in the mental lexicon (Gries & Ellis, 2015, p. 229).

Construction Grammar

More specifically while looking at the immediate subject of this thesis, construction grammar comes into play. Construction grammar entails the notion that conventionalised form-meaning pairings are relevant not only in the context of individual words but also affect higher-level constructions such as idioms and phrases (Hoffmann & Trousdale, 2013, p. 1). The pairings are described as direct mappings – they do not make use of intermediate structures (Bybee, 2013, p. 51). In this sense, construction grammar also supports the notion that lexicon, morphology, and grammar are not strictly binary and therefore separable categories, they rather advocate for the notion of a lexico-grammatical continuum (Langacker, 2008, p. 15) which is also commonly adopted in CL (Berber Sardinha, 2020, p. 2). This directly ties in with viewing language as a network: the strength of association between individual nodes of this network dynamically highlights common patterns in language which can be expected to represent such form-meaning pairings. Constructionist approaches are furthermore built on the idea that predictions and inferences are crucial for communication and that, as a consequence of this, abstractions are also essential. These abstractions need to be processed using statistical knowledge in order to enable individuals to communicate effectively in the majority of cases (Kapatsinski, 2014, p. 29). Construction grammarians (Bybee, 2013, p. 49; Herbst, 2018, p. 2) posit the following three notions which are highly relevant to this thesis:

— The notion that lexical knowledge and grammatical knowledge cannot be separated into binary categories, it rather exists on a lexico-grammatical continuum

— The notion that linguistic knowledge is not biologically determined, i.e. inborn, but rather acquired through form meaning pairings (constructions) that are stored in networks.

— The notion that linguistic knowledge is emergent, i.e. ever-changing due to new situations of use and individual experiences (a concept also referred to as exemplar theory (Bybee, 2013, p. 50).

The latter point is especially relevant to the approach taken here since it constitutes the motivation to use corpora as a means to explore collocations and mental representations thereof. Balanced corpora such as the BNC 2014 can serve as a broadly indicative model showing which types of constructions speakers of a specific language, in this case British English, are likely to have encountered and how frequently these were experienced (Herbst, 2018, p. 6).

Construction grammar, like the abovementioned approaches, is of interest as a foundation of this study since it carries far-reaching implications regarding the structure of a mental grammar – and thus, since this is happening on a lexico-grammatical continuum, also the mental lexicon. This mental grammar is firstly described as a network composed of constructions, i.e. schematic constructions (Hoffmann & Trousdale, 2013, p. 3) which is compatible with a network-based approach to collocations. Construction grammar secondly entails the idea that the frequency of co-occurrence of certain linguistic elements impacts the way in which they are perceived, processed, chunked and stored (Ellis et al., 2009, p. 107). Following this line of thought further, certain types of collocations are then also expected to be processed as one meaning-carrying unit (i.e. one processing unit) rather than as the combination of its separate constituents. Investigations of different types of collocations and their interrelations within complex networks could therefore be used as indicators for a possible topography of both a mental grammar and a mental lexicon. While there is a large amount of evidence supporting the hypothesis that frequency effects greatly impact the long-term structure of the mental lexicon, no such consensus exists regarding the precise extent of this phenomenon. Durrant and Doherty (2010, p. 145), for instance, conducted an empirical study investigating the relationship between mental representations of collocations in native speakers and the frequency with which these collocations occur in the BNC 1994. Whilst confirming findings from Ellis et al. (2009, p. 107) studies indicating that frequency effects playing a key role for language learning, Durrant & Doherty further propose a psychological reality based on frequency effects might not apply to all types of word co-occurrences. The authors report mixed results: While they found that only associated word pairs, not high-frequency collocations impact

automatic priming effects, priming effects for both types of word co-occurrences have been identified in lexical decision tasks where participants might have employed higher-order processes as part of their decision-making processes. What is of relevance to the project at hand is that different types of collocations might be processed and stored differently in the mental lexicon and that it cannot be assumed that a one-size fits all approach to collocation extraction will provide uniform and reliable results. The present approach therefore aims to capture as many candidates for psycholinguistically valid collocations as possible and to establish meaningful filtering mechanisms for distinguishing between particular types of collocations later on.

Psycholinguistics

Lastly, Psycholinguistics, which is sometimes classified as a subdiscipline of cognitive psychology, investigates the acquisition, perception, and production of language via their contributing psychological and neurobiological mechanisms (Ellis, 2019, p. 40) is also briefly characterised here. This is necessary since a large portion of the data used in this thesis, the SWOW (Deyne et al., 2019, p. 987) dataset, is a result of psycholinguistic research. Psycholinguistics as a field of empirical study has a brief but rich history, spanning back to the late 19th century when Wilhelm Wundt (1897) first set out to examine the mind via language. Some highly influential paths that have been taken since were focused on the physical properties of the brain and processes on a neural level (e.g. Wernicke (1908) and Broca (1865)), while others aimed to shed light on the linguistic properties of our minds via behavioural observations. Piaget (1936), for example, made influential contributions to the field via early child language acquisition studies. Psycholinguistics has also undergone a cognitive revolution in the past two decades and therefore fits the cognitive principles laid out above. Both neurolinguistics and developmental psychology are large and influential fields today, but the present thesis exhibits a different focus: Using large amounts of empirical data as a lens through which linguistic connections and plausible processes in our mind can be studied.

2.3.2 Functional Linguistics

Perhaps surprisingly, the second cornerstone of this thesis is Functional Linguistics. This discipline offers a robust theoretical foundation for this project as it underscores the communicative role of language (van Valin, 2003, p. 320), along with systems of meaning and applicability. It perceives syntax and pragmatics as interconnected concepts (Newmeyer, 2010, p. 302), and places particular focus on the cyclical relationship delineated in Chapter 1.1: Langue (the collection of abstract rules of a signifying system that systematically shapes our perception of language) and parole (the tangible, individual linguistic elements expressed by individual speakers) are continuously informing and shaping each other (Hasan, 2009, pp. 309–310). The functionalist view on language

furthermore emphasises the communicative context and direct language in use when carrying out linguistic analyses, a principle that aligns perfectly with corpus linguistic methodologies. In this sense, corpus linguistics is closely connected to Functional Linguistics in that functional linguistics regards language to be governed by probabilistic features (Berber Sardinha, 2020, p. 2; Halliday & Webster, 2005, p. 67) which are often the subject of CL studies. There is furthermore also a less obvious, but nevertheless strong connection between Functionalism and Constructionism since 'Facts about the use of entire constructions, including register (e.g. formal or informal), dialect variation and so on, are stated as part of the construction as well.' (Goldberg, 2003, p. 221).

Drawing on Functional Linguistics ensures that language analysis takes into account the function it serves for the speaker; this is a crucial aspect of real-world communication. This thesis utilises corpus data that inherently displays a communicative intent – various objectives were intended to be achieved by interacting with a broad range of audiences across different genres of the BNC 2014. Recognising this diversity within the corpus is essential, and functional linguistics offers a robust framework for this.

It should also be highlighted that the dedication to following neurological findings and aiming to address general cognitive processes rather than specific ones does not conflict with the goal of exploring habitual and conventionalised form-meaning mappings. On the contrary: If applied systematically, functional approaches to linguistic entrenchment and the data collected and analysed in this tradition yield valuable new hypotheses that can serve as starting points for uncovering general and neurologically valid cognitive processes. This thesis seeks to facilitate this symbiosis through applying dynamic network approaches to semantic representation.

Examining the points of connection between Cognitive Linguistics and Functional Linguistics reveals a greater overlap between the disciplines than commonly acknowledged. The commitment to following neurological findings and aiming to address general cognitive processes over specific ones does not contradict the aim to explore habitual and conventionalised form-meaning mappings. On the contrary: Functional approaches to linguistic entrenchment and the data collected and analysed in this tradition provide valuable new hypotheses that can then be used as starting points for uncovering general and neurologically valid cognitive processes. This thesis seeks to facilitate this symbiosis via dynamic network approaches to semantic representation.

2.4 Collocations

One of the most central concepts to be defined and explored in this thesis is the notion of collocations; a wide range of occasionally vague definitions and types of collocations have been discussed in existing literature (Pecina, 2010, p. 141) and this Chapter provides an overview of the

different schools of thought around collocations. The existence of this great number of at times confusing, overlapping and mutually exclusive concepts that have been labelled "collocations" can be explained through the fact that collocations are epiphenomena. Specifically, this is to say that the co-occurrence of lexical items can have a wide variety of causes that cover a broad range of different areas of human communication and perception such as underlying semantic structures that facilitate certain word combinations, idioms, ideas about stereotypes etc. (Evert, 2008, p. 1218).

Different definitions and types of collocation are presented in the following Chapter in the hopes of disentangling the described melange of concepts as best as possible and creating a clear working definition of the concept of collocations for this thesis. Chapter 5.6 provides a comprehensive overview of limitations that apply to the approach taken here, including the role of language production and language perception when viewing language as a Complex Adaptive System, as well as a discussion of longitudinal studies, and a focus on languages other than English.

2.4.1 Definitions of Collocation

This Chapter explores different definitions of collocation that have been established in previous linguistic research. It is essential to communicate clearly and unambiguously what is meant by the central terms that are used when formulating a hypothesis; in the context of collocation research, this task is even more pertinent because of a high level of inconsistency in the existing definitions of collocations.

The rather complex task of clearly presenting and assessing different rather opaque and intertwined definitions of collocation is achieved here by moving from the broadest possible scope of the term 'collocation' to a gradually more specific scope applicable to psycholinguistic research. One of the most general definitions of collocation describes collocations as combinations of a wider variety of linguistic elements beyond the word-level, i.e. combinations of smaller individual morphemes or combinations of larger syntactic constructions (Evert, 2008, p. 1215). This definition is, however, used rather infrequently in existing research.

As mentioned in the introduction, the broadest and generally accepted definition of collocation characterises a collocation as a repeated or commonly co-occurring group or set of words (Brezina, 2018, p. 59; Stulpinaitė et al., 2016, p. 31). Barnbrook et al. (2013, p. 3) add the phrase "in their normal use" here which is relatively opaque; they do, however, also cite a narrower definition of collocation which entails that the term is used to "describe an aspect of language production in which pre-fabricated chunks of language are used to build up utterances" (ibid., p. 3). This specific approach already relies on several assumptions such as the immediate relevance of collocations for

language production and their storage in the mental lexicon as "pre-fabricated chunks." – These assumptions need to be further examined and critically evaluated since they define the limitations of all research built upon them.

Another quantitatively oriented definition of the term 'collocation' requires the frequency of co-occurrence of two words to be statistically significantly greater than the expected co-occurrence if there were no relationship between the constituent words (Messaoudi, 2019, p. 222). While this definition helps clarify the concept of a collocation and ensures a more objective approach, it also indirectly links the concept of a collocation to the size of the dataset at hand. This is the case since statistical significance indicates the amount of evidence against the null hypothesis. This assumption is problematic in that it violates the maxim of exactness that any definition should follow (Bickenbach et al., 1997, p. 102) – the theoretical idea of what a collocation is should not depend on the size of a given dataset. Corpus size significantly influences the practical identification of collocations. Consider a scenario where the objective is to extract collocations from a specific domain, such as academic spoken language. In such cases, the limited sample available for this subgenre may fall short of reaching the threshold for statistical significance. An additional challenge arises from the non-random distribution of words in academic speech, which contradicts a common null hypothesis in collocational research. However, it is essential to recognise that the absence of a comprehensive dataset encompassing the entirety of academic speech does not imply the absence of collocations within this context.

Lastly, this definition also explicitly excludes "negative collocations", i.e. words that systematically occur together less often than expected. Messaoudi (2019), however, also acknowledges that there might be different reasons leading to the emergence of collocations that might not be perfectly reflected in raw co-occurrence. Examples for this are linguistic factors such as grammatical constructions or sentence structure, on the one hand, and non-linguistic factors that are grounded in the reality that the words in question describe such as a "natural" connection between the word *food* and the word *eat* on the other hand. It is impossible to separate collocations that emerged from one of these factors from collocations that emerged from another or a combination thereof using the quantitative definition of collocation.

Beyond the definitions that have been described above, one further even more narrow interpretation of collocation that does not rely on statistical evaluations exists: collocations as non-modifiable, non-compositional, and non-substitutable units, i.e. units "whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components" (Choueka, 1988, p. 612).

Pecina (2010, p. 138) and Evert; (2005, p. 17) share this definition and remark that this unpredictability of meaning leads to the necessity to explicitly incorporate collocations into lexical collections. This might also lead to the hypothesis that these collocations are assigned a special status in the mental lexicon. However, this definition, while insightful, exhibits a certain narrowness. It excludes a significant range of co-occurrence types that, despite not possessing all the canonical features, have nonetheless been universally labelled as collocations. Remarkably, even within research that cites this narrow definition for collocation, instances of other co-occurrence types are presented as examples for collocations. Consider, for instance, the mention of *Prague Castle* as a collocation in Pecina (2010, p. 141); one could persuasively argue that "Prague Castle" remains compositional—it directly refers to the castle situated in Prague, and its unambiguous meaning derives from the combination of its components.

Similarly, Choueka (1988, p. 613) mentions word combinations such as *olive oil* and *chairman of the board* in a list describing the "type of collocational expression that we would like to locate in, and retrieve from, a large corpus" (p. 614). *Olive oil* is, however, compositional since it literally denotes oil made from and consisting of olives, and chairman of the board is substitutable (*chairwoman of the board*, *chair of the board*, *member of the board*) as well as modifiable (*chairman of the company*, *chairman of the party*). In summary, even prototypical collocations seem to defy these rigid boundaries, requiring further exploration and nuanced understanding.

For the purpose of this thesis, collocations are thus simply defined as words that systematically co-occur. Distinctions will be made between syntactic collocations – words that systematically co-occur primarily due to their grammatical functions –, and lexical collocations – words that systematically co-occur primarily as singular meaning-carrying units. Formulaic language such as 'I think" or "you know" are considered syntactic grammatical units and treated as a category of their own for the purpose of this thesis. The provided definition of collocation furthermore explicitly includes negative collocations, i.e. words that systematically co-occur less often than expected. Lastly, collocations are regarded as inherently directed constructions due to the psycholinguistic relevance of the internal word order as illustrated by the difference in meaning between *white house* and *house white* or *he is* and *is he*.

2.4.2 Types of Collocation

Having defined the concept of collocations, this chapter introduces a number of subcategories of word associations that have been discussed in previous literature in the hopes of providing a unified clear and comprehensive overview. Not only is this essential to prevent additional confusion in the field, but it is also worthwhile to delve into root of this large variety in subtypes of 'collocation'

before using them as the foundation for corpus-based network generation. The emerging categories share parallels with the aspects of collocational relationships listed in Brezina (2018, p. 65): collocational strength, frequency, 'position', collocate unit, and connectivity; but go beyond these categories and root the following analyses in various research traditions. The following five spectra of collocation types are discussed in order to provide a framework for later network analyses, allowing a classification of the identified collocations into specific subtypes of collocations.

Spectrum 1: Syntagmaticity/ Paradigmaticity

The first and broadest distinction is the distinction between syntagmatic and paradigmatic collocations (Kang, 2018, pp. 105–106; Michelbacher et al., 2011, p. 246; Utsumi, 2015, p. 18). Syntagmatic word associations occur when a set of words are connected on a horizontal, i.e. morphosyntactic level such as *mother earth* or *not least*. Paradigmatic word associations can be substituted for one another within a sentence as they fulfil the same function, *black* and *white* as or *hot* and *cold* are common examples. A further subdivision into more clearly defined types of relationships can also be made, paradigmatic relationships can, for instance, be split into hyponyms, synonyms, antonyms, meronyms etc. and syntagmatic relationships could be divided by into subject/object, modification or compounds/phraseological constructions (Kang, 2018, p. 106).

Many more and even smaller subdivisions are thinkable and employed in the context of highly specialised research questions, they are, however, not be discussed here for reasons of brevity. Paradigmatic word associations are less prototypical collocations and not commonly discussed in previous literature. Despite this, the comprehensive meta-evaluation of word association classifications by Fitzpatrick and Thwaites (2020) shows that paradigmatic relationships are highly relevant in word association contexts, and represent common response types.

Several studies have investigated to what extent different approaches to employing AMs result in more paradigmatic or more syntagmatic collocations. Rapp (2002, p. 7) reports a correspondence between the order of collocations and their type: second-order collocations (i.e. collocations that are associated via one shared collocate (McEnery & Brezina, 2019, p. 103)) are found to be paradigmatic and first-order collocations (i.e. directly associated collocations, ibid.) are found to be purely syntagmatic in their dataset[5]. Rapp further hypothesises that words connected via shared collocates may be more likely to be functionally interchangeable as they are found in similar contexts. For the classic example of a paradigmatic collocation, *first* and *last*, one could intuit that *first last* or *last first* would not – or very rarely – naturally co-occur, whereas a connection of these

[5] In this study, collocations that are found to be both first-order collocations and second-order collocations were disregarded.

terms via words like *day* as in *first day* and *last day* or *time* as in *first time* and *last time* seems more plausible.

An exploration of the BNC 2014 (Version 2; Love et al. (2017) & Brezina et al. (2021)) shows that this does indeed hold true for the abovementioned examples. Looking at the collocational profiles[6] of *first* and *last* it becomes clear that the two target words do not collocate with one another, but they are, in fact, second order collocates via *time* (*last time* log Dice: 9.1, 4,368 occurrences; *first time* log Dice: 10.7, 10,226 occurrences) and *day* (*last day* log Dice: 7.0, 622 occurrences, *first day* log Dice: 7.8, 1,258 occurrences), amongst others. This case study serves to illustrate two points: Firstly, the words *first* and *last* do share an indirect, paradigmatic connection and present second-order collocations through shared collocates. Secondly, some instances of shared collocates will express stronger preferences of *last* over *first* and vice versa – a phenomenon that might only become apparent when looking at a large-scale representation of these co-occurrences (i.e. in a corpus wide collocation network).

Spectrum 2: (A)symmetry

Another parameter can be combined with the differentiation into syntagmatic or paradigmatic collocations: Symmetry (Michelbacher et al., 2011, p. 248). A symmetric collocation consists of components with no particular directionality; an example for a symmetric paradigmatic collocation would be *first* and *last* and an example for a symmetric syntagmatic collocation would be *see you Û you see*. Their asymmetric counterparts would be collocations like *brother* and *sibling* (paradigmatic) as well as collocations like *post office Û *office post*. The asymmetry here stems from the fact that a father is always a family member (i.e. *brother* can be replaced by *sibling* in most situations) whereas not every sibling is necessarily a brother; analogous to that the word *post* predicts *office* considerably more than the word *office* predicts *post*.

Spectrum 3: Lexical / Grammatical

A further distinction of collocations into the categories of "grammatical" and "lexical" collocations (Gyllstad, 2014, p. 1) has also been made. The difference here is structural, a grammatical collocation contains at least one closed class (i.e. grammatical) item paired with at least one open class (i.e. lexical) item. A lexical collocation consists of purely open class items (Evert et al., 2017, p. 532). *join in* is an example for a grammatical collocation (*in* belonging to the closed class of prepositions), *Father Christmas* is an example for a lexical one.

[6] Collocation Parameter Notation (CPN; Brezina, 2018, p. 65): Log Dice(6.0), L1-R1, C5-NC-5.

Spectrum 4: Strength of association and predictability[7]

Another dimension according to which collocations can be classified is psycholinguistic in nature: predictable and unpredictable collocations. Vespignani et al. (2010, pp. 1683–1684) use this classification and illustrate the concept using *to cry over…* and *to break the….* While *to cry over* is highly predictable in that highly proficient speakers would be expected to complete it to *cry over spilt milk*, *break the…* could be completed using a wide range of non-metaphorical options (*car*, *window*) as well as the idiom *break the ice*.

An alternative, more fine-grained way of dividing collocations in sub-groups is not only by predictability, but also by collocational strength and usage properties. Li et al. (2005, p. 142) and Stulpinaitė et al. (2016, p. 33) take this approach and distinguish between four main types: loose, strong, fixed and idiomatic. Loose and strong collocations are compositional and less strongly associated than the two other types – *heavy rain* would be an example for a weak collocation and *gather strength* would be an example for a strong one. For both strong and loose collocations, the word order can be altered, but loose collocations are more open to being modified substituted by a near-synonym – *heavy* is, for example, easily substitutable with *hard* or *pouring* in the above example. Since this distinction depends heavily on researcher intuition, loose and strong collocations are treated as one category here. Fixed collocations are not synonym substitutable; their order cannot be changed, and they are not modifiable; *council tax* would be an example of this type of collocation. Lastly, idiomatic collocations such as *kick the bucket* exhibit all the properties of a fixed collocation, but they are also non-compositional.

Spectrum 5: Range

Lastly, types of collocations can also be determined based on the window in which researchers choose to investigate them. Gablasova et al. (2017, p. 158) and Evert (2008, p. 1215) use Evert's categorization which establishes three broad categories: Surface-level co-occurrences, textual co-occurrences and syntactic co-occurrences (Evert, 2008, p. 1220). Surface co-occurrences are the most basic type and identified purely based on the proximity of different words to one another in a running text. Textual co-occurrences on the other hand are only affected by structural boundaries such paragraph breaks, sentence or utterance and clause or website boundaries. Lastly, syntactic co-occurrences consist of words that can be much further apart than, for example, surface co-occurrences but are syntactically connected. Examples for these types of collocations are verbs and

[7] It needs to be noted that Spectrum 4 is discussed here since a categorisation into types of collocation would not be complete without engaging with predictability despite the fact that the empirical part of this thesis does not contain a breakdown of the frequency of predictable vs. unpredictable collocations. This is the case since coding for predictability is at least a semi-manual process and could not be justified given the temporal and spatial constraints of this thesis.

their object noun as well as entire patterns of constructions such as Verb-Argument Constructions (VACs, see Ellis and O'Donnell (2014, p. 71) for an in depth investigation of English VACs and their effect on language learning).

2.4.3 Overview

Having illustrated the five dimensions of collocational types in greater detail in the preceding sections these classifications can be used to analyse which association types are present in both word association networks and collocation networks generated using different AMs. Figure 7 depicts a schematic overview of the discussed dimensions.

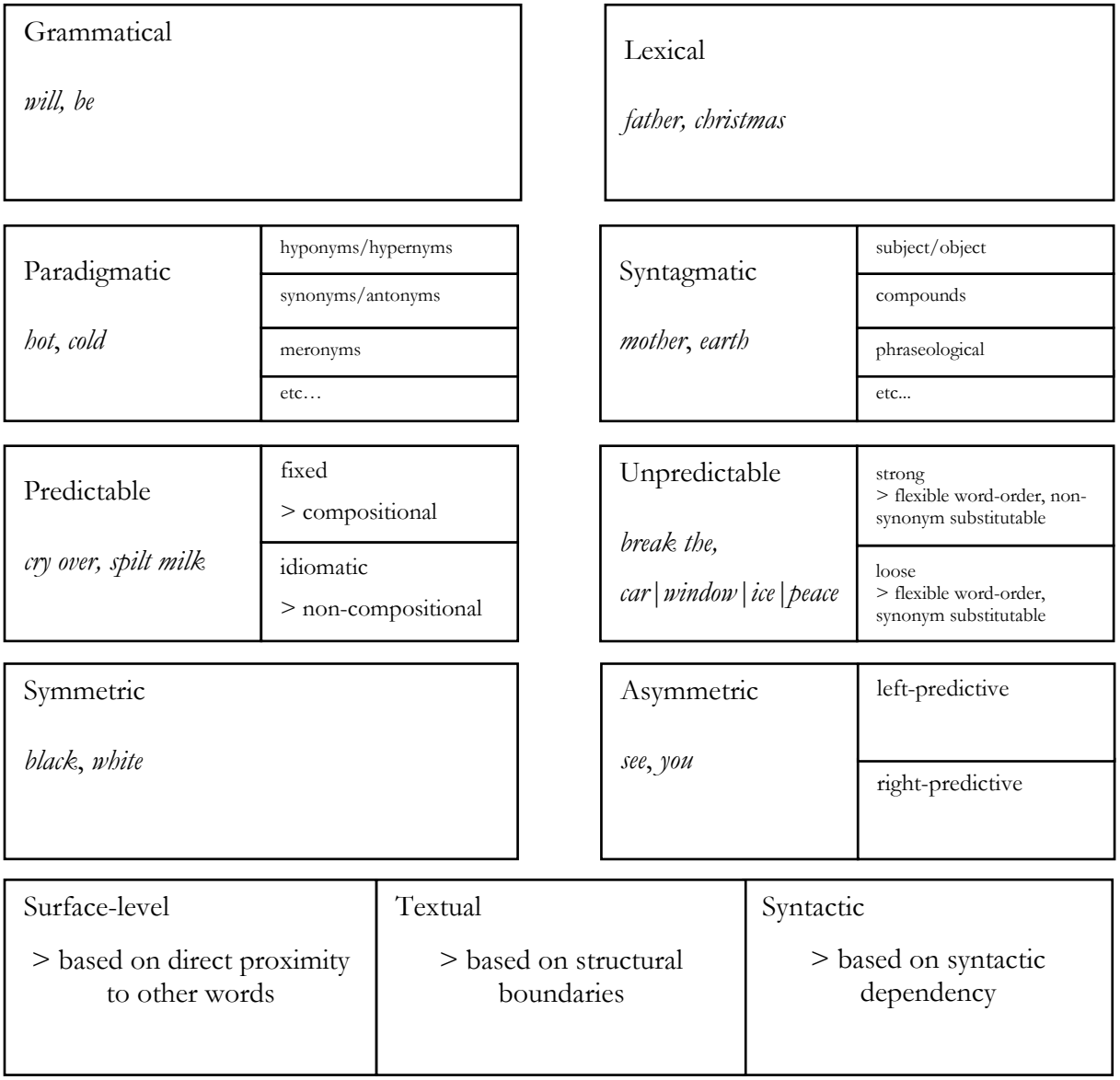


Figure 7: Illustration depicting the five spectra of collocations.

Some related concepts such as multi-word expressions (MWEs) are mentioned here to avoid confusion. MWEs are an umbrella term that encompasses a variety of different, partially

overlapping[8], elements of formulaic and/or regularised language such as idioms, proverbs, lexical bundles, compounds, binomials, phrasal verbs, and, crucially, collocations (Siyanova-Chanturia & Martinez, 2015, p. 549). The choice to examine collocations specifically in this thesis has been made since the focus lies on the *empirical phenomenon* of a collocation (as Evert (2008, pp. 1213–1214) puts it), as well as for reasons of compatibility with corpus linguistic tools and literature.

2.5 Intersections of Corpus Linguistics and Psycholinguistics: The Concept of Psycholinguistic Plausibility

Having explored the framework and theoretical underpinning as well as definitions of collocation and related phenomena, it is now possible to focus on the intersections between corpus linguistics and psycholinguistics. While the integration of psycholinguistics and corpus linguistics has been a contentious and relatively uncommon practice, there is a growing recognition of the value in merging these fields with the most common form of interaction being employing corpus data as stimuli for psycholinguistic experiments (Durrant & Siyanova-Chanturia, 2015). Besides this core area of interaction between the fields, existing research demonstrates that there is an immediate connection between a popular element of interest in corpus linguistics, collocations or word co-occurrences, and mental processing of word meanings; e.g. Mitchell et al. (2008, p. 1191) who show that neural activation when processing semantic knowledge can be predicted using word co-occurrence statistics. This research further links co-occurrence statistics for a select group of action words to the embodiment: The findings indicate that there is a special status of words directly related to sensory-motor functions since predictions on the basis of these terms are more accurate than predictions from other high-frequency words (Mitchell et al., 2008, p. 1194). More generally, studies on collocational processing show that words frequently appearing together facilitate faster processing (Carrol & Conklin, 2020; Sonbul, 2015; Vilkaitė-Lozdienė, 2019), although the advantage can be influenced by word order and language-specific conventions (see Vilkaitė-Lozdienė and Conklin (2021) for an example of Lithuanian collocation processing).

Extracting terms using corpus-linguistic methods and mapping the relationship between them via network analyses thus aligns the abovementioned approaches and fields. Along a similar line, the activation of action words in the mind via passive reading has been shown to result in activation of corresponding areas in the motor and premotor cortex (Hauk et al., 2004). For instance, the parts of the motor and premotor areas associated with hand movement are activated for processing the verb *to pick*, whereas for the verb *to kick*, activation of the parts of the areas associated with foot movement was observed. Hauk et al.'s findings also indicate that semantics can influence

[8] See e.g. Bauer (2019) for a discussion of the fuzzy boundaries between MWEs and, in this case, compounds.

overall mental processing patterns and strengthen relations between brain areas through repeated exposure. Understanding the interlinked nature of word relations and semantic connections is therefore integral for examining possible structures of the mental lexicon.

The following Chapter elucidates the reasoning for integrating corpus linguistic and psycholinguistic methodologies to characterise and assess collocation networks. The fusion of these methods and theories is realised through the notion of *psycholinguistic plausibility*. This concept employs a collection of experimental outcomes from psycholinguistics to guide and restrict methodological choices, aiming to construct a system that could potentially mirror mental linguistic operations. It is crucial to clarify that this is referred to as *psycholinguistic plausibility*, not *psycholinguistic reality*, since findings from psychology and neuroscience cannot definitely and exhaustively explain mental processing (yet or possibly ever). Given this situation, the objective is to devise a methodology that incorporates existing findings and thus models *one psycholinguistically plausible* rendition of the mental processes that might transpire during linguistic communication.

This chapter imparts knowledge on linguistically pertinent discoveries from these disciplines, thereby influencing the trajectory linguistic network studies. Key concepts that are extensively discussed include the mental lexicon (emphasizing linguistic retrieval mechanisms and language learning processes), Statistical Learning, and prevalent experimental designs in psycholinguistics with a focus on cue-response pairs.

In line with the explanations above, the comparison and interconnection of psycholinguistics and corpus linguistics via networks forms the crux of this thesis and presents a core part of its novelty. These fields are already thematically linked at a meta-level through the interaction of language production as depicted by the corpus and language perception and processing as represented by psycholinguistic data such as word associations. This is complemented by the overall structural similarities that have been identified in prior research (Steyvers & Tenenbaum, 2005, 54). Despite this, methodological decisions in corpus linguistics are rarely, and to the knowledge of the author never systematically, rooted in psycholinguistic findings. In order to illustrate how psycholinguistic research can practically influence corpus methodology, psycholinguistic findings related to Statistical Learning (with an emphasis on frequency and dispersion), phrase chunking processes, semantic representations, and the connective structure of the Mental Lexicon are presented and contextualised. Finally, the factors that render graph theoretical analyses an apt and highly effective method for comparing the resulting findings are provided. The individual graph theoretical parameters are also introduced and discussed to provide a comprehensive overview of the analytical repertoire and to demonstrate what these concepts mean in a linguistic context. Chapters 2.5.1, 2.5.2, and 2.5.3 aim to present the individual findings that are used in this thesis to

measure/quantify psycholinguistic plausibility, and will, in turn, inform and constrain all methodological decisions.

The network comparison carried out in this thesis is designed to model exposure data, that is language as it has been authentically produced, and in turn serves as a basis for language learning (corpus data), to networks designed to model stored linguistic representations (word associations). The following sections therefore explore Language learning with a particular focus on Statistical Learning, Language Production, and Linguistic Memory in the Mental Lexicon respectively and present empirical findings that can be used as the starting point for methodological constraints.

Network Representations: (Dis)similarity to/from Neuronal Networks

A last important precursory point is the question to what degree language networks can be equated with neuronal networks. Since this work is not grounded in neurology and the primary units of analyses in this thesis are (psycho)linguistic rather that neurolinguistic, brain topology and neural pathways will not be discussed here. It is especially crucial to acknowledge that many fine-grained cognitive processes are not at all well understood yet and blanket statements covering exact linguistic organisation at the neuronal level can under no circumstances be made on the basis of existing research. The networks described here are therefore in no way indicative of physical neural pathways and networks – despite the fact that evidence for a neural network structure does exist (Friederici & Gierhan, 2013, p. 250), as partially computationally simulated in Tomasello et al. (2018, p. 14). The mental lexicon, semantic maps etc. primarily aim to represent the connections between mental concepts, not physical pathways of neuronal activation.

Whilst it is important to acknowledge that neuronal systems and language systems are not the same, network science makes it possible to explore both via the same techniques. This, for instance, enables the study of language learning both from a linguistic perspective of networks reflecting learner vocabulary building and contextualisation, and also from a neurological perspective via network representations on the basis of fMRI measurements during the learning process (Bassett et al., 2011). This thesis focuses firmly on linguistic networks and is based on word associations and corpus-based collocations; all outcomes are therefore not to be interpreted as neuronal representations.

2.5.1 Language Learning Processes in the Mental Lexicon: Statistical Learning

The first section to explore in greater detail here is Language Learning. Of particular importance for this project are learning processes that lead to the memorisation of collocations and other linguistic constructions. Hebbian learning (Hebb, 2002) provides a mechanistic perspective on how

collocations could emerge during language development. When words frequently co-occur (e.g., as part of frequent collocations), their neural representations become interconnected. The strengthening of these neural connections aligns with the idea that collocations are learned through repeated exposure (Rapp, 2002, p. 1). Beyond this, the psychological law of association by contiguity states that language learning involves processes that seem to function similarly to obtaining co-occurrence statistics. This is due to the fact that experiences that come about at roughly the same time tend to create an association and evoke the other in the person that experiences them when occurring separately (Lachnit, 2003, p. 3). What links this to Chapter 2.3 is a closely related concept, experiential realism, which is a key view in cognitive linguistic frameworks: It posits that language is largely acquired by experience, and commonalities in language structures root from shared conceptual and sociocultural spaces as well as shared underlying cognitive processes (Evans, 2019, p. 193).

Statistical Learning

When exploring the interplay of psycholinguistic reality and corpus data further, the first and most prominent connection between the two fields is Statistical Learning (Ellis, 2006, p. 7; Frost et al., 2019, p. 1134; Peterson & Beach, 1967, p. 42). This concept has been researched extensively in psycholinguistics and SLA in the past 25+ years history of SL (Isbilen & Christiansen, 2022) and is also supported by, and compatible with, the central claims of construction grammar. SL has been extensively researched over the past two decades in neurology, psychology, psycholinguistics, and related fields, and select empirical findings are presented here to form a basis for methodological considerations in this thesis. While the umbrella term *Statistical Learning* encompasses a broader range of phenomena, one subtype of SL is deemed especially significant for psycholinguistic investigation (Frost et al., 2019, p. 1130): SL as an outcome of exposure to ordered auditory, visual, or tactile stimuli. Restricting the definition of SL to align with these parameters redirects the emphasis towards relationships among recurring events, thereby delving into entrenchment and more intricate mental transfers, thus probing learning phenomena that transcend mere pattern replication.

Growth (i.e. learning) processes and their potential underlying rules are relevant for both graph theory and for a thorough exploration of large-scale collocation networks in language learning research. The concept of SL builds on the idea that the frequency of a stimulus individuals are presented with as well as its surrounding context and patterns of occurrence are registered and processed as a part of the learning process. In addition to that, a consistent repetition of the stimulus consequently leads to a higher proclivity to learn it. This process is described to take place irrespective of the medium, i.e. auditory, visual etc. (Ellis et al., 2009, pp. 91–92; Rebuschat &

Williams, 2012, p. 2) and to be dependent on distinctiveness and overall contingency between a cue and its corresponding interpretation (Ellis, 2006, p. 1; Ellis & O'Donnell, 2014, p. 78).

Empirical Evidence supporting the Notion of Statistical Learning

The following sections contain a number of empirical findings that will be considered the basis of the 'status quo' of psycholinguistic research. Firstly, SL effects are reported for purely orthographical regularities as indicated through performance in wordlikeness tasks for children in Grade one to three (O'Brien, 2014), as well as wordlikeness, letter detection and reading tasks for adults after being presented with artificial scripts (Chetail, 2017, pp. 118–119). On a morphological level, (Sandoval et al., 2017, p. 8) have conducted an fMRI study using auditory stimuli that indicates that Statistical Learning is involved when learning linguistic categories, here specifically the gender of Russian nouns by English native speakers. Moreover, (Ulicheva et al., 2020, p. 13) also found evidence suggesting that long-term SL takes place when learning to categories non-words as adjectives or nouns on the basis of rules underlying English derivational suffixes (i.e. non-words ending in "-ness" are assumed to be nouns). These effects were assessed using nonword classification, spelling, and eye tracking during sentence reading. In an EEG study, Teinonen et al. (2009, p. 6) investigate word boundary detection in streams of natural speech for sleeping new-born infants. The authors' findings are in line with expectations on the basis of SL that neonates rely on translational probabilities for early language acquisition. Further findings from the field of child language acquisition indicate that children correctly perform both speech and tone element segmentation based on statistical knowledge of transitional probabilities (Saffran et al., 1996, p. 1927-p. 1928; Saffran et al., 1998, pp. 49–50).

This selection of exemplary studies from a wide variety of different linguistic subfields shows that SL effects are ubiquitous, empirically measurable, and they independently concern multiple layers of linguistic processing, such as orthography, phonology and morphology both short- and long-term. Empirical evidence further indicates that SL is built on interactions between different higher order brain areas, most prominently – but not exclusively – the hippocampus (Covington et al., 2018, p. 692; Schapiro et al., 2017, p. 12; Schapiro et al., 2016, p. 7), and modality specific (i.e. visual, auditory) areas (Sandoval et al., 2017, pp. 10–11)

Both auditory and visual SL are relevant in a linguistic context since language perception and production, respectively in the form of listening and reading, occurs in these modes. While it is tempting to conflate these two branches of SL by understanding them as a uniform instance of language learning, studies researching auditory and visual SL found fundamental differences: lifelong linear learning processes for visual SL and a plateauing in learning performance for auditory

SL in learning patterns based on the different modes (Emberson et al., 2019; Raviv & Arnon, 2018, p. 11). Siegelman and Frost (2015, p. 108) further explore the relationship between general cognitive abilities and SL performance using a range of four different SL experiments. They employ combinations of visual and auditory, as well as verbal and non-verbal tasks, paired with six cognitive tests such as, among others, syntactic processing tasks, rapid naming tasks, and verbal working memory tasks. In line with previous research, they find evidence suggesting that SL is highly componential, mode-dependent, and cannot be regarded as a unified capacity. What is more, they also found that general cognitive abilities are not predictive of SL performance or vice versa, suggesting that there is no strong link between these capabilities (Siegelman & Frost, 2015, p. 118). As a consequence of these findings the term *Statistical Learning* is understood to act as an umbrella term covering a variety of different and modality-specific SL processes (Frost et al., 2019, p. 1134).

SL shows that concepts are robustly learnable if they possess Zipfian type-token usage distributions, occupy verb forms selectively and are semantically coherent, all of which apply to collocations (Ellis & O'Donnell, 2012, p. 295). Further research also shows that, when presented with some linguistic input, humans tend to store concepts at the functional level, that is as key lemmas and event structures, rather than on a phonological and positional basis (Menn & Dronkers, 2017, p. 181). In other words, humans tend to remember key concepts of the content of a sentence rather than being able to reproduce a given sentence verbatim. They fulfil the requirements of all factors mentioned in Ellis and O'Donnell and operate on a functional, lemma-based level. This suggests large differences in the efficiency of the production and comprehension of linguistic information that have been learned such as strong collocations versus linguistic input that one has not been systematically exposed to. Indications for that would be longer reaction times and a higher cognitive load when encountering non-collocations versus collocations. In combination, the abovementioned properties make the linguistic items that lie at the heart of this dissertation, namely collocations and strongly associated words, key concepts for language learning processes. This emphasises the key role of collocations for learning processes in that they represent ideal candidates for robustly learnable items; see Chapter 2.6.

Collocations and Statistical Learning

Having explored SL in greater detail as a foundation for this chapter, it is essential to take a closer look at the overlap of collocations and SL in particular. A range of experimental studies investigate contingency learning, i.e. learning processes that involve registering the relative frequencies of certain form-function mappings, and find that statistical measures such as χ^2, r_φ, or ΔP (Ellis, 2006, p. 7) correlate with human learning of contingency (Peterson & Beach, 1967, p. 42); for an expanded discussion of the psycholinguistic validity of individual AMs see Chapter 3.2.3). The idea

of associative learning (Perruchet & Poulin-Charronnat, 2012, pp. 137–138), i.e. the notion that ideas and concepts that occur together are reciprocally reinforcing each other, is also closely linked to these findings. Adding to these observations, research indicates that the learning processes of L2 learners are characterised by induction through usage-based statistic experience (Ellis et al., 2015, pp. 357–358). Empirically identified components of this statistic experience are prototypicality, closeness of mapping and type-token ratio distributions of specific constructions and formulaic expressions. This directly applies to collocations which makes them central components of language acquisition processes (Ellis et al., 2015, p. 375). This is further underlined by corpus-based evidence implying a connection between higher proficiency and the use of – increasingly abstract – collocations that form a semantic unit (Brezina, 2018, p. 73). The phenomenon of repeated use of known patterns has also been discussed extensively in linguistic literature; Hoey (2005, p. 8), for instance, describe the construction of mental profiles for collocational patterns as lexical priming (Berber Sardinha, 2020, p. 4).

With a focus on the linguistic elements of interest here, studies based on collocation recognition support the theory of frequency-based learning. In an experimental study exploring the processing of known linguistic items involving 45 participants, Vogel Sosa and MacFarlane (2002, pp. 233–234) find that collocational frequency negatively correlates with reaction times in a word monitoring task using the target word *of*. This is an indication of inhibition effects in cases where *of* is part of a collocation. This implies both a psycholinguistically real and measurable impact of collocational frequency effects on language comprehension and production and the holistic storage of certain collocational structures in the mental lexicon.

In a language learning context, studies have examined statistical learning effects (and therefore effects on previously unknown items) specifically on collocations. Webb et al. (2013) investigate how different reading modes impact the incidental learning of collocations. They found that learners with higher prior vocabulary knowledge and congruent collocations had better learning outcomes, suggesting that frequency and repetition are crucial factors in effective collocation acquisition. Additionally, they further examined effects of reading with textual input enhancement (underlining) which were found to lead to the highest learning gains, highlighting the importance of not only repeated but also focused exposure to target collocations. More recently, a large-scale longitudinal study with 100 Vietnamese pre-intermediate EFL students conducted by van Vu and Peters (2022) revealed that, in this context, encountering collocations 15 times within a graded reader led to significant learning gains, emphasising the positive effect of repetition and input enhancement.

Summary, Limitations and Outlook

To sum up, the presented findings suggest that SL effects are ubiquitous, empirically measurable, and concern multiple layers of linguistic processing, such as orthography, phonology, and morphology. As a result, SL demonstrates a large potential for linking corpus methods with Psycholinguistics. This is the case since a corpus can not only provide frequency information regarding the occurrence of individual words and phrases; by using a corpus the researcher can also tease out statistically interesting elements due to other factors relevant to SL processes such as exclusivity, cohesiveness, abstractness, and connotativity. The corresponding methodological basis for uncovering a range of these factors in corpora is association measures, their suitability to represent these parameters is also discussed in-depth in Chapter 3.3.

As this section has demonstrated, SL is a vast field of research, and even a brief summary of some of the most relevant empirical studies in the context of this thesis is quite extensive. As Frost et al. (2019, p. 1142) suggest, SL as a general field thus requires what they call "an ecological theory of SL". This theory should focus on empirically unravelling the series of constraints that predict what will be learned, what will not be learnt, and why this is the case. Importantly, the theory should explain how learning proceeds when the stream to be learned is complex and not uniform in terms of sizes of units and the statistics of their co-occurrence.

The development of such a theory would be a significant step forward in the current understanding of SL and its role in language learning. It would not only enhance the understanding of the mechanisms underlying language learning but also provide valuable insights into the broader cognitive processes involved in learning. This, in turn, could have far-reaching implications for various fields, including neurology, psychology, psycholinguistics, and related fields. For this reason, Chapter 2.2.1 discusses the utility of network approaches in this context, and explicitly investigates which factors that have previously been identified as relevant to SL can be measured and implemented as part of collocation and collocation network research.

It is essential to consider the limitations of these lines of thought: While the aim is ultimately to disentangle learning processes in order to allow for deeper insights into how humans perceive, process and store linguistic information, there is, at present, no possibility to conclusively identify whether or not specific statistical calculations actually take part in the mind. Several attempts have been made to create models that aim to develop an explanatory power by incorporating SL processes into larger mechanisms of thought. Models based on automatic calculations of statistical regularities (SRNs) show that raw statistical computations as the basis for SL are generally plausible (Perruchet & Peereman, 2004, p. 99), but an alternative, not currently falsifiable theory would be

that real-world processes simply result in effects that happen to fit statistical models reasonably well (ibid, p. 116). More fine-grained approaches to studying the neurological processes relevant to linguistic phenomena are therefore needed to investigate this further.

A further approach, latent semantic analysis (LSA (Landauer & Dumais, 1997, p. 211)), aims to tackle observations in language learning that cannot be explained by SL alone: Linguistic knowledge acquisition seems to transcend direct experiences in that learners seem to be able to use a vocabulary beyond what they have previously directly encountered and have been taught. Landauer & Dumais, 1997, pp. 234–235 therefore assume that this can only be explained though an inference process which would then, in line with LSA, rely on a high dimensional space in the mental lexicon. While this high dimensional space cannot be directly modelled since very few of its properties have been probed empirically, network approaches could present a valuable tool for hypothesis generation, e.g. by suggesting words that are embedded in similar areas of the network as candidates for mutual exchangeability.

2.5.2 Linguistic Memory in the Mental Lexicon

A natural question emerging from studying language learning processes is how the entrenched connections are stored and, later, retrieved. This Chapter is titled *Linguistic Memory in the Mental Lexicon* since entrenched language essentially constitutes linguistic memory (Divjak & Caldwell-Harris, 2019, p. 73). A commonly referred to concept in psycholinguistic research is the way in which words are interlinked and stored in the human mind to facilitate rapid linguistic comprehension and production. This system of interlinked linguistic items is the Mental Lexicon (ML (Aitchison, 2008; Tucker & Ernestus, 2016)), a concept immediately relevant to the present thesis since such a structure of complex interrelations is a prime candidate for network representations. The ML has been extensively researched, and several especially influential findings are presented in this Chapter.

The mental lexicon is not only an interesting subject of study since it represents the internal structure of our lexical knowledge and might thus hold the key to a better understanding of human cognition and learning processes as a whole, but it has also been found to impact specific characteristics such as creativity and fluid intelligence. In their studies using network science approaches to ML data, Kenett et al. (2016, p. 377) and Siew et al. (2019, pp. 12–13) found that the structure semantic networks, specifically features such as a higher average shortest path length, impacts fluid intelligence and flexible small world properties influence the overall creativity of an individual. Examining the ML on the basis of semantic networks also helped further research into

relative processes and these explorations indicate that both executive and associative creativity exists. It thereby unifies two theoretical schools of thought (Benedek et al., 2017, p. 164).

Furthering the exploration of the ML, EEG semantic decision studies[9] have provided valuable insights into the neuronal organisation of the ML. A study conducted by Ploux et al. (2012, p. 210) discovers an organisational structure along the dimensions of animacy (i.e., living vs. non-living things) and proximity to the individual in French native speakers. This study, however, only reports ERPs in situations where individual words were presented to participants rather than larger meaning-carrying units or sentences. The hypothesis that emerged from this research suggests that the mental processing of lexical concepts that are living and close to the participant, such as *people* and *clothes*, differs from the mental processing of concepts such as *fruits* and *tools*. These findings further underscore the complexity and intricacy of the ML, highlighting its potential role not only for understanding not only linguistic comprehension and production but also broader cognitive processes. After illustrating the relevance of researching the ML, two core research areas that are directly relevant to the collocation network approach taken in this thesis are presented in the following section.

One question fundamental to gaining a better understanding of the organisational structure of the ML is to what degree collocations exist as separate entities in this space. One could, for example, imagine *orange juice* could - similarly to compound words like *likeable* - be either stored as two distinct concepts (i.e. *orange; juice* and *like; -able*) which follows the so-called morphemic model (Taft & Forster, 1975) or as one singular unit which is described in the full-listing model (Butterworth, 1983). The methodological unification of these paradigms resulted in a third model: the partial decomposition model which takes possible changes according to specific types of morphological forms into account and thus allows for both unified and composite representations of collocations. Evidence for the partial decomposition model has been found by Dasgupta et al. (2016, pp. 853–854) as part of their study on Bangla compound words; participants employed the morphemic or full-listing model depending on several underlying features such as morphological and orthographic complexity. In a further exploration of this phenomenon, McConnell and Blumenthal-Dramé (2019, p. 23) find additional evidence for the co-existence of the full-listing model and the morphemic model to the effect that words and multi-word units are processed concurrently. This carries the implication for the present thesis that it is methodologically sound to work with space separated units as the default collocational units and to create a meaningful network on this basis whilst acknowledging that an alternative and equally valid MWE-based network also exists.

[9] More on these methods and their utility and possible explanatory power in Chapter 2.5.3.

Another essential question is what the overall structure of the ML looks like, and which elements might play a special role within it. Traditional psychological views suggest that concepts would be organised hierarchically into natural categories via perceptual similarity, but there is considerable dispute surrounding this theory and an alternative structure on the basis of thematic relations has also been proposed (Deyne et al., 2016, p. 52). It is, however, important to mention that neither of these options have the explanatory power to present a coherent taxonomy of the entire ML based on one factor alone; thematic or hierarchical factors should rather be understood as the main ordering factors among many others (Deyne et al., 2016, p. 66). What can be said is that experimental evidence suggests that frequent co-occurrence of action words and physical actions carried out by the participant result in the establishment of corresponding cortical links (Hauk et al., 2004). This suggests that cortical areas for movement can be linked via repeated co-occurrence of action words which, in line, activate the corresponding neuronal assemblies. While it is unclear how generalisable this is beyond action words, it clearly demonstrates how linguistic perception of repeated cooccurrence can influence the shape of both linguistic, and crucially also non-linguistic cognitive patterns.

Figure 8, adapted from Kovács et al. (2021, p. 194) provides a visualisation of a possible meta-structure of the mental lexicon. All of these factors are necessary conflated in the present study - neither word association data nor the type of collocations allow for an exploration of these dimension separately. The resulting networks rather need to be considered the end result or overarching meta-network that emerges when all of these layers are superimposed, and the strongest links are followed. Hypothesising which of these layers are of particular importance would be mere speculation, more so since the layers displayed in the illustration are based on human classifications of the complex language system into discrete components which might, in itself, not represent neurological or psychological realities. Syntactic and semantic processing, for instance, are intertwined (Yamada & Neville, 2007, p. 177).

Stella et al. (2018) explore the connections between words encoded in a network composed of a number of layers, namely taxonomic relations, phonological similarity, word association, and synonymy. While these do not exactly mirror Kovács et al.'s theoretical layers they present a valuable approach to assessing multi-layer interactions. In their paper, Stella et al. find a central word cluster that contains words with a special function, both in terms of higher frequency of natural occurrence, ease of learnability and memorisability, and semantic richness. Their study therefore sheds light on what a core component of the Mental Lexicon spanning multiple layers could look like.

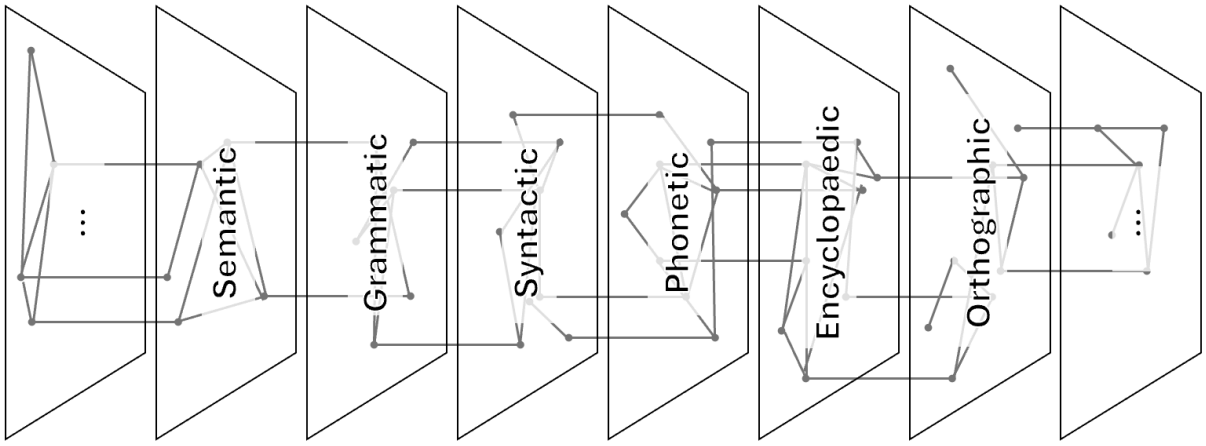


Figure 8: Possible interacting layers of the Mental Lexicon, adapted from Kovács et al. (2021, p. 194); different layers added.

Examining possible structures of the ML further Vitevitch and Goldstein (2014, p. 136) investigate a phonological network on the basis of auditory naming task data and find the following overall structure: The network consists of several clusters surrounding a central highly connected component as well as a large number of isolates. The authors report small world characteristics (defined and further discussed in Chapter 2.7.3), assortative mixing by degree, and a degree distribution following a power law within the largest connected component (Vitevitch & Goldstein, 2014, p. 132). They furthermore employ several conventional psycholinguistic tasks such as perceptual identification tasks in order to assess whether or not there is a difference in recognition based on the position of words within the network (Vitevitch & Goldstein, 2014, p. 136). The results show a measurable processing advantage for keywords over foils, indicating a clear potential of network approaches to investigate psycholinguistic phenomena. The presented findings suggest that speed and accuracy of language processing and production depend on network-based properties of certain lexical items (Vitevitch & Goldstein, 2014, p. 143).

Focusing on word association data, Deyne et al.'s (2016) work is of extraordinary relevance since this mirrors one of the datasets examined in the empirical part of this thesis. The authors investigate the structure of the mental lexicon on the basis of large-scale word association data in Dutch and identify a number of central hubs that might represent hubs in the mental lexicon overall. This would support the existence of a thematic organisation within the Mental Lexicon. The identified themes include the Dutch terms for *water*, *food*, *money*, *car* and *pain* (Deyne et al., 2016, p. 58). This finding, while potentially not being perfectly transferable to English for a variety of reasons (Deyne et al., 2019, p. 1002), implies that the central hubs in the mental lexicon might coincide with concepts that are strongly connected to (modern) basic needs. Since the abovementioned studies,

especially Deyne et al. (2016), present a large advance in the field and directly address some of the core questions that this thesis poses; this approach is developed further and extended to collocational and word association network analyses of British English in Chapter 4.

Lastly, it is important to emphasise that while the past 20 years of intensive research in this field have led to substantial advances and allowed for the exploration of some key elements of the ML, a lot of the inner workings of our linguistic memory have not yet been explored. It is further essential to view the findings that have been obtained with a certain degree of scepticism since they underlie a wide range of limitations. One particularly relevant example of this is interfering factors such as changes in emotional states; Kousta et al. (2011, p. 14) for instance present evidence for marked differences as to how emotionally charged concepts are processed in contrast to neutral ones. Beyond this, age effects also play a role both by impacting the overall structure of the ML leading to less efficiently organised and overall less connected networks with increasing age (Castro & Siew, 2020, p. 15) and by affecting the individual words present in the network via Age-of-Acquisition effects (Steyvers & Tenenbaum, 2005, p. 74). This is the case since elements that have been acquired first have been found to play a key role in the ML overall. Concept internal features such as the level of abstractness and connotativity have also been shown to impact structural properties of the ML; these are very hard to disentangle and identify computationally which further exacerbates these limitations (Vankrunkelsven et al., 2018, pp. 3–4) A further possible factor lies in the positioning of individual linguistic units within a sequence since positive processing effects have been found for components that mark a boundary when compared to components in a central position (Perruchet & Poulin-Charronnat, 2012, p. 138). These factors, and potentially more sources of interference that have not been explored extensively in previous research such as socioeconomic factors or specific personality traits in people taking part in such studies, are likely to have impacted and distorted the results of the conducted empirical studies.

Semantic maps and their structures

The layer of the ML that can be approximated most effectively using collocation networks is semantic, and a large body of research into specifically semantic maps exists. Semantic maps are virtual representations of linguistic meaning relationships, and they are often based on what is a filtered association network only containing the most lexicalised items. The nodes of semantic maps usually represent different meanings and the connecting edges signify that two nodes (most commonly nouns) are connected through a special relationship such as synonymy, antonymy, hypernymy, hyponymy, meronymy, holonomy, or even collocation (Rijk & Mareček, 2020, p. 36). The most intensively researched connection is synonymy where two different lexical items are assigned a shared meaning. The shapes of semantic maps do, however, vary considerably – some

even cover inter-linguistic relationships such as words in different languages that map onto the same meaning (Georgakopoulos & Polis, 2018) e.g. for translation studies.

In their paper, Gravino et al. (2012) explore large scale word association networks and the semantic relationships between the nodes within said networks. The authors found that specific semantic relations such as synonymy, hypernymy and hyponymy occur particularly frequently. The graph theoretical analysis further showed that preferential combinations of semantic relations exist such as a high co-occurrence of cue words and targets on the same level of specificity.

There are, however, also alternative approaches to this that allow for more varied concepts than individual words to be considered linguistic units. Such approaches permit compounds and word classes other than nouns as nodes which allows for insights into more general psychological events underlying language production, comprehension and learning processes. This is based on a number of assumptions such as the existence of a differentiation mechanism which builds on previously known words as the main learning strategy in terms of a meaning expansion mechanism (Borge-Holthoefer & Arenas, 2010, pp. 1288–1289) – in other words: it also relies on Statistical Learning.

Theoretical foundations of semantic maps

This subchapter presents the status quo in terms of theoretical frameworks produced to model the organisational structure of lexical knowledge. Three broad schools of thought are explored in greater detail: hierarchical/Aristotelian models, vector models and the more recent ACOM (Automatic Contexonym Organising Model) as first proposed in Ji et al. (2008, p. 926). It is also important to mention that verbal and visual semantic memory appear to be stored in two separate memory hierarchies, as indicated via affective priming studies carried out by Ellis and Frey (2009, pp. 479–480), and the models presented below aim to describe verbal semantic memory only. A theoretical exploration of semantic maps is required since the storage of semantic units in the ML is central to the both the question of how effectively large linguistic networks can capture this structure and what differences are expected to be encountered between corpus-based and word association-based networks.

The first model to be presented here is the Aristotelian hierarchical model. This is based on the assumption that the lexicon can be organised into a tree-type structure with each branch in the tree inheriting properties of the branch from which it stems. WordNet (Fellbaum, 2006) constitutes one exemplary application of this model; this dataset consists of a manually constructed extensive collection of lexical concepts connected via synonymy, hyponymy and other concepts (Ji et al., 2008, pp. 926–927). Individual groups are categorised into synsets (sets of near-synonymous word meanings (Ellis & O'Donnell, 2014, p. 85)) that are then interlinked on a higher level. Despite its

name, WordNet does not operate on the word level – individual meanings of words are considered discrete nodes. To give an example, the search for *language* in WordNet 3.1 results in six different synsets: language as a part of (1) linguistic communication, (2) oral communication, (3) lyric or word, (4) linguistic process, (5) speech, (6) terminology. Some issues with the Aristotelian model become immediately apparent when scrutinising these examples: (2) oral communication and (5) speech seem to be directly equivalent – only a look at the provided example sentences helps to shed light on this: (2) contains the example sentence "he uttered harsh language" which relates to an individual occurrence of a speech act, whereas the example for (5), "language sets homo sapiens apart from all other animals" refers to language abilities in a more general sense. Interestingly, however, the example for (3), "the song uses colloquial language" is almost indistinguishable from the example in (2) in that it also refers to a singular instance of spoken language. This raises questions as to whether these should be considered separate synsets – these deliberations will vary considerably based on the researcher's intuitions and are therefore rather unstable while significantly impacting the structure of the tree/network. Further issues with the Aristotelian model are manifold such as an imbalance based on word class with nouns heavily overrepresented in the inheritance trees. The model furthermore suffers from a number of definition gaps for certain more complex umbrella terms that cover a range of different members; occasionally, these members do not share any universal features. The most famous example for this is Wittgenstein et al.'s (2009) example of the concept *game* where types of games are so different that there are no features that could be applied to all of them. Lastly, data from Ji et al. (2008, pp. 926–927) suggest that there are differences in processing speed when subjects are presented with sentences containing classifications of the form "A SUB-CATEGORY is a CATEGORY" despite the fact that this would be expected to exhibit identical processing speeds regardless of the particular sub-category in the Aristotelian framework.

The second type of models, vector models, on the other hand, represent a context-based approach where every word is represented by a vector. The relatedness of two words can then be calculated via their proximity, i.e. the cosine of their vectors. Several studies indicate that this approach is more appropriate to represent human behaviour as they, for example, found that reaction time speeds in semantic priming experiments correlate with calculated vector distances (Ji et al., 2008, p. 927; Vigliocco et al., 2004). Examples for popular applications of vector models are Latent Semantic Analysis (Landauer & Dumais, 1997, p. 211), and deep learning models such as word2vec (Mikolov, Chen, et al., 2013), GloVe (Pennington et al., 2014), and BERT (Devlin et al., 2018). These models are all data-driven and can thus, unlike WordNet, dynamically update their knowledge as new data becomes available. A major shortcoming of this approach is the lack of

interpretability of the vectors themselves. While they show impressive performance, they are not suitable as potentially explanatory models. Chapter 3.2.2 explores the suitability of vector-based approaches for word association extraction in greater detail.

Thirdly, corpus-based semantic maps (Matusevych & Stevenson, 2019) are presented, one of which is the geometric model (Ji et al., 2008, p. 928). The novelty of the geometric model is that it aims to represent the intrinsic structure that connects different concepts and could therefore provide results similar to vector-based approaches, but it can also be used to map word senses. It follows the broader idea that the meaning of a word is usage-based and it relies on Semantic Atlases (Ploux et al., 2010, p. 356) and corpus data. The model works on two semantic levels; the first are semantic points or cliques which are established on the basis of contexonymy, i.e. they consist of closed groups of words that are linked to every other word in the group. The second component used in the geometric model is the notion of a semantic area which covers the semantic points identified for a specific word in question. Each concept is thereby assigned a multidimensional space in which its associated cliques are positioned; this can then be divided up into zones and compared to the zones of other concepts. In practice, this is achieved computationally through filtering, clique computations, factor correspondence analysis and hierarchical clustering (Ji et al., 2008, pp. 928–929).

2.5.3 Retrieval Processes in the Mental Lexicon

Following a discussion on language learning and the static structure of machine learning this chapter explores processes associated with retrieving linguistic information. Psycholinguistic experiments aim to observe and interpret unconscious linguistic phenomena or those that occur too rapidly to be processed in real-time (Menn & Dronkers, 2017, p. 167). Various experiments have been used to study language retrieval, ranging from low-resolution approaches like association experiments, which require participants to produce words based on a target word (Ji et al., 2008, p. 930), and lexical and auditory decision tasks, stimulus-response pairs, or eye-tracking tasks to advanced brain imaging techniques such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG). The latter techniques offer high-resolution insights into parameters like reaction times and cognitive load, reflecting the mental processes underlying language competence. Consequently, a brief explanation of these methods is provided alongside linguistically relevant findings in this Chapter.

fMRI and EEG studies allow for non-invasive monitoring of specific brain regions in order to map certain regions of interest and infer new insights into the anatomy and organisation of language (Friederici & Gierhan, 2013, p. 250). ERPs (Event Related Potentials) re specific patterns of

electrical activity that are time-locked to a particular sensory, cognitive, or motor event and are derived from the EEG by averaging the brain's response to repeated occurrences of the same event. The usual procedure associated with EEG studies in Multi-Word-Extraction and collocation research is as follows: Participants are fitted with electrodes on different parts of the scalp to measure neurophysiological activity in the brain. The signals registered by each electrode is then amplified and recorded for analysis. This method is non-invasive and does not cause any health risks to participants. In linguistic research, the next stage of the experiment involves prompting the participants to execute a linguistic task such as reading or listening to a text or producing speech or writing. The changes in the recorded electrical signals as a consequence of the linguistic operation in the participant's mind can then be recorded and later analysed.

Looking at psycholinguistic findings regarding collocations in the mental lexicon, Siyanova-Chanturia et al. (2017) exemplify the power of these fine-grained, neurophysiological approaches in providing concrete empirical evidence. Event-related potentials (ERPs) have been recorded in EEG studies involving 30 participants (L1 speakers of English) for this study. They were presented with commonly co-occurring words such as "knife and fork" as well as control phrases consisting of a similar, plausible, but less typical word combinations, here "spoon and fork" as well as a third combination of a semantically unrelated and relatively implausible word with fork, here "theme and fork" (Siyanova-Chanturia et al., 2017, p. 114). The study put a particular focus on electrophysiological responses occurring during two time-windows after being presented with the stimulus, 250-350ms and 350-450ms.

These time windows are of particular interest since they have been established as relevant to predictive mechanisms in neurophysiological research. The first timeframe of particular importance is linked to N400, a single-phase negativity occurring between 200ms and 600ms and peaking around 400ms after stimulus onset predominantly in centro-parietal areas of the brain (Kutas & Federmeier, 2011, p. 623). N400 effects have been found to be reactions to linguistic stimuli, but they also occur in the context of face and gestural processing as well as, amongst others, in mathematical recognition. In a linguistic context, this negativity effect has been found to be a marker for prediction activity. Findings from recent N400 studies support the theory that N400 effects are a result of pre-activation effects in the brain and that these predictions are an essential element of sentence comprehension overall. Existing research also indicates that sentence comprehension relies on permanent updates as new information is ingested (Szewczyk & Schriefers, 2018, p. 682).

The second measure explored here is P300, a positive wave first recorded at around 250ms after stimulus onset and peaking at around 300ms after stimulus onset P300 (Vespignani et al., 2010,

p. 1685). This effect is attributed to general context updating mechanisms (Donchin & Coles, 1988, p. 417), for example when checking whether an expected word such as an antonym is present or if a word or sentence is correct (Vespignani et al., 2010, p. 1683). It also precedes more elaborate semantic representations. P300 have in practice been used to test for prediction mechanisms when participants are presented with the first part of idioms and other types of collocations. P300s were found in situations where a specific completion of a collocation is expected, i.e. after the collocation has been recognised, whereas N400s were observed for probabilistically unexpected words before the recognition point (Vespignani et al., 2010, pp. 1696–1697).

Returning to the elements of interest, collocations, and Siyanova-Chanturia et al.'s (2017) study at hand, the findings indicate that measurably different mental processes underlie the perception of existing collocational patterns when compared to new material. The study furthermore implies that frequent collocations and infrequent word co-occurrences differ systematically since the former display better semantic integration and lower cognitive load (Siyanova-Chanturia et al., 2017, p. 121). This brief aside demonstrates the potential of neurophysiological studies when it comes to explanations and detailed descriptions of language processing in the mind – however, this equally underlines how essential corpus databases are for this endeavour since the example phrases for Siyanova-Chanturia et al.'s experiment are corpus-derived. While brain imaging approaches are not employed in the present thesis, the development of a method to model corpus data to fit cue-response associations presents a starting point for a large number of further studies in this area and the graph-theoretical exploration of these models themselves provides insights into the possible activation patterns and paths of information spread.

Other studies recording ERPs showed that predictable (classes of) nouns can be pre-activated before they actually occur in the text (Szewczyk & Schriefers, 2018, p. 665). In this instance, the study at hand considerably furthers knowledge of collocational processing in the mind indicates that there are mental retrieval processes that set collocations apart from other linguistic elements.

Another factor that influences the overall retrieval performance in linguistic tasks is the concept of spreading activation (Deyne et al., 2016, p. 72). This theory infers that concepts neighbouring a concept in use (e.g. something that has just been perceived or talked about) will be activated alongside the originally perceived concept. This then facilitates the retrieval of the neighbouring concepts and thereby decreases reaction times and effort, before the effect fades over time (Siew et al., 2019, p. 9). Interestingly, this constitutes a direct connection to network approaches to linguistic networks: Spreading activation can be simulated in networks using random walks from a starting node to its neighbours. Experimental studies have shown that this process has the potential to predict human memory retrieval behaviour (ibid.). In an examination of reaction times for

phonological items, Chan and Vitevitch (2009, p. 1936) theorise that items with a low clustering coefficient will receive a higher relative spreading activation since they do not underlie the same amount of competition as items with a high clustering coefficient. This is hypothesised since a high clustering coefficient inevitably leads the linguistic item to pass on some spreading activation to a large number of other items. At the time of writing, the author is unaware of studies aiming to replicate this process for semantic networks.

Other retrieval processes have also been explored, or show potential to be explored, using graph theory. One of these is to do with mental navigation which takes place as part of the language process; this process directly impacts cognitive load and is limited by it. In graph-theoretical terms key nodes would, for instance, be expected to be an example for very central, easily accessible items and the small-world properties of the collocation networks could be explained through their efficiency (Borge-Holthoefer & Arenas, 2010, p. 1284). Small world networks (see Chapter 2.7.3), by definition, allow for quick and efficient navigation between nodes in large networks.

A related meaningful avenue for categorising and analysing nodes and collocates is the encoding/decoding effort they require. Grammatical and thus less context-bound collocations would be expected to play an important part in language production as they can be seen as serving as the linguistic base of a sentence due to their low encoding effort and overall low strain on memory resources. Lexical items that possess a higher encoding effort can then be added to the grammatical structures to reduce ambiguity and decoding effort (H. Chen et al., 2018, p. 8), thus minimising cognitive load overall. Large Part-Of-Speech tagged collocation networks paired with psycholinguistic experiments to measure reaction times could play a key role in investigating this relationship further.

Another important factor that plays a role in accessing the mental lexicon is cognitive load, i.e. the working memory resources used in a particular situation (Navarro et al., 2020). This is an important metric for measuring the psycholinguistic reality of collocations and it equally partly explains their existence. Measuring cognitive load is in practice commonly achieved using eye-tracking techniques to investigate the frequency, number, and duration of fixations, and, in this case, also saccade jumps (Keating, 2013, p. 74). Limitations based on cognitive load, are also likely to impact the maximum window size (i.e. the maximal number of individual lexical items grouped together) for word dependencies in general and therefore also for collocations (H. Chen et al., 2018, p. 17).

Lastly, one particular study going beyond the retrieval stage and investigating the role the ML can play in language production processes, Kang (2018, p. 110), has a large overlap with the project at hand. Their work investigates the how well primary responses on the basis of word association

tasks from the Edinburgh Association Thesaurus (Kiss et al. (1973), aiming to represent the ML) maps onto a list of top collocates from the BNC 1994, aiming to represent language production. The authors found that around half (52%) of the primary response words were included in the top 50 paragraph-wide collocations identified using simple LL, t-scores and local MI. This observation strongly implies a meaningful connection between the ML and language production – especially considering the statistical probabilities at play when observing an exact match between 26 out of 50 words between the word association tasks and the collocate list since any conceivable word is a possible candidate.

2.6 Summary: Key Findings from Psycholinguistics

After an in-depth exploration of psycholinguistic findings relating to language learning, the shape (i.e. nature and properties) of the mental lexicon, and language retrieval and production, the key linguistic features that emerge as having an impact on word association are summarised and reviewed in this Chapter. A large number of variables influencing language processing have been identified in psychology and psycholinguistics such as word frequency, age of acquisition, cohesion, lexical category, contextual variation, chunking (Divjak & Caldwell-Harris, 2019, p. 71), spelling-to-sound consistency, imageability, semantic richness, orthographic length, phoneme length, syllable length, number of morphemes, syntactic class (Balota et al., 2012, p. 90). Factors such as recency, the probability of occurrence – as indicated by previously experienced occurrences –, reliability – as indicated by the previously experienced ratio of correct interpretations to misinterpretations –, and context – as indicated by previously experienced co-occurrences – are further crucial for linguistic learning processes (Ellis, 2006, pp. 5-6, 15). In more specific terms, subfactors such as salience, prototypicality, generality, and redundancy have been found to affect learning alongside external influences impacting the mental capacity of the learner. Examples for this are factors such as automaticity, blocking, overshadowing and transfer (Gries & Ellis, 2015, p. 229). This is noteworthy with regards to the present project since it implies that the statistical probabilities of collocations present in the language we are surrounded by and exposed to (as represented by a substantial and balanced corpus) will have a direct and significant impact on the mental organisation of linguistic knowledge.

The picture is, however, yet more complex since the different factors listed here can interact with one another in non-trivial ways; age of acquisition effects can, for example, vary depending on frequency effects as well as lexical category membership (Tribushinina & Gillis, 2017, pp. 23–24). While it is not viable to observe and control for all of these effects comprehensively in a single study, factors of particular importance for a given research question can be selected and used as an initial point of exploration. For this reason, more immediately measurable factors like frequency,

cohesion, probability of occurrence, lexical category, context, chunking (this, essentially, refers to MWE membership or "collocativity"), and semantic richness are explored as part of this thesis.

A summary of research regarding these factors aids methodological decision making on two levels: Firstly, it describes the constraints of psycholinguistically motivated methods of collocation extraction, particularly AM selection (see Chapter 3.2), and contextualises the results from collocation analyses. Secondly, the psycholinguistic research discussed above provides pointers for methodological guidelines and best practices for corpus construction more broadly.

Psycholinguistically Motivated Constraints for AM selection

The most obvious psycholinguistic feature that plays a role in measuring association, and a feature that is highly relevant in all three sub-domains of language use (learning, memorisation, and production), is frequency. Language learning studies show that low frequency correlates with the non-adoption of linguistic features present in the target language (Ellis, 2006, p. 19), and frequency has been found to be the best predictor in word recognition tasks (Balota et al., 2012, p. 101). On the basis of Statistical Learning mechanisms, continuous re-activation of a high frequency linguistic item is expected to lead to entrenchment (Brysbaert et al., 2017). This is true, albeit to a different degree, in the context of a speaker's native and second language (ibid.). Frequency itself is not a homogenous measure, not just the frequency of co-occurrence, but also the frequency of a category a word belongs to, and the frequency of a cue within that, have been found to be influential (Tribushinina & Gillis, 2017).

The term *word frequency effect* refers to the observation that high-frequency words are processed more efficiently than low-frequency words (Brysbaert et al., 2018). Previous research also points to individual differences regarding said frequency effects (R. A. I. Davies et al., 2017), meaning that they present at different word frequency ranges for people with different degrees of language exposure. When word recognition is analysed, frequency of occurrence is one of the strongest predictors of processing efficiency with high-frequency words being processed faster than low-frequency words. Equally, frequency has been shown to positively affect memory performance with higher recall values for higher frequency words. Interestingly, recognition tasks which involve discrimination of previously shown stimuli from lures show the opposite effect – low frequency words lead to better performance in a recognition task (Yonelinas, 2002); this can be partially explained via inhibition effects. In a language learning setting, low frequency, low contingency, and low probability of feature selection are key factors that are observed to correlate with the non-adoption of linguistic features present in the target language (Ellis, 2006, p. 19). Lastly, frequent terms have also been found to be more robust to language production errors (Balota et al., 2012,

p. 92). For this reason, frequency is a key factor in determining which units should be represented in psycholinguistically motivated methods of collocation extraction and association measures that show frequency effects are particularly well-suited to achieve this goal.

However, while frequency overall is an important predictor for language acquisition, looking at word frequencies alone cannot possibly account for the full complexity of language perception and production (Tribushinina & Gillis, 2017, p. 19) and a wealth of other factors such as cohesion, exclusivity, translational probability, and connectivity are also both influential and measurable.

Looking beyond raw frequency, co-occurrence and contextual embedding has been found to influence behavioural outcomes more than frequency alone (Divjak & Caldwell-Harris, 2019, p. 66). In the field of child language development, transitional probabilities have been found to be a predictor for word segmentation learning (Rebuschat & Williams, 2012, p. 2; Teinonen et al., 2009, p. 6). Transition probability effects have also been found in adults (Perruchet & Poulin-Charronnat, 2012, p. 119), e.g. in reading time studies (Smith & Levy, 2013). Transitional probability can be approximated using a specific AM which focuses on forwards predictability: $\Delta P_{forward}$ (see Chapter 3.2.3). For the purpose of exploring languages other than English as well as in approaches that are not collocational in nature, backwards probability, $\Delta P_{backward}$, may also be relevant.

Looking at further cognitive mechanisms that have been found to strongly impact cue retention in the context of language acquisition, contingency or cue reliability/exclusivity (Gries, 2012, p. 49; Tribushinina & Gillis, 2017) plays a major role. This factor is so important that the name of the entire field of contingency learning is coined by it; the idea here is that exclusivity and repeated exposure lead to more robust language learning and storage in the ML. This effect has even been described as more dominant than frequency effects overall (Ellis & O'Donnell, 2014, p. 78) and dominant reliability effects have been found in second language acquisition contexts (Ellis et al., 2015, pp. 357–358). Specific AMs such as (log)Dice use this concept in order to approximate the associative strength between words (see Chapter 3.2.3).

After employing word association measures that can optimise for as many known psycholinguistically relevant factors as possible, connectivity measures can be extracted from the resulting network. This is a major contribution the application of network approaches to linguistic data can make in the area of interpretable psycholinguistic models of language processing. Connectivity is crucial since phenomena such as spreading activation (Collins, 1975; Siew et al., 2019, p. 9) are known to chiefly influence linguistic recall (Pecina, 2010, p. 141; Stella et al., 2018, pp. 7–8). Chunking lowers processing effort - the strongest collocational (dispersion-robust) units

would be the best candidates for that (Divjak & Caldwell-Harris, 2019, p. 71). What is more, the influence of connectivity cannot be overstated since evidence for the theory of preferential attachment (Mak & Twitchell, 2020, p. 1067) shows that a strongly interconnected word is more likely to gain even more connections with time, further easing recall. To a degree, connectivity can thus also be seen as a proxy for lower overall cognitive load, a factor that plays a significant role in language processing (H. Chen et al., 2018, p. 17; Navarro et al., 2020) but is impossible to quantify directly without brain imaging data. Beyond this, connectivity is also known to be subject to complex and non-linear age effects (Zortea et al., 2014, p. 90) making future explorations of idiosyncratic large linguistic networks using the pipeline provided in this thesis an ideal testbed for examining the aging effects on the mental lexicon.

A further fundamental factor impacting AM generation is positionality. Retaining directionality of collocations, i.e. separating frequencies of occurrence found for *I*, *am* or *am*, *I,* is essential for psycholinguistically plausible collocation extraction. Event-Related-Potential studies show that recently perceived syntactic features influence perception and processing of following words (Yamada & Neville, 2007, p. 177). This phenomenon only exists linearly meaning that words which were uttered in the beginning of a sentence are impacting the processing of following words, whereas the following words cannot influence the initial perception of a previously uttered word. Sentence comprehension studies strengthen this point by illustrating that sentence comprehension is procedural (Szewczyk & Schriefers, 2018, p. 665). This evidence leads to the methodological imperative of preserving directionality when extracting psycholinguistically plausible word associations.

A key question that cannot currently be answered on the basis of existing research is the degree to which these subsystems interact; it is not possible to assign a clear-cut percentage of influence to each of these parameters. The development of different methodological approaches that target the variables of interest here (frequency, exclusivity, centrality, directionality, probability of feature selection) individually, however, presents an immediate way forward. The experimental component of this thesis entails a detailed juxtaposition of 15 different corpus-based networks that are generated on the basis of different AM approaches and combinations thereof in the quest of exploring how these systems interact.

Corpus Construction

After developing specific suggestions for AM identification grounded in psycholinguistic research, it is equally crucial to consider broader guidelines for corpus construction; these are presented herein. First, when working with general and reasonably large corpora, single occurrences should

be disregarded resulting in a minimum co-occurrence frequency of two. This is the case since it cannot be attested if one-off events lead to entrenchment (Ellis, 2006, p. 19), and thus to being added to the mental lexicon. Secondly, it is crucial to bear in mind that all layers language can be experienced through (e.g. orthographic, phonological, morphological, syntactic, semantic etc.) are relevant to Statistical Learning processes. This directly informs corpus collection and pre-processing measures since it makes it desirable to retain attributes relating to the original spelling should the corpus analysis require stemming/lemmatisation, etc. and makes a strong case for systematic POS-tagging and syntactic tagging. This thesis is largely working with lemmatised representations of words for comparability with the psycholinguistic dataset; ideally, retention of all of these layers would be desirable. Thirdly, findings presented in this Chapter further indicate fundamental differences between auditory and visual Statistical Learning (Sandoval et al., 2017, pp. 10–11; Vitevitch & Goldstein, 2014, p. 143). This could result in differences in networks based on spoken data (e.g. the Spoken and Written-To-Be-Spoken subgenres of the BNC 2014) and networks based on language consumed in a written format (e.g. the Academic and Literature subgenres of the BNC 2014) and has to be carefully documented in the corpus construction process. Lastly, word associations are primarily representing semantic relations – this means using association measures that favour items from the lexical end of the lexico-grammatical continuum are expected to be particularly well-suited for this RQ. This also motivates choosing a sentence-span for generating AM values since semantic relations are not well captured by limiting the observation to the immediate vicinity of nodes of interest.

In conclusion, the exploration of psycholinguistic findings highlights the multifaceted nature of language processing and the numerous variables that influence word association. Key factors such as word frequency, age of acquisition, cohesion, lexical category, and contextual variation play significant roles in shaping the mental lexicon. Additionally, the interference generated via interactions between these factors along with subfactors like salience, prototypicality, and redundancy, underscores the complexity of linguistic learning and processing. This intricate interplay of variables necessitates a nuanced approach to studying language, emphasising the importance of frequency, cohesion, probability of occurrence, lexical category, context, chunking, and semantic richness in psycholinguistic research.

2.7 Graph Theory

Having explored the linguistic framework, definitions of the subject of study, collocations, and psycholinguistically relevant factors for extracting them, this chapter marks the introduction of graph theory for large linguistic networks which serves as the basis for the majority of later data visualisation and analysis. Graph theory is a mathematical field of studies encompassing the generation of abstract graphs and an extraction, description and categorisation of their properties (Biemann, 2012, p. 19). Broadly speaking, graphs consist of two types of information: information on the items in the dataset of interest – referred to as nodes – and the relationships connecting said nodes – referred to as edges (Biemann, 2012, p. 20). Based on these two types of information are two theoretical lines of thinking: network science and connectionism. While network science is focused on the overarching structure and its influence on specific processes (and therefore heavily node-based), connectionism is edge-based and aims to investigate or model processes directly (Castro & Siew, 2020, p. 3). In linguistic terms, connectionist approaches would investigate specific language learning or word retrieval mechanisms, whereas network science approaches would investigate the overall structure of linguistic constructs such as collocational relationships and the word associations in the present thesis. The focus of this thesis lies on a network-scientific approach. Most evaluations of graph theoretically relevant properties conventionally happen on a macroscopic level (Thurner, 2009, Chapter 2.1).

Applying graph theoretical concepts to linguistic data is worthwhile since the generated networks can be compared to other datasets which can modelled using networks and classified according to their abstract properties which enables a richer understanding of the interplay of large, complex, and dynamic systems such as language. The datasets that serve as the basis for the can furthermore be subdivided into meaningful groups, i.e. according to genre or mode, and then graph theoretically analysed and compared to each other in order to extract structural differences (Bales & Johnson, 2006, p. 451). Chapter 4.2.1.4 of this thesis implements this approach and entails an in-depth description of a model pipeline that serves to contrast word association and collocation networks.

Over the years, different branches of graph theory have emerged, most prominently quantitative and classical graph theory. The major difference between these two approaches is their primary focus: classical graph theory is mainly descriptive, primarily aiming to present graph decompositions, embeddings, structural properties, and characteristics whereas quantitative graph theory takes a measurement approach and aims to quantify structural information of networks with the possibility of going beyond concrete evidence and employing statistical models (Dehmer et al., 2017, 575–576). Examples of methods that belong to the toolkit of quantitative graph theory are comparative graph theory which aims to assess how structurally similar a number of networks are

and graph characterisation which is used to describe networks on a meta-level, e.g. via entropy measures. Since the main focus of this thesis lies on comparing psycholinguistic and corpus-based networks, quantitative graph theory is of particular interest for this endeavour and a range of relevant graph characterisation properties are described further in Chapter 2.7.1.

Traditionally graph theoretical methods have been used to explore a wide range of different types of data in fields as varied as bioinformatics, neuronal networks modelling, and computer science (Bader & Hogue, 2003; Dehmer & Emmert-Streib, 2009; Dimitropoulos et al., 2009). In the past 20 years, graph theory has also been incorporated in different branches of social science research. In an early review of articles with a focus on semantics that also involve graph theoretical explorations, Bales and Johnson (2006, p. 451) find that the majority of this very early research investigates real-world networks based on linguistic databases such as corpora and dictionaries thus laying the foundation for the incorporation of graph theoretical methods into the field of corpus linguistics.

A range of different linguistic features can serve as the basis for linguistic network constructions such as semantic, psycholinguistic, lexical, phonological and orthographic features. (Trautwein & Schroeder, 2018, p. 12), and all kinds of co-occurrences of linguistic elements such as words or phrases (Biemann, 2012, pp. 40–41). Within these broad types of networks, different metrics act as the basis for edges that connect individual items such as Association Measure values for co-occurrence networks, association weights or reaction times for psycholinguistic networks, the number of shared characters or spelling variants in orthographic networks, and homophones and near-homophones in phonological networks. From this point onwards, the focus strictly lies on psycholinguistic and AM-based co-occurrence networks since other kinds of networks, linguistic and otherwise, lie outside of the scope of the present thesis.

When considering the use of this new application of graph theory to large amounts of (corpus)linguistic data, it is essential to recognise that its effectiveness hinges on the underlying statistical foundations. Specifically, in the present context, these statistics represent word co-occurrences and word association weights. The meticulous selection of association measures and their corresponding parameters assumes paramount importance. Consequently, Chapter 3.4 of this thesis extensively discusses these critical considerations.

Moreover, it is equally crucial to bear in mind that any statistical analysis inherently constitutes an argument (Hodges, 1996). In the present thesis, this argument asserts plausible structures within linguistic cognitive processes and implies possible parallels between mental language and communicative language. This claim rests upon several factors: the data selection, alignment with

existing psycholinguistic experimental evidence, the use of particular graph theoretical parameters for evaluating similarities, and their nuanced interpretation. The outcome of this large-scale exploration of linguistic networks through graph theory aligns with Hodges' classification scheme which categorises it as an active hypothesis generation argument.

2.7.1 Graph Theoretical Parameters of Interest

This chapter provides a comprehensive examination of graph theoretical parameters with potential for application to linguistic networks. The psycholinguistic relevance of a wide range of these parameters has been tested empirically as described in Chapter 3.2. Additionally, further parameters that exhibit strong potential but have yet to be thoroughly investigated in psycholinguistic research are also discussed.

A number of parameters that play an important role in quantitative graph theory and that serve as the basis for comparisons of the corpus-based and psycholinguistic networks[10] in this thesis are presented in Table 1 below. One of the major contributions of this thesis is the accessible and clear framing of these parameters and their utility in a linguistic context. The features presented here are relevant on a macro-, meso- or microscopic level and explored further in the respective subchapters. It is common for parameters to be analysed on both the macro and the micro level; this is the case since averages derived from node-focused metrics such as degree or centrality can be harnessed to characterise the overall network on a macro-level. A 'translation' into linguistic terms for many of these parameters is provided in the table below, and practical examples showing what these parameters look like in a linguistic context are provided in Chapter 4.3 where the results of this thesis are presented.

Table 1: Graph Theoretical parameters and their possible levels of analysis; corresponding linguistic/collocation-based parameters provided in brackets. When no correspondence is provided the explanations are non-trivial and provided in the relevant sub-sections below instead. No meso-category included here since this is reserved for clusters as discussed in Chapter 2.7.1.2.

Parameter	Macro	Micro
number of nodes (~ number of words)	☑	☐
number of edges (~ number of collocations)	☑	☐
number of connected components (~ number of fully connected collocation networks)	☑	☐
number of strongly connected components	☑	☐

[10] Corpus-based networks are here understood to be networks generated on the basis of corpus-data, in the case of this thesis collocations, whereas psycholinguistic networks are networks generated on the basis of psycholinguistic data, in the case of this thesis word associations.

(~ number of fully and strongly connected collocation networks)		
number of self-loops (~ number of self-collocates)	☑	☑
characteristic path length; the average of the shortest path (~ average number of words on collocational paths that connect two collocates)	☑	☐
diameter; the longest length between two nodes (~ highest number of other collocating words between collocates)	☑	☐
network radius; the minimum of non-zero eccentricities – i.e. maximum non-infinite lengths of a shortest path – in the network (~maximum non-infinite number of words on collocational paths that connect two collocates)	☑	☐
degree; number of target nodes connected to the source node (~ number of collocates)	☑	☑
network density; the normalised average number of neighbours (~normalised average number of collocates)	☑	☐
network centralisation; the centralisation of the network connectivity	☑	☐
closeness centrality; measures the potential speed of information spread	☑	☑
betweenness centrality; measures how relevant a node is for overall connectivity	☑	☑
clustering coefficient; probability for a given node's neighbours to be interconnected amongst themselves (~probability for a given word's collocates to collocate with one another)	☑	☑
eigenvector centrality; measures the influence of a node. The score of each node is proportional to the sum of the centrality of its neighbours	☑	☑
transitivity; relative number of triangles in the graph compared to the total number of connected triples of nodes	☑	☐

2.7.1.1 Micro-Level

The first level to be explored here, the micro level, contains parameters that characterise individual nodes, in the present context thus individual words or lemmas. The parameters discussed in greater detail here are degree centrality, closeness centrality, betweenness centrality, eigenvector centrality, clustering coefficient, and self-loops.

Degree Centrality

A core graph theoretical parameter to be examined is the degree *k* which represents the number of target nodes connected to a source node (H. Chen et al., 2018, p. 8). In a directed network it is possible to measure both the in-degree and the out-degree of a node resulting in a Degree Centrality (DC) measurement. When applied to collocation data, degree explorations highlight systematic commonalities of words that have an exceptionally large number of collocates which is relevant for assessing possible biases towards certain types of collocation due to the selected AM. When applied to word association data a list of words with high *k*-values shows which words exhibit the most associative connections to other words; the properties of these words could shed light on what makes a word particularly associatively rich which heavily influences the shape of the ML.

The out-degree of a node in cue-association networks is of particular interest since it reflects the number of distinct associations formed based on a given cue. Similarly, in collocation networks the

out-degree of a node represents how many distinct other words succeed the node. Set size has been found to serve as a subtle proxy for quantifying the features associated with concrete nouns. Pexman et al. (2003, p. 844) report a facilitation in semantic processing for nouns with a higher number of features such as *deer* as compared to nouns with less features such as *curtains*. Furthermore, these feature-rich words were found to facilitate reading by activating related features. It has been hypothesised that this effect might stem from spreading activation, either directly on a word-level or through the activation of sensory and conceptual information.

To illustrate the nature of these features the feature list for the high feature noun, *deer*, is provided as listed in McRae et al. (1997, p. 112) whose research the Pexman et al. study builds on:

> *deer <is herbivorous><has antlers><lives in the woods><lives in the wild><a mammal><an animal><is brown><has hooves><has four legs><has fur><has legs>*

Upon examining this example, it becomes evident that many of the words contained in these features – such as *wild*, *antlers*, *woods*, *animal* etc. – would likely emerge as responses when prompted with the cue *deer*. Furthermore, these terms can also be expected to contribute to collocational representations. Nelson et al. (1987, pp. 133–134) also investigate category size and find differences in processing depending on set size (number of features) when asking participants to indicate whether or not a cue belongs to the same category as a previously seen target, thus indicating that the degree of a word impacts its mental processing.

By conducting corpus linguistic analyses focused on nouns with substantial set sizes, it may be possible to uncover terms that exhibit enhanced semantic processing efficiency and immediate interpretability. One exemplary area of immediate applicability of the set size metric is its employment for identifying the most easily accessible nouns in a reference corpus in a language teaching context; the pipeline developed for this thesis enables identifying and exploring these words.

Degree Centrality has further been explored in the context of adult free association norms (Mak & Twitchell, 2020, p. 1067). The outcomes of three sub-experiments indicate that words were more likely to be recalled if they were initially associated with a cue word with a high DC, meaning it was more interconnected within the network, supporting preferential attachment models (see more on the theoretical background of preferential attachment in Chapter 2.5.2).

Closeness Centrality

The next metric of interest, Closeness Centrality (ClC, Sabidussi (1966)) is measured on the basis of the inverse of the average shortest path between the source node and all other nodes (Metcalf & Casey, 2016; Siew et al., 2019, p. 5).

$$ClC(n_0) = \frac{n-1}{\sum_{n_1 \in G} d(n_0, n_1)} \qquad \text{Equation 1}$$

It signifies how quickly information from the source node could spread through the network. In collocation terms, a word with a high closeness centrality signifies a particular capability of the word at hand to connect different collocational contexts with one another. In word association networks, a high ClC word indicates that a word is crucial for connecting associations and thus impacts the overall information flow.

A small sample network is employed throughout this chapter to illustrate different types of centrality measures. In Figure 9, the highest ClC is exhibited by node 5. This is due to the fact that the sum of shortest paths leading to all other nodes from 5 is the highest (14), which leads to a closeness centrality of 0.64 since $ClC('5') = \frac{9}{14} \approx 0.64$.

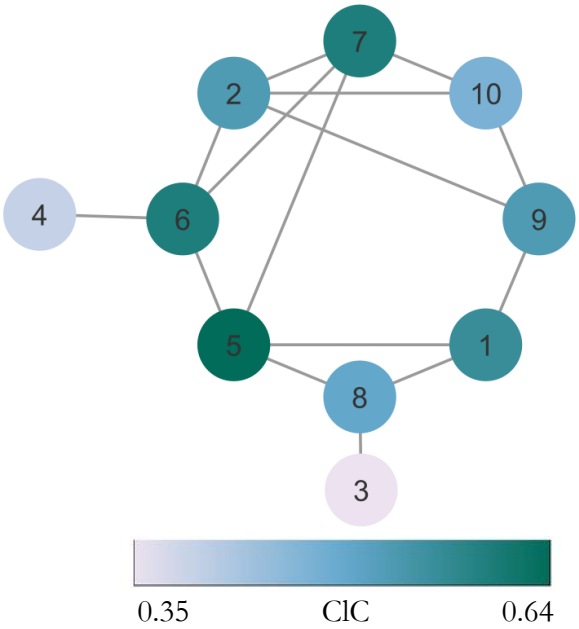


Figure 9: Closeness centrality in a small, unweighted model network.

The application of closeness centrality can provide insights into the interconnectedness and influence of linguistic items on the entire network. High closeness centrality indicates that a word has close relationships with many other words, reflecting its central role and influence within the

network. However, it is important to note that closeness centrality should not be interpreted as a direct proxy for information flow since it is based on shortest paths, but not necessarily most frequently taken paths (Borgatti, 2005, p. 59). Instead, closeness centrality highlights the overall influence and connectivity of a word; in social network research this metric has been used to highlight the degree to which different disciplines influence each other most strongly via co-citations (Ni et al., 2011).

Betweenness Centrality

Betweenness Centrality (BC; Freeman (1977); Anthonisse (1971), Equation 2) on the other hand is based on the sum of shortest paths (σ) between any two nodes n_1 and n_2 that the node of interest n_0 lies on divided by the total number of shortest paths between any n_1 and n_2 (H. Chen et al., 2018, p. 7).

$$BC(n_0) = \frac{\sigma_{n_1,n_2}(n_0)}{\sigma_{n_1,n_2}} \qquad \text{Equation 2}$$

This signifies how much the interactions between all other nodes depend on the node in question and can therefore be used to flag up candidates for cluster-central and long-range nodes. Betweenness centrality is very important in a linguistic context since it can identify what has been termed 'long-range nodes' (Bordag, 2003, p. 330) or hubs (Veremyev et al., 2019, p. 5) and it has been employed as a semantic salience marker in collostructional analyses (Dekalo & Hampe, 2017, p. 165). These nodes represent shortcuts between clusters of nodes through their high connectedness (Bordag, 2003, p. 330). They are of specific interest to linguists since they have previously been identified in co-occurrence networks and they have been reported to be unevenly represented based on the word class of the node, with common verbs, articles and function words being the most common word classes of long-range nodes (Bordag, 2003, p. 330). Long-range nodes are interesting beyond these observations since they can act as hubs that connect different contexts within a corpus or psycholinguistic network. Long-range nodes might furthermore show potential in meaning disambiguation (Nazar, 2011, p. 163) since different meaning groups present in clusters can be identified through analyses of the fragments of a cluster after the removal of a connecting long-range node.

A similar, but not identical concept is a bridge (also referred to as isthmus). This describes a node that would, upon its removal, fragment the graph into multiple components (Oxley, 2014). Bridge nodes are thus extreme forms of long-range nodes; applied to collocational data these represent words that present the only contextual connection to a distinct other set of collocations, indicating that they might exhibit a key role for topic shifts.

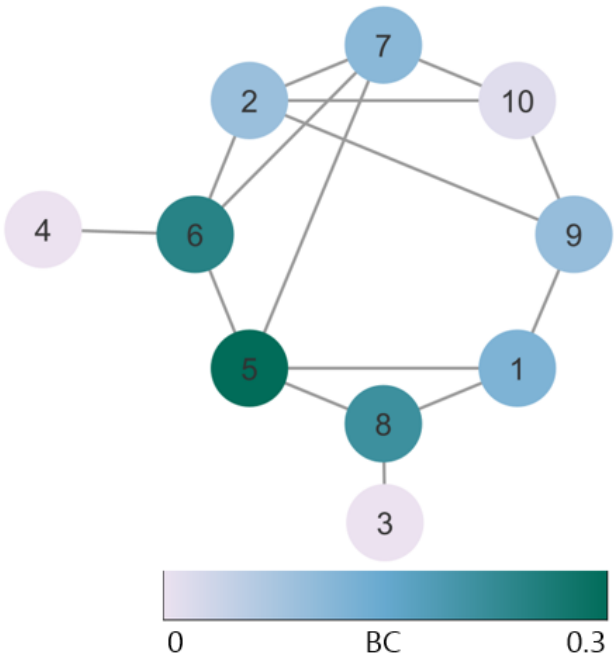


Figure 10: Betweenness centrality in a small, unweighted model network.

In the simplified graph above, 3 and 4 have the lowest BC since no shortest path between two other nodes traverses through them. 5 has the highest BC (0.3) since it lies on the most shortest paths between nodes. It effectively creates a shortcut between the top and bottom half of the graph via (5,7) and lies optimally between two high density areas (4,6 and 8,3). In this thesis, the default algorithm embedded in NetworkX (Hagberg et al., 2008) is employed to compute BC as efficiently as possible. What sets Betweenness Centrality apart from Closeness Centrality is that it measures how immediately important the word is to the remaining network whereas Closeness Centrality measures how important a word is to the efficiency of topic traversal or information spread.

Eigenvector Centrality

The last centrality measure employed as part of this thesis is Eigenvector Centrality (EC; Bonacich (1972), Pradhan et al. (2020). EC, similarly to BC and ClC, is a measure of the influence of a node in a network. Here, relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. In a linguistic context, a high EC indicates that a word is connected to many other influential words who themselves display high scores. In a collocational context, the 'long-term influence' of a word is thus expressed via its EC value. EC is particularly interesting since it links conceptually to preferential attachment (Castro & Siew, 2020, p. 16; Mak & Twitchell, 2020, p. 1059; Sheridan & Onodera, 2018, p. 1), the tendency of influential words to be directly connected to other influential words.

Clustering Coefficient

A further property of interest is the clustering coefficient (ClCoef). This measure represents the ratio of existing connected edges between neighbouring nodes to possible connected edges between neighbouring nodes (Borge-Holthoefer & Arenas, 2010, p. 1268; Steyvers & Tenenbaum, 2005, p. 46; Watts & Strogatz, 1998, p. 441); it can therefore also be expressed as the probability for two nodes being neighbours. The clustering coefficient can take values from 0 to 1 (the latter would be present in a fully connected network). When applying this concept to linguistic data, the clustering coefficient signifies the extent to which words are forming collocational cliques. In simpler terms, clustering coefficient is a measure of how interconnected the collocates of a specific node word are. If the clustering coefficient is high, it indicates that the collocates are strongly clustered and belong to the same tight-knit context. A low clustering coefficient, on the other hand, suggests that the collocates branching from a node have few or no shared collocational links, representing distinct contextual embeddings. This concept can be applied to cue-association data, where clustering coefficient values measure the likelihood of responses given for the same cue being given in response to each other. In other words, high clustering coefficient values indicate a shared associative embedding, while low clustering coefficient values suggest a greater variety of

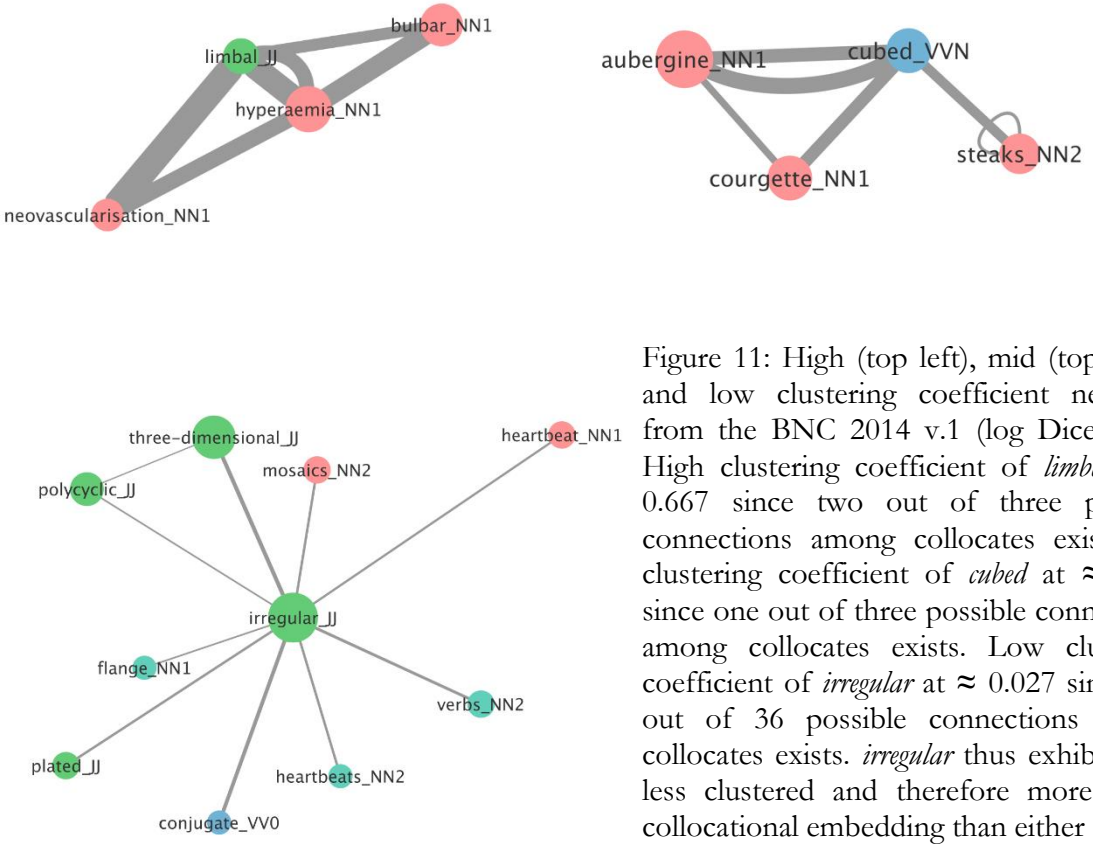


Figure 11: High (top left), mid (top right), and low clustering coefficient networks from the BNC 2014 v.1 (log Dice ≥ 10) High clustering coefficient of *limbal* at ≈ 0.667 since two out of three possible connections among collocates exist. Mid clustering coefficient of *cubed* at ≈ 0.333 since one out of three possible connections among collocates exists. Low clustering coefficient of *irregular* at ≈ 0.027 since one out of 36 possible connections among collocates exists. *irregular* thus exhibits as a less clustered and therefore more varied collocational embedding than either *cubed* or *limbal*.

associations. This concept is illustrated in Figure 11 which is taken from the log Dice ≥10 BNC 2014 collocation network. The figure displays sample nodes with high, mid-high, and low clustering coefficient values, along with their first-order collocates.

Empirical findings from explorations of clustering coefficients provide a strong argument for using graph-theoretical parameters in psycholinguistic research. Goldstein and Vitevitch (2014) find that words with a higher phonological clustering coefficient are beneficial for long-term learning. This fits in with the view that spreading activation in closely connected parts of the mental lexicon leads to a strengthening of the learned word, possibly since low connectivity around the learned word might result in little back-flow of spreading activation and rather a dispersion through the rest of the network. Further research also indicates that different linguistic processes are affected by clustering coefficients in an even more nuanced way: High clustering coefficients have been found not to benefit, but to inhibit visual word responding (Yates, 2013, p. 1653). The same underlying process that increases learnability, a high spreading activation amongst all neighbours, might result in this negative effect since the word in production might compete with closely related neighbours (Karuza et al., 2016, p. 632). This is, of course, purely hypothetical at this stage and ultimately needs to be tested neurolinguistically before causal relationships can be assumed with reasonable certainty.

Self-Loops

Lastly, self-loops are briefly mentioned. This metric is binary and indicates whether or not a node is connected with itself. In collocation networks, words which are self-loops collocate with themselves (e.g. *hear*, *hear*). This is very frequent when employing sentence-wide collocation windows as is the case in this thesis. Existing small-scale collocation visualisation tools such as GraphColl (Brezina & Platt, 2024) often struggle with the visualisation and ranking of self-loops since this forces words to appear twice in the network, once as the centre of the graph and once as a collocate. As can be seen in Figure 11 when examining the node *steaks_NN2*, the visualisation chosen here evades this issue by indicating the self-loop via a circular edge. In word association networks, self-loops are expected to be found less frequently since they would have to result from a participant directly responding to the cue word with the word itself.

2.7.1.2 Meso-Level: Clusters

Clusters, here defined as groups of nodes with an exceptionally high internal interconnectivity, are introduced as the key graph theoretical construct on the meso-level. In existing network literature, dense clusters are also referred to as cliques (Veremyev et al., 2019). Clusters are of particular interest since they can logically be considered optimal for language processing due to shortening

access paths between different clusters (Trautwein & Schroeder, 2018, p. 3) and they might be the key to enabling swift language comprehension and production. A large number of different extraction methods exist for identifying clusters from a larger network; the one examined in greater detail as part of this thesis is Molecular Complex Detection (MCODE, Bader & Hogue, 2003) clustering.

MCODE, as the name suggests, stems from the field of molecular biology and has been developed to harness density connectivity in order to identify particularly densely connected regions in a network as clusters. Figure 12 depicts the steps taken when extracting MODE clusters. Essentially, each node in is assigned a weight based on its degree centrality. Using the highest weight node as the starting point, the algorithm then traverses the network to identify densely connected regions. A cluster is formed by including nodes that are connected to the starting node and have a weight above a certain threshold, in this thesis the threshold 20% of the degree centrality of the connecting node has been chosen. Once all remaining connected words lie outside the threshold the algorithm moves on to use the next highest degree centrality node as a starting point until the entire network is clustered.

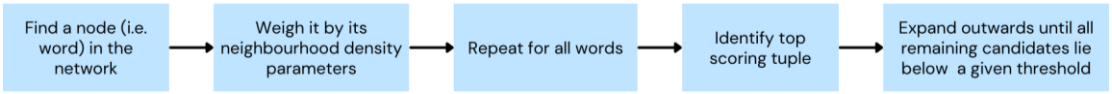


Figure 12: Simplified representation of the MCODE algorithm.

Findings from linguistic cluster analyses show promising results in terms of increased objectivity and precision as compared to conventional manual analyses (Gries & Stefanowitsch, 2010, p. 86). Clusters consist of a number of neighbours, i.e. nodes that share an edge (Steyvers & Tenenbaum, 2005, p. 45) and they can, in theory, be maximal when every node is fully connected to every other node in the cluster. This is rarely the case in linguistic co-occurrence networks. Insights from qualitative analyses of linguistic clusters may provide new information as to which linguistic units are commonly processed together (and might elicit one another) on the basis of psycholinguistic networks and as to which linguistic units are commonly produced and encountered together on the basis of text-based networks. Clusters can furthermore aid on a macro-level comparison of networks since each individual network can be characterised both in terms of the number of interconnected clusters it produces, as well as by the content of the emerging clusters. Lastly, clustering techniques show promise in applied linguistics as a starting point for refining and expanding existing graph-based word sense disambiguation methods as described in Klapaftis and

Manandhar (2008), both via allowing for analyses containing custom parts of speech other than nouns, and via the application of new clustering algorithms and dynamic visualisations.

Employing cluster analyses can provide a valuable addition to the conventional exploration of words that are connected via a specific number of other words; in collocation-based datasets these are commonly referred to as second-order collocates (connected through one shared neighbour and therefore collocates of collocates (McEnery & Brezina, 2019, p. 103)) and third-order collocates (connected through two shared neighbours and therefore collocates of collocates of collocates). In contrast to this approach which depends on the researcher's intuition as to which search word to choose and how many 'orders' of separation to explore, the clustering approach is based on structural properties emerging from the data itself. This means that strongly connected and highly exclusive n-gram structures can be identified without predefining *n*. Chapter 4.3 explores results from several clustering operations based on a word association network and the BNC 2014. These results illustrate that a particular strength of LLN clustering approaches lies in the possibility for labelling the individual clusters via their highest betweenness centrality nodes, thus avoiding opaque cluster labelling as is required when applying, for example, Principal Component Analyses.

2.7.1.3 Macro-Level

Thirdly, the macro-level of graph theoretical parameters is explored here. The largest number of graph-theoretical parameters is extracted to describe the overall network shape, and the macro level plays a pivotal part in the structural comparison between collocation networks and psycholinguistic word association networks. The properties to be explored are directedness, the number of nodes, edges, and (strongly) connected components and self-loops. Beyond these relatively simple measures, the characteristic path length, transitivity, density, diameter, and radius of the network can also be measured. Lastly, all properties explored in the micro-level are also represented on the macro level since their network-wide distribution can be explored and used to characterise network properties for later comparison.

A first parameter that greatly influences all other analyses is directedness. When constructing large linguistic networks, emphasizing the preservation of directionality is paramount. This principle extends from the recording of collocational relationships, a topic further elaborated upon in Chapter 3.2.4. Undirected networks do not allow for examining information flow and thus lack a core feature of complex, dynamic representations of language. Within the context of graph theoretical explorations of collocation networks, directed graphs thus play a pivotal role and are used wherever possible throughout this thesis.

This introduction to macro-level graph theoretical properties opens by first looking at reasonably simple measures: The total number of nodes in a network represents the total number of linguistic units present, in the context of the present study these are represented by collocates in the BNC 2014 that reach a certain AM rank on the one hand and cues and responses from free association tasks above a threshold on the other. The number of edges – which in represent weighted AM or stimulus-response links in this work– serves to describe the overall connectivity of a network. More refined measures such as characteristic path length, radius and diameter help explore the network shape in greater detail.

Characteristic path length is a measurement that captures the average of the shortest paths between individual nodes in the network (H. Chen et al., 2018, p. 6). A shortest path is defined as the minimum number of nodes that have to be passed through to reach a target node from a given source node. Characteristic path length in the present context is indicative of how heavily interconnected the respective network is and how large the average distance between certain collocates or items of the mental lexicon is. A large value implies a dispersed network with long shortest paths which might lead to a lesser disposition for swift topic shifts.

The diameter of a network is based on the longest shortest path between two nodes in the network (Peruani, 2009; Trautwein & Schroeder, 2018, p. 3). While this naming convention might make the concept seem contradictory, it entails identifying the shortest paths between any two nodes in the network and then identifying the longest one among these shortest paths across all nodes in the entire network. In a linguistic context, a small diameter indicates that the network is quite compact, with any word being relatively close to any other word. This allows efficient traversal from one concept to another and aids quick access of information. A large diameter suggests that the network is spread out, which could mean that individual topics are less closely interlinked, and some areas of the collocation/association network are more remote.

In a similar vein, the network radius (Gould, 2012, p. 36) is measured as the minimum non-zero eccentricity of any node. Eccentricity is defined as the longest distance from any node to any other node; the radius is defined as the minimum of these. In other words, the radius, defined as the minimum eccentricity, is the inverse of the diameter, defined as the maximum eccentricity. In the present context, a small radius implies that there is a centrally located word which acts as a hub for reaching other nodes quickly. A large radius, on the other hand, would suggest that even the most central linguistic units are still far removed from other units which could hinder efficient information flow.

It is important to note that diameter and radius can be measured on the basis of the edge weight rather than just the raw number of steps required. This is of particular interest since the edge weights for all networks analysed as part of this thesis are linguistically relevant: they either represent the weighted association strength identified via frequency of cue-association or AM strength. Presenting diameter and radius of a given network in combination therefore allows for characterising some of the most extreme path lengths and describing the network shape effectively.

Another core feature is the distribution of degrees over the whole network (Steyvers & Tenenbaum, 2005, p. 47). In a linguistic context, an even degree distribution means that there are no major differences in how well interconnected different words or stimuli are, other shapes of degree distribution would indicate that there are – as suspected when dealing with language – differences in how well interconnected certain words are. Degree distribution is an important factor for small worldedness and scale-free networks, classifications further explored in Chapters 2.7.2 and 2.7.3. When examining uneven degree distributions in large linguistic networks, the natural next step is then carrying out an analysis of words with exceptionally high degree as described in Chapter 2.7.1.1 with the aim of identifying latent patterns that cause these differences via correlating factors such as POS membership, special meaning, length of word, total number of occurrences of a word, amongst other factors.

Next, density is defined as the normalised average number of neighbours in the network (Jun Dong & Horvath, 2007, p. 2). A density value of 0 indicates that there are no edges between nodes, a density value of 1 would be achieved if each node is connected to each other node. Self-loops are not taken into consideration when computing density. For linguistic networks it is recommended to present the network density alongside the degree distribution to further illustrate the overall interconnectedness of collocates or stimuli/responses.

The last parameter that describes interconnectedness is network centralisation. Network centralisation relies on measuring the sum of the differences in the centrality of the most central node versus all other nodes; this sum of differences is then compared to the largest possible sum of differences and results in a ratio that may take values between 0 and 1. Network centralisation therefore effectively measures how uniform the connectivity of the whole network is; a network with one distinct centre or focal cluster that connects outwards to all other nodes – often somewhat resembling a star – displays a high network centralisation (approaching 1) whereas a network with evenly distributed edges and no particular centre displays a low network centralisation (approaching 0). In a linguistic context, a high network centralisation means that there is a distinct centralised node or cluster that connects all of the collocations or stimuli. There are a range of centrality

measures that can be used as the basis for calculating network centralisation such as closeness centrality and betweenness centrality.

Taken together degree distribution, density, and network centralisation allow for characterising and comparing the most relevant edge properties of the networks at hand and thereby identifying how interconnected the networks are, how the differences in connectivity are distributed, to what extent the network tends to contain nodes with different degrees, and to what extent the network is clustered around a core element.

Other measures that have been discussed on a micro-level in Chapter 2.7.1.1 are also of interest on a holistic level; two of these are betweenness centrality. Nodes with a high betweenness/ closeness/ eigenvector centrality play a special role as long-range nodes in the network. An examination of the overall number and types of long-range nodes present in a graph can be employed for graph characterisation. In a linguistic context, this is of special interest since the distribution of long-range nodes indicates how efficiently topics are connected, and how many words are especially structurally important to the overall design of the network. Similarly, the overall clustering coefficient values present in a network are therefore a measure for how structured a given network is on the whole (Deyne et al., 2016, p. 56).

2.7.2 Scale-Free Properties and the Power Law Model

After looking at individual graph theoretical features, it is also important to focus on overarching characterisations of the shape of entire networks; metrics that are of relevance here are scale-free properties and, closely connected to that, power law models. In this chapter, an exploration of the utility of these metrics is provided alongside a definition and explanation of these graph theoretical concepts aimed at a linguistic audience and a brief discussion of the limitations of said approach.

Scale-Free Properties and the power laws in degree distributions allow for comparisons between the structural design of different systems on a meta-level – a core feature necessary to contrast and analyse different kinds of linguistic networks such as collocation networks and word-association networks as well as networks observed in other domains. This is of interest since power laws govern distributions of a wide variety of real-world networks and this phenomenon has been investigated in many different disciplines as varied as linguistics (see below), human geography (i.e. populations of cities), electrical engineering (i.e. power outage severity) and geophysics (i.e. earthquake intensities) (Clauset et al., 2009, p. 663).

The investigation of power laws in linguistic networks is relevant to (psycho-)linguistic research for two main reasons: Firstly, it allows for observing practical limitations of linguistically relevant

cognitive processes regarding memory use and storage of lexical information (Morais et al., 2013, p. 143). If there were no such limitations, fully connected component where every word is immediately connected to every other word in a maximally large lexicon and word retrieval is instant regardless of the overwhelming size of the network would be expected. The second relevant insight power law analyses can provide regarding the structure of linguistic information storage concerns growth mechanisms. Not every degree distribution can result from any growth process - this leads to the development of testable hypotheses as to possible language learning mechanisms after power laws have been identified for a particular linguistic network.

Now that the utility of these concepts has been demonstrated, a more detailed definition and explanation of the related concepts of scale-free properties and power laws follow. In essence, scale-free properties are present in a network when the distribution P(k) of the node k's degree, i.e. the probability for the node k to have a certain degree, is governed by a power-law (see Figure 13 and H. Chen et al. (2018, p. 5)). More specifically, in linguistic networks this means roughly following the shape a power law defined by $k^{-\alpha}$ with $2 \leq \alpha \leq 3$ (Siew et al., 2019, p. 6). A degree distribution of this shape – specifically with an α of 2 or higher – consequently proposes that the distribution stretches towards infinity (or is *scale-free*).

A relationship similar to the prototypical shape displayed in Figure 13 reflects the fact that there is a high number of nodes with a limited number of connections to other nodes such as specialist words that only occur in certain contexts, and a small number of nodes that are heavily interconnected – these are likely to be mostly grammatical units and other words that act as hubs through connecting a large number of other units. Although scale-free networks are very tolerant to random removal of edges, if deletion is directed to the most connected edges the network gets broken into pieces (Ferrer-i-Cancho & Solé, 2001, p. 2265). It is important to remember that the displayed plot is an ideal representation that follows a specific power law perfectly and therefore only helps to identify the most probable node/degree distribution of a real datapoint. Real data cannot be expected to follow this prediction with perfect accuracy and the suitability to use one type of power law over another needs to be researched carefully before generalising statements can be made.

After defining this concept some findings from existing studies are presented to contextualise the observations in this thesis. Rank-frequency distributions in English corpora have been found to underlie power laws both for individual lexical items in the shape of Zipf's law (Zipf, 1949) and n-grams (Biemann, 2012, p. 40). Power laws in these word co-occurrence networks have, however, been reported to be language specific since generalisations regarding the power laws that govern German, Icelandic and Italian co-occurrence networks do not closely resemble power laws present

in the equivalent BNC 2014-based English networks (Biemann, 2012, p. 48). Beyond corpus data, degree distributions in a variety of real-world networks as well as semantic networks based on word associations and thesaurus-based data have also been reported to exhibit scale free (H. Chen et al., 2018, p. 5; Steyvers & Tenenbaum, 2005, p. 43). However, in their re-analysis of existing cue-association networks, an English network based on the *University of South Florida Norms* (USF; Nelson et al. (2004)) and a Dutch association network (Deyne & Storms, 2008), Morais et al. (2013, p. 138) found that there is compelling evidence for alternative explanations to the previously attested strict power-law degree distributions. Following Morais et al.'s methodologically refined degree distribution analysis, the best form to describe the degree distributions of cue-association networks has been identified as power-law distributions with an exponential cut-off. It needs to be explicitly acknowledged that, while this is the best fit for the observed networks, a range of other plausible options also exists.

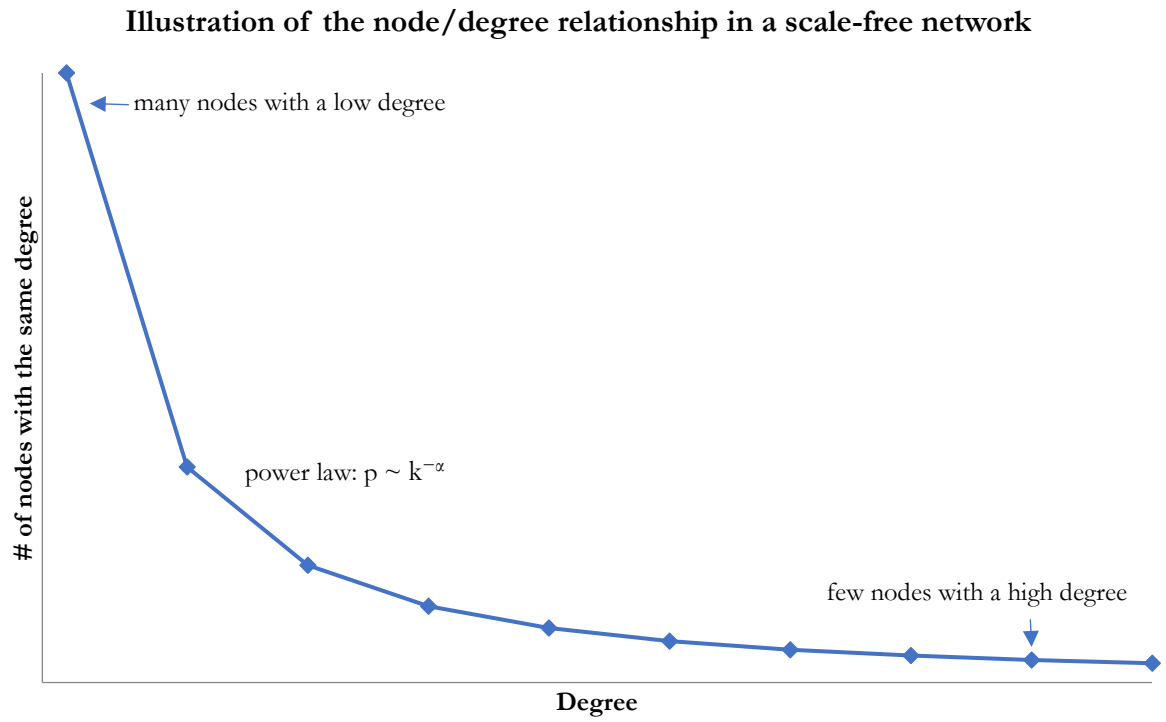


Figure 13: Prototypical shape of a plot showing the number of nodes sharing the same degree and their respective degree when a network is governed by a power law of the shape $p \sim k^{-\alpha}$.

One theoretical explanation behind an observation of power laws of this shape in network data lies in the concept of preferential attachment. In general, this describes a process whereby nodes get connected to other nodes proportional to their current connectivity meaning that nodes with a high connectivity experience a proportionally high connectivity when they connect to other nodes in a circular and self-perpetuating manner (Castro & Siew, 2020, p. 16; Sheridan & Onodera, 2018, p. 1). In linguistic terms, preferential attachment can, for instance, be explained through the fact

that the acquisition of a new word is easier when it is linked to another known and heavily interconnected word in language learning networks as well as the idea that words that have the capacity to convey different meanings in different contexts might be more flexible and applicable to new contexts as well, further increasing their interconnectedness. In a psycholinguistic context, recall and recognition of response words when presented in a pair with a well-connected cue word were found to be facilitated versus response words paired with less well-connected cue words which matches a preferential attachment process (Mak & Twitchell, 2020, 1059; 1067). On the one hand, preferential attachment holds explanatory power for networks exhibiting ideal, non-truncated power-laws and provides a logical and seemingly plausible explanation for the shape of the Mental Lexicon. On the other hand, however, more refined power-law analyses such as the ones carried out by Morais et al. (2013, p. 143) consistently show truncated power-laws – this means that network growth solely based on preferential attachment is not consistent with the structures observed in real semantic networks. In Chapter 4.3.2, this is further explored by means of the collocation and word association networks generated as part of this thesis.

Whilst the abovementioned linguistic findings imply the existence of a real power law that is followed by the observed distributions and certain theories provide explainable reasons for this being the case, it is important to discuss and acknowledge the limitations of this feature as well as possible alternative explanations. One component of the limitations is statistical in nature, the other conceptual. In terms of statistical problems, fitting a power law and assessing its appropriateness demands more complex processes than are often applied in order to limit errors that are systemic to, for instance, the evaluation of goodness of fit such as least squares approaches, relying on straight lines emerging from histograms on a doubly logarithmic plot alone to assert the presence of a power law and setting an arbitrary lower bound.

Given that the available data is not a comprehensive representation of, for instance, the entirety of modern British English, even the most ambitious assertion must be constrained to identifying an observed distribution that is consistent with a distribution following a power law rather than claiming that it truly is governed by this law. Ultimately, it can, however, not be claimed that the unobserved data would equally follow this power law; for a more detailed discussion of these issues see Clauset et al. (2009, pp. 663,666,690).

The other, methodological problem that arises when assessing the applicability of a power law and scale-free properties to node-degree distributions in language data is the large influence the data collection exerts over this distribution. When determining a best fit of a power law for this kind of data, the large number of nodes with a very low degree plays a significant role but there is reasonable doubt that this represents language accurately. When a corpus is collected, a large

number of words that occur on the "borders" of the corpus only are necessarily collected. Examples for this are words that have been mentioned once only in a conversation such as names or placenames, specific dates or neologisms. These words would be assigned a very low degree, largely based on their low frequency in the corpus, while they are in reality very likely to be used in different contexts and display a much higher degree. This is the case since the corpus can be seen as the sample aiming to represent a much larger population, here the entirety of the English language. Chapter 4.2.1.1 discusses the rationale for corpus selection and limitations of this approach in greater detail. At its core, the issue therefore lies in the fact that the presence of a node with a high degree in a corpus does imply that, on a population level, this high degree would also be present, but the absence of a low degree may or may not simply be the result of the small sample size versus the astronomic population size. It is therefore imperative to question all implications that arise from power laws that are strongly dependent on the true number of nodes with a very low degree.

2.7.3 Small Worlds

Another commonality of many real-world networks with linguistic networks is small world properties. Small worldedness as a feature of a network's structure is relevant for two main reasons: it firstly carries implications regarding the overall robustness of the network against random node deletion Siew et al. (2019, p. 6). In a linguistic context, small-worldedness thus indicates to what degree a network will be able to withstand forgetting individual words. Small-worldedness secondly also carries implications as to what paths the processing and recall of linguistic information is likely to take. It can also be seen as an approximation of the efficiency of a network's information spread. Beyond this, it also carries a broader meaning: Small world properties optimise the trade-off between high connectivity and efficiency in information flow (Beckage & Colunga, 2016, p. 9; Deyne et al., 2016, p. 48). As such, they might naturally occur in linguistic data due to the nature of communicative language itself: There is a large pool of words to choose from in a conversation and small world properties might be the key that enables the communicating individuals to balance low cognitive effort with acceptable precision and understandability.

In fact, small world properties and the resulting phenomenon of centralised, strongly connected hubs have also been found in phonological networks and shown to influence linguistic processes: In the three experiments of their spoken word recognition study, a naming task, a lexical decision task, and a serial recall task, Siew and Vitevitch (2016) found that words that stem from the large connected component at the centre of the phonological network exhibiting small-world properties are recalled less reliably than words taken from the islands disconnected from the central hub of the network.

It is now essential to define small networks and explore what other types of networks can be distinguished from it. The following section therefore gives a brief overview over two other key types of networks, regular lattice networks and random networks, before providing a definition for small world networks. A regular lattice network is a network where every node is connected to all its nearest neighbours; this network is non-random and Figure 14 provides both a grid and a ring visualisation of such a network.

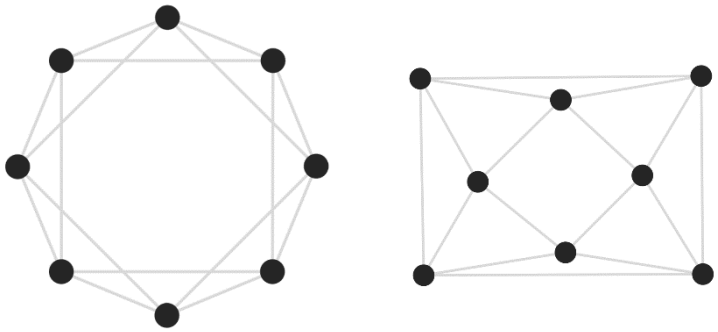


Figure 14: Grid and circle representation of the same regular lattice network.

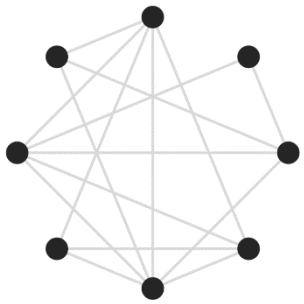


Figure 15: Visualisation of an Erdős & Rényi random network with the same number of nodes and edges as the regular lattice network.

A random network on the other hand is characterised by the absence of an underlying system that governs which neighbours a node has (if any) resulting in a choice of any neighbour with a constant probability. There are different types of random networks such as the one described by Erdős and Rényi (1960, p. 17). This particular type of network (see visualisation in Figure 15) is the result of a random choice of edges for a start node of the network where the probability for choosing any next edge is also equally distributed over all possible edges each turn.

A small-world network is distinctly different from the above since it is a network with a high clustering coefficient and a relatively small shortest path length (Bales & Johnson, 2006, p. 463; H. Chen et al., 2018, p. 6; Utsumi, 2015, p. 4). In this sense, they lie between the two extremes of regular and random networks (Watts & Strogatz, 1998, p. 440). When linguistic networks exhibit a

small world structure, this means in practice that any two words are relatively closely connected despite a high number of topic-specific and decentral hubs. Looking at the overall shape of a network in a linguistic context allows for assessing whether or not they display small world and scale free properties and would thus benefit from the ease of information spread and robustness described for these types of networks.

The exact methodological definition of a small-world network is, however, not as straightforward as one might assume with different researchers using a number of different formulae to assess whether a network fits this label. The issue with the definitions is similar to the problems encountered with definitions of collocation: terms such as "high" and "relatively small" are by no means precise, objective or directly measurable – they only carry meaning when the researcher applies certain thresholds or uses a reference network to extrapolate the differences and claim that one possesses more small world properties than the other.

In practice, the averages of the clustering coefficients and the shortest path lengths of a network are often compared to a random network as described above which is taken to serve as a baseline; if the network of interest displays a significantly higher clustering coefficient and a significantly shorter average shortest path length it is taken to have small world properties.

Telesford et al. (2011, p. 369) use this to define a variable that helps quantify the small-worldedness of a network: the small world measurement ω Equation 3). ω ranges from -1 to 1; the closer the resulting values are to zero, the stronger the indication that the observed network displays small world properties.

$$\omega = \frac{L_{rand}}{\mathrm{L}} - \frac{C}{C_{latt}} \qquad \text{Equation 3}$$

In linguistic research, networks based on almost all levels of language such as semantic, phonological and orthographic networks have been shown to exhibit small world features (Trautwein & Schroeder, 2018, p. 12). More specialised models such as distributional semantic models (DSMs), i.e. corpus based networks consisting of a vector-based representation of word meanings, have also been found to display small-world properties such as a high clustering coefficient, a small shortest path length, and high connectivity (Utsumi, 2015, p. 9). There is further evidence based on neuron synchronization measurements indicates that the human brain might physiologically also exhibit a small-world architecture (Watts & Strogatz, 1998, p. 442). Looking at collocation in particular, the required features for a small-world network have been found to be more clearly pronounced in co-occurrence based graphs than in graphs based on pragmatic similarity (Cecchini et al., 2018, p. 768). This suggests that long-range nodes and shortcuts might

be more pronounced in language in use, whereas pragmatic categories overall seem to be more distinct and less well interconnected. Beyond this, a range of different word association graphs have also been shown to exhibit small-world and scale-free properties (Utsumi, 2015, p. 1).

Apart from small worldedness and power laws, a number of further network properties in complex networks are of general interest: Networks with a skewed degree distribution, otherwise special community structure (i.e. hubs as explored above), or distinctive mixing patterns. Distinctive mixing patterns emerge when specific nodes have a high proclivity to be connected to other nodes with similar (or dissimilar) properties only (Karuza et al., 2016, p. 630). For reasons of brevity these cannot be explored in greater depth as part of this thesis.

2.8 Conclusion

This Chapter has explored the intersection of corpus linguistics and psycholinguistics through the lens of network approaches, providing a comprehensive analysis of linguistic networks, collocations, and word associations. The research has demonstrated the utility of network methodologies in various areas of linguistic research, particularly in the study of collocation networks and psycholinguistic networks.

The theoretical underpinnings of semantic representation were examined, with a focus on cognitive and functional linguistics. This led to the dissection of the concept of collocations, and, due to the multitude of schools of thought surrounding this concept, various types of collocations were identified and analysed across multiple spectrums, including syntagmaticity/paradigmaticity, (a)symmetry, lexical/grammatical, strength of association, predictability, and range. This classification is necessary for the grouping of collocational patterns emerging from the corpus-based networks in Chapter 4.3.3. Beyond this, the core concept of this thesis, psycholinguistic plausibility, has been explored via a comprehensive examination of language learning processes in the mental lexicon, with a particular emphasis on Statistical Learning and its role in collocations as well as linguistic memory in the mental lexicon, semantic maps and their structures, and retrieval processes in the mental lexicon. Lastly, a Chapter on graph theory, intended to be somewhat of a linguists' guide to this field, was provided. This represents a significant part of this research since all graph-theoretical parameters of interest at the micro-, meso-, and macro-levels are not only listed but explained in detail in a linguistic context. In conclusion, this Chapter has contributed to the understanding of linguistic networks and their role in language learning and processing. It has highlighted the potential of network approaches in linguistic research and has opened new avenues for future research in this field by providing a solid, accessible methodological foundation for graph theoretical explorations. The findings of this research have implications for both corpus linguistics

and psycholinguistics, and they underscore the importance of interdisciplinary approaches in the study of language. Equipped with this theoretical, empirical, and methodological foundation, it is now possible to explore the research questions explored in the following Chapter.

2.9 Research Questions

Having explored the theoretical framework alongside the status quo of research surrounding the applicability of linguistic networks, collocations, intersections of psycholinguistics and corpus linguistics, as well as graph theory, this Chapter concludes the literature review by posing the questions that frame the remainder of this thesis. An examination of the aforementioned fields clearly shows that LLN approaches could provide significant contributions to the wider field since networks excel where other methods of information summarisation and visualisation are not as appropriate: in representing complex, dynamic, and adaptive systems. Indeed, language constitutes a prime example of this category of data. Integrated network explorations, especially in the context of generating corpus-wide collocation networks, thus present a gap in existing literature that this study aim to address explorations of such networks can provide quantifiable insights into psycholinguistically plausible structures of mental representations. Analysing these constructs further holds the potential to uncover previously unexplored representations of language.

RQ1: How can current approaches to AM extraction be harnessed in a psycholinguistically plausible manner?

RQ1 is aligned with the aim described in Chapter 1.2 to evaluate existing approaches to collocation identification. Answering this question entails an in-depth evaluation of the existing approaches including the individual parameters used for collocation extraction alongside the underlying cycle linguistic research that limits the viability of different methodological choices. The outcome of this question is aimed at providing researchers with clear, psycholinguistically informed recommendations as to which features AMs need to possess so they can be considered potential candidates for generalisation towards mental linguistic processes, and which current approaches fail to meet this standard.

Answering this question requires a solid knowledge of graph theory and the intricacies of available graph theoretical parameters with regards to their usability in linguistic contexts and the computational feasibility of their implementations. The author is not aware of any currently existing approach attempting to systematically and thoroughly assess existing broad evaluations of AMs with regards to their psycholinguistic plausibility. The outcome of RQ1 further directly informs all future collocation network generation steps since only psycholinguistically plausible association

measures are considered as the basis for networks that are later contrasted with word association networks.

RQ2: Which AMs lead to collocation networks that best approximate the content and structure of large word association networks?

RQ2 then entails applying the identified set of psycholinguistically plausible Association Measures and their parameters (filters, thresholds, window spans, directionality) to the BNC 2014 and comparing the results to the SWOW-based word association network. On the basis of this, different approaches to collocation identification can be ranked and assessed in terms of their similarity to word association data.

This approach necessitates a full data agnostic network generation pipeline that can take in both corpus data and dynamically extract collocations on the basis of individual choices made by the researcher and individual words from word association experiments which are then weighted and transformed into network representation. The input data needs to be structured and processed rigorously and in the replicable manner; a full Python-based pipeline for this process is developed as part of this thesis and can be accessed in Appendix A.

Examining the networks that are the outcome of this process fills a methodological research gap when it comes to assessing the generalisability of results from collocation explorations to mental associations of the identified collocational structures. This exploration is of use to the wider field – even if no set of parameters results in a collocation network with an exceptional similarity to word associations – since it lays out the strengths and constraints of cognitive generalisations on the basis of communicative language. Promising initial results in terms of a goodness of fit between the word association network and the collocation network would open a large number of opportunities for modelling likely common thought processes using corpus data and introducing more layers, for example containing phonological or orthographic information, into the multidimensional network representation of language.

RQ3: Is there a general structural difference between usage-based network patterns and patterns found in word association networks?

RQ3 compares collocation networks (representing usage-based networks) to word association networks in order to assess how large their overlap is both in terms of the content in the shape of corresponding connections between words, and in terms of their structural properties. This research question is of particular interest since corpus-based networks contain common structures of communicative language, whereas word association networks contain common structures of

linguistic thought. The differences between these two systems have seldom been researched quantitatively (a notable exception being Mollin (2009, p. 175)) and even less work[11] exists on a level that takes whole clusters of words and their position in a larger context into consideration. The network approach taken here allows for explorations of this kind. As part of this research question, differences and commonalities between these systems are explored from a cognitive and functional perspective in order to inform further studies. And the overall structural differences between usage-based network patterns and patterns found in word association networks can be explored.

An evaluation of structural difference between usage-based network patterns and patterns found in word association networks necessitates the following points:

1. General considerations
    - Are the networks intuitively interpretable?
2. Empirical evaluation of structural comparisons (Macro-level)
    - What are the properties (size, connectivity, etc.) of the resulting networks?
    - How large is the percentage overlap between unweighted nodes and edges in the different collocation networks and the word association network?
3. Empirical evaluation of qualitative comparisons (Meso-level and Micro-level)
    - How similar are the emerging key clusters?
    - Which words display a particularly high linguistic availability as indicated by their high degree in the respective networks?
    - Which words are candidates for kernel lexicon membership?
    - Do network-central hubs qualitatively differ between these networks in terms of the high/low frequency of represented words (as expected on the basis of Veremyev et al. (2019, p. 3)?

The three layers of this exploration, beyond the first general question, require different analytical toolkits. This thesis introduces a full processing pipeline including all code required to generate the results and replicate the procedure for different datasets. The layers to be explored are the macro, i.e. network-wide level where similarity is measured using simple percentage overlap, but also graph theoretical metrics such as NetSimile (Berlingerio et al., 2012), and Adjacency Spectral Distance (Wilson & Zhu, 2008) for structural comparisons. On the meso-level, clusters are identified using MCODE (Bader & Hogue, 2003) for the word association network and several collocation

[11] A notable exception here is Deyne et al. (2021), albeit taking a promising, but less interpretable word embedding angle.

networks created on the basis of different psycholinguistically plausible AMs. Clusters are presented in order to extract particularly strongly connected associative domains. These are visualised using edge-weighted spring directed layouts (Kamada & Kawai, 1989). This mode of presentation is particularly suitable for analysing clusters since it visually displays individual words that are strongly linked through a high association value closer to one another than words that are more loosely linked. Another motivation for choosing this method of visualising network data is its link back to the concept of psycholinguistic plausibility: core cognitive processes such as spreading activation (Collins, 1975) could be modelled using network approaches. A clear difference between existing domain extraction approaches and the present thesis is the overarching goal of cognitive plausibility of the clustering process itself as compared to more exclusively result-oriented approaches.

Lastly, on the micro-level individual high degree and long-range nodes are explored since they fulfil a special function in the respective networks, and in a word association context represent candidates for kernel membership. Concordance lines are provided for the respective high-scoring nodes from BNC 2014 networks which makes it possible to explore the effect the AM choice has on the nature of the extracted collocations and their overlap with patterns emerging from word association networks.

As described further in Chapter 5.7, the resulting methodology can be employed beyond the comparative focus of this thesis in more traditional linguistic fields of study such as genre and register research for topic extraction, and it can be used to provide automatically extracted metadata to enrich sociolinguistic and pragmatic corpus studies.

In conclusion, the questions asked in this thesis seek to bridge the gap between psycholinguistic and corpus linguistic research by developing a method for creating and analysing large-scale, corpus-wide collocation networks that are psycholinguistically plausible, exploring the underlying assumptions, and putting the resulting methodology into practice. The research questions addressed in this work focus on the psycholinguistically plausible generation of these networks, the comparison of usage-based network patterns with those found in word association networks, and the exploration of the structural and qualitative differences between these systems. The findings from this research contribute to the understanding of language perception and production by highlighting structural similarities and differences regarding the nature and connectivity of lexical items in a word association and collocation setting. Beyond this, they also pave the way for future studies that aim to further refine and apply statistical methods to operationalise possible mental processes. This thesis underlines the importance of methodological innovation in linguistics and highlights the potential of network approaches in uncovering previously unexplored mental

representations of language. The insights gained from this work offer promising directions for future research, including the potential for conducting psycholinguistic experiments to understand how predictive large-scale networks are of individual mental processes, and the possibility of highlighting specific terms and finding sets of semantically related items. As such, this thesis represents a step towards the interdisciplinary application of corpus linguistics and psycholinguistics, offering a new lens through which to view and understand the complex dynamics of language.

3 Methodological Evaluation and Innovation (RQ1)

3.1 Introduction

The aim of this chapter is to address RQ1 by exploring the possibility of implementing a new method to display corpus-wide collocation networks on the basis of current findings from psycholinguistics, and thus in line with the cognitive commitment (Lakoff, 1991, pp. 53–55), as outlined in Chapter 2.3.1. The findings presented and discussed there are used in the present Chapter in order to assess and evaluate all steps necessary for the generation of corpus wide collocation networks, i.e. the calculation of contingency tables and Association Measure (AM) values for each possible collocation, and a selection of relevant factors representing psycholinguistic features. A particular focus lies on the evaluation of the psycholinguistic plausibility of different association measures, for two reasons: Firstly, AMs provide the proxy for psycholinguistically relevant concepts; the overall shape of the network along with all graph theoretical features extracted later on depend on them. Secondly, this work contributes to the advancement of traditional corpus linguistics, independent of network approaches. The recommendations made for choosing specific association measures over others are applicable to any methodology that involves extracting collocations. This work aims to address the recognised research gap surrounding a detailed and accessible discussion of the assumptions underlying the use of different AMs in corpus linguistics (Gries, 2012, pp. 47–48). In applied research, collocations are commonly employed as a proxy for contextual embeddings, which are then generalised to represent the attitude or stance of a speaker (S. Chen, 2013; Galasinski & Marley, 1998). These objectives clearly discourage the use of measures that are based on assumptions known not to apply to linguistic data.

The definition of psycholinguistic plausibility here entails that, on a theoretical level, assumptions that are made as part of different association measure calculations do not contradict current theories and experimental findings from psycholinguistics. As explored in Chapter 3.4, particularly unifiability with theories of Statistical Learning and the Mental Lexicon alongside evidence from reading time and cue response studies, are used to inform what is and what is not psycholinguistically plausible.

The Chapter opens with an introduction that provides a definition of Association Measures (AMs) as well as a critical discussion of existing methodological discrepancies when constructing contingency tables. This is followed by a brief classification of AMs into several groups based on their origins or the statistical line of thinking they most closely align with. The next chapter contains descriptions and accessible interpretation of various association Measures that serve as the basis for corpus linguistic collocation statistics via revealing both lexical and syntactic properties (McEnery & Brezina, 2019, p. 97). This includes MI Scores, Poisson-likelihood and

Hypergeometric Likelihood, Poisson and Fisher's Exact Test, χ^2, Log likelihood, ΔP, r_φ, (log)Odds Ratio, (log)Dice, T-score, and Z-score. Presenting these metrics goes hand in hand with an explanation of their basic underlying principles which plays a pivotal role when judging them regarding possible psycholinguistic plausibility. This chapter then also discusses other parameters in collocation extraction, including directionality and symmetry, window spans, and the unit of analysis in general. The last Chapters use the obtained findings to explore the practical applicability of psycholinguistically plausible corpus-wide collocation networks. The reader is provided with a flowchart that is intended to aid decision-making as to which AMs show promise for psycholinguistic validity – and which can only be used in practical applications and opaque models rather than in exploratory research. This Chapter, which includes clear recommendations and guides to selecting appropriate methodological tools, is particularly important given the lack of standard approaches in the field in order to lay solid foundations for further research situated at the intersection of CL and psycholinguistics, and, naturally, corpus-based large linguistic networks.

3.2 Towards Psycholinguistically Plausible Association Measures

This Chapter explores a wide range of Association Measures utilised for collocation identification. Here, associated words are defined as words connected through "significative cooccurrences between words in the same sentence" (Ferrer-i-Cancho & Solé, 2001, p. 2261) and they are taken to indicate some form of cohesion (Kolesnikova, 2016, p. 341). The effectiveness of automatic collocation extraction largely hinges on the specific research question being addressed. Although automatic collocation extraction is an extensive research field, no single method has yet yielded results comparable to manual human collocation extraction (Garcia et al., 2019, p. 56). Nonetheless, substantial advancements have been made over the previous decades, and a network approach to collocation identification aspires to take a further step in the towards enhancing the precision and efficiency of the process.

In a network context, the exact statistic used to investigate the *significative cooccurrences* plays a central role for the overall shape and properties of the final network since it explicitly defines what criteria a collocation must fulfil. In this sense, AMs and their parameters such as window spans, thresholds, directionality etc. are merely a way of precisely defining what types of co-occurrences are considered to be collocations. Due to this function, different AMs can be a better or a worse fit for specific research questions and their applicability will depend on all parameters that have been pre-defined. It is particularly important to acknowledge that the performance and goodness of fit of specific AMs is always also highly dependent on the type of corpus at hand with respect to its sampling scheme since features like overall size, complexity, uniformity etc. (Evert et al., 2017,

p. 531). All of these features carry implications as to which basic hypotheses can be formulated and what assumptions can reasonably be made when choosing how to quantify associations.

Before AM options are discussed in detail, it is important to consider on low-level decisions that impact all AMs. One particularly important question is whether it is deemed appropriate to use a certain threshold (be it based on data or intuition), to apply a cut-off point due to collocation ranks or to include all identified collocations regardless of the AM value. This decision should depend on the view a researcher holds with regards to seeing collocations as a binary phenomenon or a spectrum. The binary phenomenon standpoint naturally results in the application of a threshold to differentiate between the two categories, collocations and non-collocations. A minimal co-occurrence threshold of 5 is commonly recommended for this approach (Evert, 2008, p. 1244). The view of 'collocativity' as a spectrum, on the other hand, naturally leads to ranking the results regardless of raw AM scores (Evert, 2008, pp. 1216–1217). Given the support for the notion that collocations are the result of gradual repeated exposure in Statistical Learning mechanisms, this thesis adopts the spectrum approach. Consequently, it applies a rank-based cutoff at a predetermined percentile across all datasets and sub-corpora.

Finally, the specific selection of AMs to be examined requires justification, and a guide for meaningful AM selection is central to this Chapter. This thesis generally prioritises explanatory power over computational efficiency to ensure psycholinguistic plausibility. This choice distinguishes the present project from more result-oriented (and often commercial) projects that investigate linguistic phenomena. The aim of this thesis is not to model language perception and language production in the most efficient manner, but to enhance understanding of these concepts and identify relevant factors that influence these processes. Although this makes the methodology developed in this thesis less competitive with the performance of e.g. Large Language Models there is a greater potential benefit in causally understanding the underlying procedures. This is the case since this approach maintains interpretability and facilitates the development of more refined and realistic models in the future.

3.2.1 Contingency Tables

The design of the contingency tables used as the basis for the calculations as well as the standard nomenclature is discussed here in order to present clear and unified equations for calculating the association scores. This is necessary since there are inconsistencies in the construction of contingency tables across the field of corpus linguistics (see Table 2), and it is aligned with the emphasis of this thesis on methodological rigour to propose a single theoretically sound approach for linguistic contingency tables.

The observed frequencies of co-occurrences are henceforth described as O11, the observed frequency of the first collocate with any other word as O21, the observed frequency of the second collocate preceded by any other word as O12 and all co-occurrences of words that are not part of the collocation O22.

Analogous to that, the corresponding expected frequencies are described as E11, E12, E21 and E22. This is standard practice and analogous to the approach taken in Evert (2005, p. 337).

Table 2: Contingency table employed in this thesis, henceforth referred to as "LLN method".

O11 – frequency of tuples that are word 1 followed by word 2	O12 – frequency of tuples that are word 2 not preceded by word 1	R1
O21 – frequency of tuples that are word 1 not followed by word 2	O22 – frequency of total tuples that are neither starting with word 1 nor ending in word 2	R2
C1	C2	N – total number of tuples

As mentioned above, looking at popular tools such as WordSmith[12] (Scott, 2024) or SketchEngine (Kilgarriff et al., 2015) reveals that contingency tables are generated differently in different applications. In some cases, they are populated using counts that depend on raw frequencies of words within the corpus rather than their frequency as a constituent element of a potential collocation. According to the 'Formulae' page of the official WordSmith website, the contingency tables would have the following shape.

The motivation for using Table 2 over Table 3 is the fact that the 'atomic units' in Table 3 differ from cell to cell. In this case, a mix of tuple counts (for O11) and raw word counts results in a failure of the table to sum to N. The resulting expected values are thus also drastically different. Methodological considerations such as this are paramount for the development of holistic collocation networks since they influence different AMs to a different degree and would thus skew all results. The processing pipeline made available in the Appendix A of this thesis allows researchers to extract more than 10 different AMs on the basis of the correctly summing Table 2.

[12] https://lexically.net/downloads/version_64_8/HTML/formulae.html

Table 3: Contingency table used in WordSmith (Scott, 2024). Differences in bold. A similar approach is used in SketchEngine (Kilgarriff et al., 2014).

O11 – the joint frequency of two words	O12 – frequency of word 2 in the corpus (when it isn't part of the collocation)	R1 - frequency of word 2 in the corpus
O21 – frequency of word 1 in the corpus (when it isn't part of the collocation)	O22 – total tokens – O12 – O21	R2
C1 - frequency of word 1 in the corpus (when it isn't part of the collocation)	C2	N – total number of words

To illustrate how these seemingly small differences produce significantly different results the following mini mock-corpus is analysed using both approaches.

Mini-Corpus:

> how er how long is it since that was a dock? well it's many hundreds of years ago isn't it? is it? is it that long? mm I would say so.
>
> (BNC 2014 v.2, Sp1m2f167.xml)

Table 4 shows the directional (left-to-right) counts for each sentence-wide word combination contained in this mini-corpus. For *it, is* and *long, is* the two contingency tables shown in Table 2 and Table 3 are obtained and presented in Table 5. It becomes evident that any field other than the co-occurrence count itself differs, in this case most strongly regarding O21, R1, and C1. The frequencies in the contingency table used for this project are strictly lower.

In many real-world scenarios, these results may differ from each other to a lesser degree as a result of working with high-frequency items, but it is essential to be aware of the differences when carrying out any interpretations on the basis of the so obtained scores or when analysing scores obtained from a variety of different corpus tools.

Table 4: Tuple counts for the Mini-Corpus.

node	**collocate**	**count**	**node**	**collocate**	**count**
hundreds	of	1	how	er	1
of	years	1	er	how	1
years	ago	1	how	long	1
ago	is	1	long	is	1
is	n't	1	it	since	1
n't	it	1	since	that	1
is	it	3	that	was	1
it	that	1	was	a	1
that	long	1	a	dock	1
mm	I	1	well	it	1
I	would	1	it	s	1
would	say	1	s	many	1
say	so	1	many	hundreds	1
how	er	1	SUM		28

Table 5: Contingency tables resulting from counts for *is*, *it* (tables on the left) and *long*, *is* (tables on the right). Differences shaded.

is, *it* (LLN)		
3	1	4
2	22	24
1	27	28

is, *it* (WordSmith)		
3	1	4
2	27	29
5	28	33

long, *is* (LLN)		
1	0	1
1	26	27
2	26	28

long, *is* (WordSmith)		
1	1	2
3	28	31
4	29	33

3.2.2 Classification of Methods Used for Collocation Extraction

In this Chapter, an overview of the different types of AMs in order to classify and contextualise the selected examples in the following sections is presented. Broadly speaking, the existing collocation extraction strategies can be categorised into six approaches: Statistical AMs, further subdivided into measures reporting effect-size or statistical significance, information-theoretic measures (Evert, 2005, p. 77), linguistically informed, rule-based approaches, deep-learning models such as word2vec (Kolesnikova, 2016, p. 342), and hybrid approaches.

The first approach to collocation extraction to be presented here is statistical AMs measuring effect size. A large number of very popular and commonly used AMs in corpus linguistics such as Jaccard and Dice Coefficients, Odds Ratio belong to this group of measures making use of maximum likelihood estimates; for a more comprehensive discussion of these measures see Evert (2005, pp. 84–86). The advantage of these methods over models based on statistical significance is that they strive to assess the strength of collocational adhesion instead of measuring the volume of

information available to signal the presence of a collocation initially (regardless of how strong or weak this may be).

Table 6: Systematic strengths and limitations of different approaches to collocation extraction.

Collocation extraction method	+	-
Statistical AM - effect-size	Evaluates degree of collocation	Fails to assess the significance / reliability of results Often not linguistically informed
Statistical AM - significance	Good performance when benchmarking against collocation dictionaries	Fails to measure strength of collocation Often not linguistically informed
Statistical AM - information theory	Evaluates how well the number of occurrences of the node reduces the uncertainty about the number of occurrences of the collocate Good performance when benchmarking against collocation dictionaries	Also ignorant regarding the non-randomness of language
Rule based approach	Linguistically informed	Requires accurate POS tagging Relies on pre-defined word class combinations
Deep learning models	Seamless adaptability to new data Possible implementations go far beyond identifying word associations	Often black-box models which leads to limited falsifiability and potential ethical problems when used as the basis for real-world decision making Underlying assumptions cannot be aligned with findings from psycholinguistic research
Hybrid approaches	Combination of approaches can combat drawbacks of individual approaches	Less transparent and possibly hyper-complex

In simpler terms, effect-size methods might be more suitable for the theoretical framework of lexical collocations compared to significance level methods, which are likely to be more related to grammatical collocations. Both of these types of collocations are crucial for evaluating the psycholinguistic reality of collocation networks. This is because Statistical Learning, as measured in child language acquisition experiments, depends not just on high repetition frequencies (beneficial for significance-based AMs), but also on reliability, i.e. the ratio of correct to incorrect

interpretations of a collocation (Ellis, 2006, pp. 5f; 15) as a result of contingency learning (Peterson & Beach, 1967, p. 42). While effect-size approaches are highly effective for identifying lexical collocations, which are particularly important to a large proportion of applied research, they are also limited by the variation of reliability of their results.

"Significance"-based Approaches

A closely related approach that has already been mentioned in the previous section, significance-based approaches to collocation extraction, also relies on observed and expected frequencies of word combinations. Measures such as log likelihood and χ^2 fall into this category. The strengths and limitations of what has been called 'significance-based' AMs are almost diametrically opposed to those based on effect size: they attempt to assess the amount of evidence against the null hypothesis , but fail to quantify how strong this association is (Wermter & Hahn, 2006, p. 786). In other words, significance-based tests are related to the amount of evidence (usually large in corpora) of observed differences which leads to the possible claim that if a result is significant at a given alpha level, the likelihood of encountering a different result is below the chosen error level for a large number of repetitions of the experiment (Wallis, 2013, p. 352).

This is particularly problematic if collocations are ranked according to their significance-based AM value since the ranking does not indicate any particular strength of collocation (as would be desirable), but merely the significance of there being a difference of any scale. In practice, significance-based AMs have been shown to perform well in identifying collocations that are also included in collocation dictionaries with high precision and recall (Evert et al., 2017, p. 537), but they are in essence statistical and data-driven rather than linguistically informed.

Information Theory

Statistical AMs based on information theory follow a different approach; the most popular examples are MI-based scores (MI, MI^2, MI^3 etc.). Mutual information describes the level of uncertainty about a collocate based on a node and vice versa. It presents – under the assumption that the occurrence of the node is independent of the occurrence of the collocate – the inherent dependence in the joint distribution of node and collocate. In practice, this means that MI-based scores measure the homogeneity of the observed contingency table versus the homogeneity of the expected contingency table. Moving away from MI-scores specifically, information theoretical approaches more generally rely the underlying concept of surprise and information grain (Shukla et al., 2012, p. 172) which serve as a basis for calculating entropy. The more surprising an event (i.e. the more options there are for possible outcomes and the less probable the observed outcome is), the higher its informational value. The surprisal of encountering a particular collocation in a

corpus with 5000 equally probable lemmas would therefore be larger than the surprisal of encountering a particular collocation in a corpus with 1000 equally probable lemmas. Entropy then aims to measure how much informational value is gained by learning about the outcome of an experiment; applied to linguistics entropy measures how "surprised" the researcher would be to encounter the frequency observed for a given collocation given the expected co-occurrence of the same collocation. Another information theoretic measure which is at the same time a popular metric in psycholinguistic research is forward transition probabilities (Smith & Levy, 2013) and backward transition probabilities. Since these metrics present one of the few existing well established touching points between the fields of psycholinguistics and corpus linguistics (McConnell & Blumenthal-Dramé, 2019, p. 2) they are described in greater detail in the next Chapter.

Rule-based Approaches

While arguably less mainstream in corpus linguistic research, rule-based approaches such as Limited Syntagmatic Modifiability (LSM) and Limited Paradigmatic Modifiability (LPM) also exist. Both of these approaches are linguistically informed and directly rely on concepts such as non-adaptability of collocations or terms. In this sense, this approach is less data driven and more data informed than the others: It restricts the results on the basis of an existing theory rather than test or formulate a theory on the basis of the data encountered 'in the wild'. LSM is defined as "the linguistically motivated statistical association measure for a generic collocational syntactic target structure POS" (Wermter, 2008, p. 109). The example phrases *on the table* and *on the whole* illustrate this: *on the table* occurs 224 times in the Spoken BNC 2014, and 42 modified versions of this (e.g. *on the operating table*, *on the kitchen table* etc.) exist in the corpus for *on the table*. In contrast to this, *on the whole* occurs 52 times in the corpus, and no modified versions of this can be found. Before correcting for frequency effects, the LSM score of the former would therefore be lower (0.836) than of the latter (1) since it is more modifiable. The underlying idea of this procedure links in very well with the notion that not only the frequency of occurrence of a term is relevant to Statistical Learning processes, but also coherence and exclusivity (Dehmer et al., 2011). It is, however, essential to note that this approach heavily relies on a predefined POS structure of the particular collocation type in question which makes it less universally applicable. While not viable as the primary measure for holistic network approaches such as the one taken in this thesis, it holds great potential for providing considerable refinements of existing collocation tables when modifiability and compositionality are of interest.

LPM operates on a similar basis and is used to evaluate if a multi word unit is a term, i.e. one meaning-carrying unit, via assessing the "modifiability of the paradigmatic context for a particular

word token within a potential terminological expression" (ibid, p. 114). This is of interest in the context of this thesis since Wermter's definition of terms coincides with the very broad definition of collocation followed in this thesis. The procedure for obtaining LPM scores is examining an n-gram of interest (e.g. a trigram consisting of three NPs) in terms of the replaceability of each individual component. To exemplify, for the term "Client Server Model" the frequencies of occurrence for "_ Server Model", "Client _ Model" and "Client Server _", as well as "_ _ Model", "_ Server _", "Client _ _", and lastly any combination of any three items will be used to calculate the LPM score of the term. This approach relies by default on an even more restricted POS structure, namely terms consisting of NPs only, since it has been designed with the identification of noun-based terms specifically – it is theoretically possible to include other parts of speech in the analysis. Overall, this approach, again, shows great promise for filtering a subsection of collocations or terms, but ultimately heavily depends on the researcher's intuitions as to which combinations constitute acceptable "terms".

Deep Learning Models

Deep learning models entail creating word embeddings which can be described as vector values that, taken together, serve to represent a given word. Common deep-learning based approaches such as word2vec use moving window spans to create these embeddings and optimise their "phrase learning models" using co-occurrence counts (Mikolov, Sutskever, et al., 2013, p. 6); thus ultimately relying on collocations for their predictions. Another type of language model that has recently received a lot of attention is Generative Pretrained Transformer (GPT) models such as T. B. Brown et al. (2020). Instead of focusing on word embeddings, GPT models use transformer architectures and self-attention mechanisms to understand the context of a word in relation to every other word in a sentence (Yenduri et al., 2023). While these models are less transparent than, e.g. word embedding based models, they involve predicting the next word in a sentence at the pretraining stage of the learning process and thus also partially relies on collocational relationships albeit much less directly than, e.g. word2vec.

A fundamental critique of such models is their reliance on abstract embeddings or transformer structures to yield convincing outcomes, despite the inherent non-interpretability of these embeddings or learning processes. Specifically, issues are the inability to fully understand what each model component signifies and encodes, coupled with the likelihood that these values are outcomes of skewed and weighted amalgamations of various factors, including the specific training dataset used. This makes them incompatible with the reality of human word processing, which is characterised by words having distinct, meaningful properties that represent them. The process of training models such as word2vec involves steps such as identifying skip-grams or relying on

common bags of words and performing negative samplings (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013), both of which also have no foundation in psycholinguistic research. Network representations of word-embedding (i.e. vector-based) models have furthermore been found to differ qualitatively from a large number of other language-based networks such as traditional semantic networks (i.e. WordNet). While vector-based networks also display small-world properties, the hub nodes are dominated by low-frequency words; in semantic networks, hub nodes overwhelmingly consist of high-frequency words (Veremyev et al., 2019). In conclusion, Generative Pretrained Transformer models depend too strongly on finetuning, training data selection, and overall opaque processes for furthering psycholinguistic research and vector-based linguistic networks do not display purely linguistic relations and are therefore also not suitable to represent psycholinguistically plausible connections.

Hybrid Approaches

Lastly, hybrid approaches used to identify word associations are discussed. There are multiple ways in which the abovementioned collocation extraction methods can be combined; one of these approaches is identifying suspect collocates through a method with high recall and low precision and filtering the results using one or more secondary methods as criteria later on. By the standard of psycholinguistic validity and logic, less restrictive AMs such as effect-size based measures could be used as a basis for flagging up potential collocates, and the results could be refined using a linguistically coherent rule-based approach such as LSM.

Another intuitively promising strategy is combining the rank orders achieved through different AMs and using this averaged rank list as a basis for the network generation. There are previous meta-studies such as Garcia et al. (2019, p. 57) make use of this approach. The authors posit that no single AM is capable to deliver both satisfactory precision and recall and they therefore called for this combination of measures. However, unsatisfying results and signs of overfitting have been encountered in a comprehensive study following this approach (Michelbacher et al., 2011, p. 270). The exact approach taken in this thesis is heavily adapted from this initial idea and described in detail in Chapter 4.2.1.4.

3.2.3 Descriptions of Individual Association Measures

After having explored the different types of available AMs and their properties, the following chapter provides brief descriptions of a range of AMs; these have been selected either on the basis of their popularity in Corpus Linguistics (MI, χ^2, etc.) or due to their promising properties with regards to psycholinguistic plausibility ($\Delta P_{forward}$, r_φ, etc.). These AMs do not represent the full range of AMs ever used in corpus linguistics (an ever-growing number) due to spatial limitations. The

decision to discuss the rather technical topic of statistical properties and assumptions of AMs in the main body of the thesis has been made since a commonly recognised problem in the field (Gries, 2012, pp. 47–48) is the use of statistics without a thorough motivation or methodologically adequate presentation of the results. The following sections are therefore also written to explain the AM equations to non-mathematicians alongside an exploration of underlying assumptions and the individual strengths and weaknesses of different AMs, all of which are relevant when assessing their potential for psycholinguistic validity in Chapter 3.3 and 3.4.

MI Scores

One of the most commonly used AM for collocation identification is MI scores. MI scores are based on a comparison of the expected frequency of a word combination (assuming a random distribution of words over the entire corpus) to the actually observed frequency of this combination. They therefore tend to overrepresent niche phrases involving overall rare lexical items (Simpson-Vlach & Ellis, 2010, p. 493).

$$MI = \log_2 \frac{O_{11}}{E_{11}} \quad \text{Equation 4}$$

$$MI^2 = \log_2 \frac{{O_{11}}^2}{E_{11}} \quad \text{Equation 5}$$

$$MI^3 = \log_2 \frac{{O_{11}}^3}{E_{11}} \quad \text{Equation 6}$$

$$MI^4 = \log_2 \frac{{O_{11}}^4}{E_{11}} \quad \text{Equation 7}$$

There are, however, two fundamental issues inherent to MI scores: Firstly, the focus on effect size exclusively without taking into consideration how statistically relevant the observed values are (Evert, 2008, p. 1227) and, secondly, the abovementioned assumption of a random distribution of words. The last point is particularly problematic since language is heavily influenced by grammatical rules and conventions such as a particular word order, adposition for case-marking in English and closed-class grammatical items in general. This inevitably results in word distributions that are markedly differ from randomness, directly violating the core assumption of MI scores. Beyond this, qualitative evaluations such as Evert and Krenn (2001) show poor overall performance of MI and a particular weakness at identifying Preposition-Noun-Verb triples. MI^2 has further been criticised for introducing a largely arbitrary skew (Gries, 2022b, p. 20), MI^3 and MI^4 suffer from the same problem. With regards to psycholinguistic plausibility, MI scores have thus been removed from the list of most suitable candidates for comparing collocation networks with psycholinguistic ones.

Poisson-likelihood and Hypergeometric Likelihood

In this section two examples from the family of likelihood-based scores are discussed: Poisson-likelihood and hypergeometric likelihood – both of which are inherently members of the group of statistical AMs that tests for significance only. The fact that these measures cannot be used to distinguish between positive and negative associations by default is noteworthy and requires careful filtering for applied purposes.

Poisson-likelihood has been chosen over multinomial or binomial likelihood here since it is based on the assumption of an underlying Poisson distribution rather than a completely random one. Poisson is a discrete probability distribution which approximates the probability of encountering an observed number of co-occurrences within a corpus given an expected number of co-occurrences. It operates based on the assumption that encountering a collocation does not in any way change the probability for encountering the same collocation again. While this assumption is overall less problematic than the assumption of linguistic randomness e.g. in MI scores, it still poses serious problems on the level of an individual discourse. If only a single discourse is investigated, the occurrence of a specific collocation once does influence the probability of it occurring again – i.e. in spoken discourse when a conversational partner makes use of the exact wording in their response or when a speaker reiterates their own point. On a larger scale, however, this issue is less severe. The occurrence of a specific collocation within the electronic language section of the BNC 2014 does not necessarily influence the probability of it occurring in a different genre such as official documents or newspapers.

Generally speaking, the Poisson likelihood approach (Kolesnikova, 2016, p. 335) is more realistic than many other AMs considering the nature of linguistic data and collocations in particular since estimating collocation frequencies essentially is Large Number of Rare Events (LNRE) modelling – for a more detailed discussion of the accuracy of AM assumptions in linguistics see the section on ΔP in this Chapter. A further advantage of the Poisson approach is the reduction in computational and componential complexity since it acts as an approximation of binomial distribution with the same mean which is much more costly to compute.

$$\text{Poisson likelihood} = \frac{\mathrm{e}^{-E_{11}} E_{11}^{O_{11}}}{O_{11}!} \qquad \text{Equation 8}$$

The second measure belonging to this family that is of interest here is hypergeometric likelihood (Kolesnikova, 2016, p. 335). This variant operates using row and column totals in addition to the observed frequency of co-occurrence only. It fundamentally relies on the hypothesis that the observed components of the suspect collocation are independent of one another rather than

directly relying on maximum likelihood. The idea behind hypergeometric likelihood is treating the observed frequency as a specific number of successes in a random draw (without replacement). It quantifies the likelihood with which the observed frequency was to be expected.

$$\text{hypergeometric likelihood} = \frac{\binom{C_1}{O_{11}} \times \binom{C_2}{O_{12}}}{\binom{N}{R_1}} \qquad \text{Equation 9}$$

Both Poisson likelihood and hypergeometric likelihood have a very high computational cost when encountering large values, for example when calculating the individual numbers of possible combinations; calculating $E_{11}^{O_{11}}$ for Poisson likelihood with very large values is also not possible on the machine available for this thesis since several operations for large observed and expected frequencies exceed the maximum of 4300 digits per integer in Python. A very rough extrapolation on the basis of 5 tuples that correspond to the minimum O_{11}, the 25th, 50th, and 75th percentile, as well as the maximum O_{11} indicates that the hypergeometric likelihood compute time required on the machine available for this thesis would exceed 200 hours. While the other shortcoming of these measures, their bidirectionality, can be remedied by indicating negative association when the expected frequency is higher than the observed frequency, the computational cost of hypergeometric likelihood and Poisson likelihood is prohibitive, and the measure will not be included in the network comparison at this time. This metric can still be applied in a context where the corpus size and thus the observed and expected frequencies are small or where normalisation is possible. Due to the large scale and comparative nature of the work presented here this is not possible for this project.

Poisson and Fisher's Exact Test

Measures based on hypothesis tests such as Fisher's exact test and Poisson have also been presented as possible candidates for collocation identification (Oakes, 2020; Rajeg, 2020) and are thus briefly explored here. Poisson is similar to the Poisson likelihood measure explored above in the sense that it is merely summed to include all values above the observed frequency up to an upper bound of infinity. In simple terms this means that Poisson provides researchers with a probability for all outcomes where frequency of co-occurrence is *at least* as large as the observed frequency of co-occurrence. Poisson is commonly favoured over a binomial measure for the same reasons as outlined in the section on Poisson-likelihood and Hypergeometric Likelihood in this Chapter.

$$Poisson = \sum_{k=O_{11}}^{\infty} e^{-E_{11}} \frac{(E_{11})^k}{k!} \qquad \text{Equation 10}$$

The second commonly used hypothesis-test-based measure is Fisher's exact test. The feature that makes this test stand out is its precision since it is, unlike other hypothesis-test-based measures, not reliant on the assumption that the observed relative frequencies in the respective sample (i.e. in the respective corpus) accurately represent the entire population (i.e. the entirety of the English language).

$$Fisher's\ exact\ test = \sum_{k=O_{11}}^{\min\{R_1,C_1\}} \frac{\binom{C_1}{k} \times \binom{C_2}{R_1-k}}{\binom{N}{R_1}} \qquad \text{Equation 11}$$

In general, all hypothesis-test-based measures share two major limitations: Firstly, Larger values are, perhaps counterintuitively given other AMs, indicative of a failure to refuse the null hypothesis that the observed frequency of co-occurrence is the result of a random distribution. Secondly, and unlike effect-size based measures, they furthermore give no indication as to how strong the collocation is; it merely informs how sure one can be that the frequency of co-occurrence is not random. This is especially problematic since an assumption of randomness is generally not applicable to language.

While Fisher's exact test and Poisson were discussed as a theoretical option here, they will not be implemented in the network generation system due to its unjustifiably high computational cost in addition to the universal drawbacks of hypothesis-test-based measures. The computational load may, however be reduced in the future by way of utilising deep learning techniques (Shan et al., 2017).

Chi Squared

Another AM of interest is χ^2 (Chi Squared), like log likelihood (discussed below) this ultimately tests for statistical significance of the observation made and cannot be interpreted as an effect size measure. For this reason alone, it is not suitable to be used as a standalone AM for representing psycholinguistic relationships between linguistic units; there would be no indication of the actual degree to which the observed collocational frequency differs from the expected collocational frequency under H_0.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad \text{Equation 12}$$

A simplified version of the above standard χ^2 calculation is used in line with Evert (2004).

$$\chi^2 = \frac{N(O_{11} - E_{11})^2}{E_{11}E_{22}}$$ Equation 13

It is, however, important to highlight that the question which χ^2 aims to answer when exploring collocational relationships is a more appropriate one than for Poisson or Fisher's exact test: The χ^2 statistic tests if the distribution of collocate frequencies in the corpus differs depending on whether or not the node precedes the collocate. Put differently, χ^2 tests if the relative proportions of collocates following the node and not following the node are the same (as assumed by the H_0). An examination of the collocational relationship of *black box*, for example, χ^2 would offer insights into whether or not the relative proportion of *box* or *any word other than box* following *black* differs from the relative proportion of *box* or *any word other than box* following *any word other than black*. Furthermore, χ^2 , unlike t-tests, does not rely on normally distributed probabilities (Kumova Metin & Karaoğlan, 2011, p. 179).

Due to the squared difference between observed and expected values χ^2 is never negative and thus a two-sided measure; it is therefore advisable to transform χ^2 scores according to Evert's procedure (Evert, 2008, pp. 1227–1228). This is recommended since it is, as previously discussed, relevant for collocational analyses to detail if the detected relationship is one of 'attraction' or 'repulsion' between node and collocate. A measure like χ^2 does not provide this distinction by design since both 'attraction' and 'repulsion' lead to a rejection of the null hypothesis of independence.

It is furthermore important to note that theoretically motivated thresholds for χ^2 values exist; since distributions with one degree of freedom are applicable due to the typical shape of contingency tables for collocations, as seen in Lindley and Scott (2018, Table 8) relevant values are the following:

Table 7: P values corresponding to χ^2 values for collocation identification.

P	1	0.5	0.1	0.05
χ^2	6.635	7.879	10.83	12.12

An adaptation of χ^2, χ^2 with Yates' correction for continuity, is also commonly used in corpus linguistics (Evert, 2008, p. 1235). This correction is applied in order to reduce errors for small values (< 5) in the contingency table, which is a common scenario when investigating collocational relationships. Yates' approach is not entirely undisputed in the statistical community (Hitchcock, 2009, p. 9) but nevertheless reasonably popular.

$$\chi^2_{Yates} = \sum_{i,j} \frac{\left(\left|O_{ij} - E_{ij}\right| - 0.5\right)^2}{E_{ij}} \qquad \text{Equation 14}$$

For this thesis, Yates' correction for continuity is generally not applied due to the reasonably large frequencies stemming from the relatively large amount of data. While some smaller individual subsections of the corpus are also investigated, the correction of continuity has the potential to create a false sense of precision. In order not to represent erroneous χ^2 statistics, therefore a value of 'N/A' for contingency tables with values < 5 is set.

While X2 approaches have been criticised before, for example by Dunning (1993a, p. 64) wo notes that it greatly overestimates rare collocations, χ^2 shows promising results for collocation identification on the basis of the BNC1994 when measured using collocation dictionaries as the gold standard (Evert et al., 2017, p. 537) and is thus included in the list of candidates for psycholinguistically plausible AMs.

Log likelihood

Another commonly used AM is log likelihood (LL), a two-sided likelihood ratio test. It is important to emphasise that LL, χ^2, and any other statistic described in this chapter, are evaluated as methods for *collocation* extraction, not as a means to identify lexical differences between corpora. While χ^2, LL and similar measures have been used for this purpose, base assumptions such as independence of observations are usually violated and the use of such measures is generally discouraged (Bestgen, 2018, p. 36; Brezina & Meyerhoff, 2014).

In a collocational context, LL relies on the observed and expected frequencies of the collocation at hand and all other combinations of its collocates. Evert (2008, pp. 1227–1228) provides a simplified and fully equivalent formula of Dunning's (1993b) original formula where O_{ij} and E_{ij} represent the four respective fields in the contingency tables for observed (O) and expected (E) occurrences. This will also be used as the basis for all LL calculations in this thesis.

$$\mathrm{LL} = 2 \sum_{ij} O_{\mathrm{ij}} \ln \frac{O_{\mathrm{ij}}}{E_{\mathrm{ij}}} \qquad \text{Equation 15}$$

The theoretically established cut-off points equivalent to common z-score cut offs are |LL|>3.84 and |LL|>10.83 (Evert, 2008, pp. 1227–1228).

Since one of the aims of this thesis is making the calculations underlying these measures more transparent to non-mathematicians in order to facilitate network interpretation later on, the

calculation is explained in greater detail here: At the core of LL is the natural logarithm (ln). This can be understood as a measure of the time required to reach a specific value, assuming a pattern of cumulative growth. When the observed value (O_{ij}) matches the expected value (E_{ij}), their ratio is one. This scenario corresponds to an ln(1) of 0, indicating no growth from the base value (i.e., 100% of the base value) and hence, no time required to grow. If the observed value exceeds the expected value, the ln provides insights into the duration needed to achieve this value, based on Euler's number. This scenario then yields positive values. Conversely, if the observed frequency falls short of the expected frequency, the ln indicates the time in the past when this fraction would have been achieved, resulting in negative values. Taking the full formula for log-likelihood into account, this means that LL values represent twice the sum of the product of the observed frequency and the expected time for cumulative growth to reach the ratio of observed to expected frequencies across all fields in the contingency table. Applying the ln to values close to 1 thus generates non-linearly larger absolute results than applying the ln to values further away from 1, i.e. very small or very large ratios of observed to expected results.

This measure deserves special attention since it has shown a relatively good performance when working with low expected frequencies, and it does, unlike for example MI scores, take statistical significance into consideration. Previous studies assessing how well collocations identified via log likelihood match up with association based scores – which is particularly relevant the aims of this thesis– showed that log likelihood outperforms a wide range of other AMs (Kang, 2018, p. 97). Log likelihood also outperforms 19 other AMs (among them χ^2) in a meta-studies evaluating the results obtained from using these AMs to identify collocations in the BNC 1994 against the Oxford Collocations Dictionary for Students of English as a reference (Evert et al., 2017, p. 538; Uhrig et al., 2018, p. 111).

In addition, log likelihood can be used as an approximation of the more computationally expensive Fisher's exact test (Evert, 2008, p. 1235; Gries, 2013, p. 148). This is relevant on a theoretical level, since it is based on a hypergeometric distribution rather than the assumption of a normal distribution – generally speaking this models drawing random word combinations from the corpus (similar to drawing prototypical balls from an urn (Gries & Ellis, 2015, p. 237)).

While the abovementioned points make log likelihood a good candidate for psycholinguistic plausibility, two key limitations need to be acknowledged. Log likelihood, again, tests for statistical significance, not effect size. As such, it only provides the researcher with information about the quality of the data available and the certainty with which claims can be made, but it does not indicate the strength of collocation per se. Another limitation is the strong tendency of LL to overrepresent collocations of constituent words with high frequencies over low frequency constituents even when

the log ratio of the collocation in question stays constant. Gries (2022a, 5–6) therefore considers it a "measure that reflects mostly frequency and also some association" rather than an association measure. It is further imperative to apply the abovementioned correction for bidirectional measures to LL-scores. All LL-scores in this thesis have been transformed to be monodirectional.

ΔP

One completely different Association Measure both in terms of its theoretical foundations and its data structure for collocation analyses is delta P (ΔP). ΔP is a directional AM and it can be defined as the probability of word2 given -word1 (i.e. the probability of word2 in the absence of word1) subtracted from the probability of word2 given word1. This concept is closely linked to the translational probabilities employed in psycholinguistic research as mentioned in Chapter 2.5.1. ΔP measures covariation; at ΔP = 0 no collocational pattern is found as the presence or absence of a word1 does not influence the behaviour of word2 in this case (Ellis, 2006, p. 11). A positive association will approach 1, a negative association -1.

Generally, the equation for calculating ΔP is as follows (Gries, 2013, p. 143; Perruchet & Poulin-Charronnat, 2012, p. 123):

$$\Delta P = p\,(outcome \mid cue_{present}) - p\,(outcome \mid cue_{absent}) \qquad \text{Equation 16}$$

In a collocational context, the following three sub-cases are used. The first of the equations describes $\Delta P_{forward}$ (i.e. how good of a predictor is word1 for word2 following it), while the second describes the inverse, $\Delta P_{backward}$ (i.e. how good of a predictor is word2 for the word preceding it). Lastly, ΔP is calculated as the difference between the two.

$$\Delta P_{forward} = \frac{O_{11}}{R_1} - \frac{O_{21}}{R_2} \qquad \text{Equation 17}$$

$$\Delta P_{backward} = \frac{O_{11}}{C_1} - \frac{O_{12}}{C_2} \qquad \text{Equation 18}$$

$$\Delta P = \Delta P_{forward} - \Delta P_{backward} \qquad \text{Equation 19}$$

In evaluative studies exploring the similarity between $\Delta P_{forward}$ and $\Delta P_{backward}$ such as Gries (2013, pp. 146, 148), differences between the two statistics were found for about 25% of the resulting bigrams when using the Spoken BNC 1994. This indicates that 75% of bigrams represent conflated, bidirectional ΔP values.

A feature that distinguishes ΔP from a whole range of other AMs is that it does not function as a significance test and can therefore not provide insights into the robustness of the findings. A

drawback on the one hand, it also avoids relying on problematic – or simply factually wrong – assumptions such as proposing that language is random on the other hand. ΔP has been found not to be highly frequency dependent (Gries, 2022b, p. 7).

What makes ΔP further stand out from most other and more commonly used AMs is the potential for ΔP to provide a way to align textual data analysis with psycholinguistic findings. Its underlying concept, namely contingency, is also features heavily in theories regarding learning processes (Ellis, 2006, p. 10; Gries, 2013, p. 152). Bearing in mind the generalization commitment ΔP also deserves special interest since, as Shanks (1995) theorises, it is plausible that a high contingency as expressed through ΔP combined with a reasonably high overall frequency of occurrence – which translates into exposure to the repeated, contingent pattern – influences human associative judgements on a universal scale. Beyond this, ΔP can also be used for more specialised research questions e.g. surrounding noun phrase construction since its directional nature gives an indication as to which word in the construction should be considered the headword (Gries, 2013, pp. 152, 156). This might be particularly relevant in language learning and psycholinguistic contexts.

While ΔP is an established metric in psycholinguistics, especially in a language learning context, a few studies such as McConnell and Blumenthal-Dramé (2019, p. 22), in this case in a reading time study, found ΔP to have poor predictive power. The conflicting evaluations make it all the more important to explore large collocation networks based on ΔP / $\Delta P_{forward}$ / $\Delta P_{backward}$ and discuss the emerging structural differences and similarities when compared to word association networks. The network approach is also particularly apt since ΔP has been found to work well when exploring multi-word units involving more than two lexical items (Gries, 2013, p. 154); networks have the capability to represent coherent non-fixed length high-ΔP 'chains' since all high ΔP scores directly link the relevant words with a greater force than low ΔP scores.

r_φ

Another AM of interest for this exploration is Pearson's r (r_φ) which essentially expresses the geometrical mean of both the forward ΔP and the backward ΔP (Perruchet & Poulin-Charronnat, 2012, p. 124); making it a bidirectional measure.

The corresponding formula is given below:

$$r_\varphi = \sqrt{\left(\frac{O_{11}}{O_{11}+O_{12}} - \frac{O_{21}}{O_{21}+O_{22}}\right)\left(\frac{O_{11}}{O_{11}+O_{21}} - \frac{O_{12}}{O_{12}+O_{22}}\right)} \quad \text{Equation 20}$$

r_φ might present a valuable addition to raw ΔP scores in situations where comparisons with other bidirectional measures are necessary. r_φ also shares ΔP's potential for psycholinguistic relevance;

providing better results than forward ΔP and backward ΔP individually in a syllable processing evaluation (Perruchet & Peereman, 2004, pp 104, 111). It needs to be acknowledged, however, that the linguistic element of interest in their specific study is not collocations and collocational relationships; this rather merely presents evidence for general psycholinguistic relevance of the measure. The underlying principles that have been elaborated on in Chapter 2.5.1 are expected to hold true (albeit quite possibly in an adapted form) despite the different nature of the element under observation. This representation of collocational strength that correlates not only with the frequency of occurrence of the data, but also the contingency of the individual collocate in question fits the theoretical foundation of this thesis very well and makes r_{φ} a good candidate for creating networks with potential for comparability to psycholinguistic data.

(log)Odds Ratio

This section discusses two more recently suggested AMs, OddsRatio and logOddsRatio. These measures have been suggested as independent from frequency effects, and thus a true measure of association ***only*** (Gries, 2022a, p. 14, 2022b, p. 5; Gries & Durrant, 2020).

$$oddsRatio = \frac{O_{11}}{O_{21}} / \frac{O_{12}}{O_{22}} \qquad \text{Equation 21}$$

$$logOddsRatio = \ln \frac{O_{11}}{O_{21}} / \frac{O_{12}}{O_{22}} \qquad \text{Equation 22}$$

The example below illustrates the approach taken to calculate OddsRatio and logOddsRatio values. This entails observing the ratio of the number of observed tuples consisting of node and collocate versus observed tuples beginning with the node but not ending in it. This ratio is then compared to the ratio of tuples beginning with anything but the node and ending in the collocate and all tuples neither beginning with the node nor ending in the collocate.

good, morning (observed 20 times)	good, ~~morning~~ (observed 70 times)
~~good~~, morning (observed 50 times)	~~good~~, ~~morning~~ (observed 1000 times)

In this case, the ratio of *good*-starting tuples that are *good morning* is $20/70 \approx 0.29$. The ratio of tuples that end in *morning*- of all other tuples that neither begin with *good* nor end in *morning* is calculated at $50/1000 = 0.05$. This leads to an OddsRatio of $0.29/0.05 = 5.8$. This value is larger the larger the first ratio and the smaller the second ratio. Applying the natural logarithm to results in a value of ≈ 1.76 in the present example.

(log)Dice

The last canonical statistical measure to be considered here is (log)Dice (Dice, 1945).

$$Dice = \frac{2O_{11}}{R_1 + C_1} \qquad \text{Equation 23}$$

Promising results have been achieved using Dice to detect collocations for lexicographic purposes and when compared to a collocation dictionary (Evert et al., 2017; Kolesnikova, 2016, p. 340; McKeown et al., 1996, p. 7). The measure has a few other particularly beneficial properties considering psycholinguistic plausibility: It does not rely on the base assumption of a random distribution. Instead, Dice measures the proportion of the doubled collocation frequency to the sum of the row and column total. In practice, this means that Dice evaluates the total exclusivity of the collocation. A top-scoring collocation for Dice in the complete BNC 2014 is *troilus_N|criseyde_N* reaching a value of 1. Examining the corresponding contingency table below (

Table 9) it becomes clear that the values are symmetrical, and *troilus_N|criseyde_N* are perfectly exclusive; there are no occurrences of one without the other. The minimum value observable for R_1 is O_{11} (since the total occurrence of the word *troilus_N* for example cannot be lower than the frequency of *troilus_N* as part of the collocation, and the same is true for *criseyde_N* and C1. The less exclusive node and collocate are, the smaller the Dice value. Another important feature of (log)Dice is that it is independent of frequency effects. This is desirable for AMs to maintain a separation of frequency from association, and to enable independent analyses of either property of a given collocation (Gries, 2022b). Dice values obtained for a more or less frequent, yet still perfectly exclusive, collocation (e.g. *saltzmann_N|tincture_*N) are identical (see Table 10) which sets it apart from any AM that relies on expected frequencies which, in turn, take the total number of tuples into account.

Table 8: Concordance lines for *troilus_N|criseyde_N* (Dice; sentence-span; all sections, frequency threshold 1wpm in each BNC 2014 subsection).

SENTENCE ID	CONCORDANCE LINE
FICTHIS10_1255	I put it between the pages of Troilus and Criseyde.
FICTHIS10_717	The only book I wanted, Troilus and Criseyde, had not been among the things delivered to me at the Tower.
FICTPOE5_370	Troilus and Criseyde meet that night.
FICTPOE5_760	Troilus and Pandarus go to meet Criseyde.

Table 9: Contingency table for *troilus_N|criseyde_N* (Dice; sentence-span; all sections, frequency threshold 1wpm in each BNC 2014 subsection). Shaded fields used for Dice calculation.

	criseyde_N	criseyde_N	
troilus_N	4 [O_{11}]	0	4 [R_1]
troilus_N	0	212,863,972	212,863,972
	4 [C_1]	212,863,972	212,863,976

Table 10: Contingency table for *saltzmann_N|tincture*_N (Dice; sentence-span; all sections, frequency threshold 1wpm in the relative section). Shaded fields used for Dice calculation.

	tincture_N	tincture_N	
saltzmann_N	9 [O_{11}]	0	9 [R_1]
saltzmann_N	0	212,863,967	212,863,967
	9 [C_1]	212,863,967	212,863,976

A $\log_2$ transformed variation of this, log Dice is also commonly employed.

$$logDice = 14 + \log_2(\frac{2O_{11}}{R_1 + C_1}) \quad \text{Equation 24}$$

log Dice operates on a fixed scale of -∞ to 14 (Gablasova et al., 2017, p. 164; Messaoudi, 2019, p. 224; Rychlý, 2008, p. 9) which makes it easily comparable and aids the visualisation process. The application of the logarithm re-shapes the AM values to roughly follow a normal distribution. While having originated from a different statistical school of thought, the Jaccard coefficient is fully equivalent to Dice via a monotonic transformation. This is mentioned here since both Dice and Jaccard have been used extensively in collocation research.

$$Jaccard = \frac{O_{11}}{O_{11} + O_{12} + O_{21}} \quad \text{Equation 25}$$

T-Score

The T-Score is designed to determine the confidence with which it can be claimed that there is some association between words. T-Score historically has been (Gries, 2022b, p. 14; Kang, 2018, p. 91; McEnery et al., 2006) and still is a popular choice in corpus linguistics, a cutoff of two or higher is normally considered to be statistically significant (ibid.). T-Score has been described as mirroring human intuition particularly well (Kang, 2018, p. 91). It is calculated by subtracting the expected frequency from the observed frequency and then dividing the result by the standard deviation.

$$T - Score = \frac{O_{11} - E_{11}}{\sigma}$$ Equation 26

This measure does, however, ultimately also rely on the hypothesis that bigrams are generated randomly (Kumova Metin & Karaoğlan, 2011, p. 179). T-Score has further been found favour frequent combinations to the extent of being almost entirely predictable from frequency only (Gries, 2022b, p. 16), but studies evaluating it in psycholinguistic settings report poor performance (McConnell & Blumenthal-Dramé, 2019, p. 18)[13].

Z-Score

The Z-Score is, in terms of computation, is quite similar to t-score presented above. It determines the number of standard deviations from the mean frequency.

A commonly employed cut-off value for z-scores is $|z|>1.96$ (Evert, 2008, p. 1227)

$$Z - Score = \frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$$ Equation 27

While both Z-Score and T-Score are one-directional measures and thus do not conflate positive with negative association, they are both following the assumption that linguistic events are random as further explained in Chapter 3.3. Z-Scores and T-Scores are therefore not considered as good candidates for psycholinguistic plausibility.

3.2.4 Other Collocation Extraction Parameters

After having explored AMs and their theoretical foundations and suitability it is important to also consider other parameters that greatly influence the result of the collocation extraction process. Parameters to be considered here are directionality, symmetry, window span, and, most foundationally, unit of analysis. Since each of these factors chiefly influences the emerging networks and should always be considered before carrying out collocation analyses the following three Subchapters are devoted to explaining the role they play in collocation extraction processes.

Directionality and Symmetry

Directionality is highly relevant for the suitability of individual AMs for comparisons with psycholinguistic data. AMs that are not calculated on the basis of directed counts do not allow for distinguishing e.g. between occurrences of *white house* and *house white* and are thus only of very limited usability for comparisons to structures of the ML. Additionally, the English language has been

[13] A notable limitation in McConnell & Blumenthal-Drame (2019) is their online data collection method, which offers less control compared to in-person experimental settings. Their participants were also encouraged to take their time when reading which deviates from most other experimental setups in processing research.

found to be overall right-predictive as supported by elicitation experiments (Michelbacher et al., 2011, p. 259), which would inevitably skew the outputs should directionality not be taken into account. Another example demonstrating the importance of retaining directionality information is more complex: Participants have been found to prefer using the form "X is virtually Y" with X representing a sub-category or example and Y being a prototype based on rating-based and choice-based experiments spanning both production tasks and comparative tasks. This general tendency was found for a wide variety of contexts such as shape recognition, similarity ratings of letters, numbers and colours, countries, and other categories (Tversky, 1977, p. 337). One example Tversky provides for this is a much higher proclivity of participants to produce "103 is virtually 100" when compared to "100 is virtually 103". In Tversky's own words, we can therefore assume that "similarity is not necessarily a symmetric relation". Practically speaking, a comparison of cue-response pairs generated (and thus directional) networks and undirected collocation networks is therefore futile.

The aforementioned right-predictivity of English may raise questions as to why log transformed backward translational probabilities in the shape of $\Delta P_{backward}$ are considered in this thesis. Backward predictions are of interest since empirical studies indicate their psycholinguistic relevance in that lexical processing depends on backward integration difficulty. This concept refers to the difficulty of integrating words with previously encountered information, which generates a measurable delay in lexical processing (McConnell & Blumenthal-Dramé, 2019, pp. 5–6). Since there is a large number of empirical studies (Ellis, 2006, p. 11; McConnell & Blumenthal-Dramé, 2019, p. 2) demonstrating perceptual effects based on forward-looking translational probabilities ($\Delta P_{forward}$), both networks are considered as possible candidates for psycholinguistically plausible collocation networks here.

While a special type of network containing information pertaining to backward and forward translational probabilities at the same time may be desirable, it is not trivial to represent this visually in comparative linguistic graphs. This is the case since the weight of a collocational relationship between two words indirectly determines their plotted distance relative to one another in edge-weighted graphs. While it is possible to draw two separate edges between two nodes (in the above example one edge from *house* to *white* and one and from *white* to *house*) and assign these their respective weights as per their collocational strengths, it is impossible to unambiguously determine their distance taking the full complexity of both values into account; see illustration in Figure 16. Another added layer of complexity stems from the fact that some colocations will be unidirectional (such as *white house*) whereas others will be bi-directional (such as *straight up* and *up straight* as part of the phrase *stand|sit up straight*) thus making it impossible to solve this problem by simply averaging

both directional weights without skewing the systematically results leading to tighter edges between bidirectional colocations. While this is less problematic in many non-linguistic graph theoretic applications, it is highly relevant when representing psycholinguistic and cognitive phenomena. In this thesis, forward and backward translational probabilities are thus represented in separate networks.

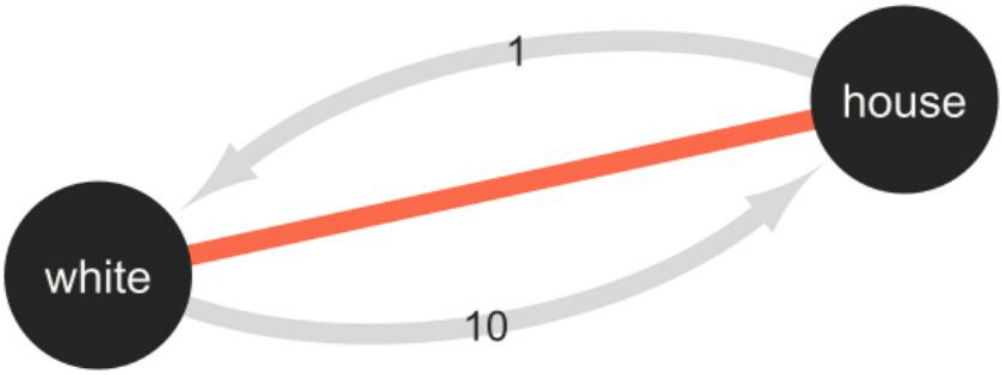


Figure 16: Issues regarding the visual representation of di-graph distances for bidirectional collocations. Should the distance marked in red represent 1, 10, or the average of 5.5 units?

Window Spans

After selecting an AM based on the research question and corpus a hand, it is important to choose the parameters for the individual AM with equal care. The choice of window span is another factor allowing the capture of different types of collocations and relationships between words. Looking at the disciplinary tradition, most early studies have rely on collocation window spans of two, mainly for practical reasons. (Ferrer-i-Cancho & Solé, 2001, p. 2262). However, the technological advancements of the past two decades meant that other approaches have also become viable. A large number of more recent studies chose window spans of 3 to 5 words (Evert et al., 2017, p. 532) and some studies explicitly exclude 2-word collocations since they expect the bulk of the relevant bigrams to be indirectly evident in larger collocations (Simpson-Vlach & Ellis, 2010, pp. 491–492). Yet more exhaustive approaches look at sentence or paragraph-level co-occurrences in order to identify larger, meta-level relationships and to evade missing collocations that have been spatially separated due to grammatical constraints or sentence structure. Issues would, for instance, arise when analysing "He **kicked** the proverbial **bucket**" or "He was riding **from dawn** – when the message reached him – **to dusk**." using a window-size that only captures up to five words either side of the word in question. These contrasting approaches to window-size illustrate a central issue in terms of the comparability of studies investigating collocations: The richness in available

parameters and underlying definitions of collocations leads to a rather fragmented field consisting of individual, widely incomparable studies.

Based on the aims of this study to investigate psycholinguistically plausible phenomena, this thesis uses sentence-span windows for optimal recall despite this choice being suboptimal for comparability with existing collocation studies. This decision is partially also motivated by ERP studies which show that semantic and syntactic processing are intertwined during on-line processing (Yamada & Neville, 2007, p. 177). One way of ensuring this semantic relationship is accounted for without implementing full syntactic tagging is generating collocation statistics for all elements within the same sentence (albeit maintaining directionality). It is essential to note here that the sentence-span tuples generated for this thesis are not weighted according to their closeness to the node word. This follows the same underlying principle as not applying window-span corrections: Applying these measures would lower overall interpretability since the effect of the graded weighting or window-span correction is not immediately predictable. It cannot be assumed that the true collocational closeness of sentence-span tuples reflects a linear picture; the examples above illustrate that sentence final elements might exhibit stronger collocational connectivity than other words that spatially lie between node and collocate. In order to avoid these distortions, no weighting is introduced.

Unit of Analysis

Lastly, the unit of analysis chosen to apply AMs to is inevitably crucial for the outcome of collocation explorations. Common choices for collocation extraction are either the original types, normalised types, lemmas, semantic categories, or – in the case of colligations – grammatical categories. Which of these units is most suitable depends entirely on the purpose of the research. Lemma-based collocations, for instance, are a good indication for more general semantic connectivity (Brezina, 2018, p. 64) since they combine all inflectional forms of a head word into discrete bins. POS-tagged lemma-based collocations are another popular choice. They are, however, highly dependent on tagging accuracy and fail to capture fine-grained differences in usage patterns and, in turn, differences in collocational relationships. A large number of studies therefore do not lemmatise the corpus before applying AMs (Dayrell, 2007, p. 381). While a range of customisation options are available via the LLN processing pipeline included in the appendix of this thesis, the decision was made to use lemma as the base unit for all comparisons between the BNC 2014-based and the SWOW-based networks. This decision has been made to ensure a semantic focus rather than a full investigation of morphological variation. While a full investigation on the basis of lemma and POS membership would have been highly desirable when considering that, e.g. cut_VERB and cut_NOUN fulfil different communicative functions, this was not

possible in the present study. Initial assessments of possible POS-tagger accuracy on the sparse data obtained from the word association component which largely consists of one-word responses were discouraging. The POS-tagged BNC 2014 pipeline is therefore only provided as a resource for future work but not used for network-level comparisons.

3.3 Identifying Word Association Measures with Psycholinguistic Validity

After a discussion of the most relevant strategies and AMs for generating collocation networks suitable for meaningful comparisons to word association networks, the objective of this chapter is exploring the possibility of creating large corpus word collocation networks using techniques that align with psycholinguistic findings. The result of this chapter is a comprehensive elucidation of the foundational premises of various association measures, and an assessment of whether these premises make each association measure unsuitable for approximating cognitive processes, language comprehension and production. To support ongoing corpus linguistic research, whether it involves network analyses or not Figure 17, a flowchart for selecting association measures that are plausible from a psycholinguistic perspective is provided. While this chapter covers a large number of AMs, several existing and commonly used AMs were not considered as candidates for this study due to their limitations; these are explored in this Chapter.

The flowchart indicates that black box methods, such as word vectors/embeddings often used in popular machine learning algorithms like BERT (Devlin et al., 2018), GloVe (Pennington et al., 2014), or Word2Vec (Mikolov, Chen, et al., 2013) are not conducive to advancing psycholinguistic knowledge. The first level of the decision-making process therefore concerns whether or not the method used is explainable and reasonably transparent. Deep learning models and other black-box methods based on word embeddings, which are not built on logical principles, fail to provide this interpretability. Beyond the immediate explainability issues arising from the design of the method, additional interpretability problems arise when examining the decisions made in tuning mainstream processing pipelines. Various degrees of brute-forcing and randomness are a standard practice when using GloVe or Word2Vec and related systems that rely on opaque word embeddings.

A further related aspect is the fact that the good performance of such models does not necessarily reflect unique findings based on the sophistication of vector space models. First order co-occurrences have been used to extract structural semantic representations with a similar success rate (albeit in a larger dataset) (Louwerse, 2021). This means that the underlying pattern picked up by vector-based representations can be accessed with a much-improved interpretability via word co-occurrences. In contrast to this, harnessing the adaptability and immediate interpretability of simple AMs therefore holds great potential for any assessment of potential causal relationships

between word properties, their frequencies, and their semantic representation. These issues are not limited to Word2Vec, underlying processes of BERT are also not fully understood (Kovaleva et al., 2019). Since the primary purpose of carrying out comparisons between textual and the underlying linguistic processes actually work – cannot be achieved using opaque models that do not rely on neurologically plausible approaches, they have been excluded as candidates for measuring psycholinguistically valid linguistic relationships.

Following this distinction, another fundamental property that might lead to a direct exclusion of certain measures needs to be considered: The possible conflation of positive and negative association into a single score. Certain two-sided measures such as Likelihood Ratio and LL will, by default, result in high scores if there is a strong association between a potential node and collocate is identified, regardless of whether this difference is an effect of node and collocate co-occurring considerably *less* often than expected. Measures that follow this approach, mainly likelihood-based measures as well as measures such as χ^2, struggle with the distinction between positive and negative association when analysing small numbers of occurrences (Evert, 2005, pp. 242–243) and are therefore particularly problematic. Without corrections (some of which are proposed in Evert (ibid., p. 76)), these will be considered unsuitable for psycholinguistically valid investigations since they would greatly distort the shape of the resulting network. Focusing on purely positive collocations or, preferably, assessing positive and negative collocations individually, are the only psycholinguistically sound options for word association identification.

The next step towards identifying valid measures is checking whether or not directionality can be captured by the respective method. If this is not immediately the case but theoretically achievable using corrections such as the procedure proposed in Michelbacher et al. (2011, p. 255), the measure can still be considered potentially suitable for psycholinguistic purposes. A measure that captures asymmetry by default is, however, preferable since this does not result in ranking the obtained results and bypasses potential issues around ties and scaling of the ranked results.

The next stage of the decision-making process is slightly more complicated: It is essential to consider whether or not the method in question relies on POS tagging. If it does, this might help improve the overall accuracy by enabling the researcher to exploit complex linguistic dependencies and potentially also tagging that includes specific metadata. The major drawback of this procedure is the fact that no automatic tagger, some might argue no tagger in general, will achieve perfect accuracy. If the errors introduced into the dataset through tagging are systematic, there is a risk of distorting shape and properties of the entire network based on resulting effects. While POS-tagging is therefore highly problematic in the given psycholinguistic context, an investigation of specific

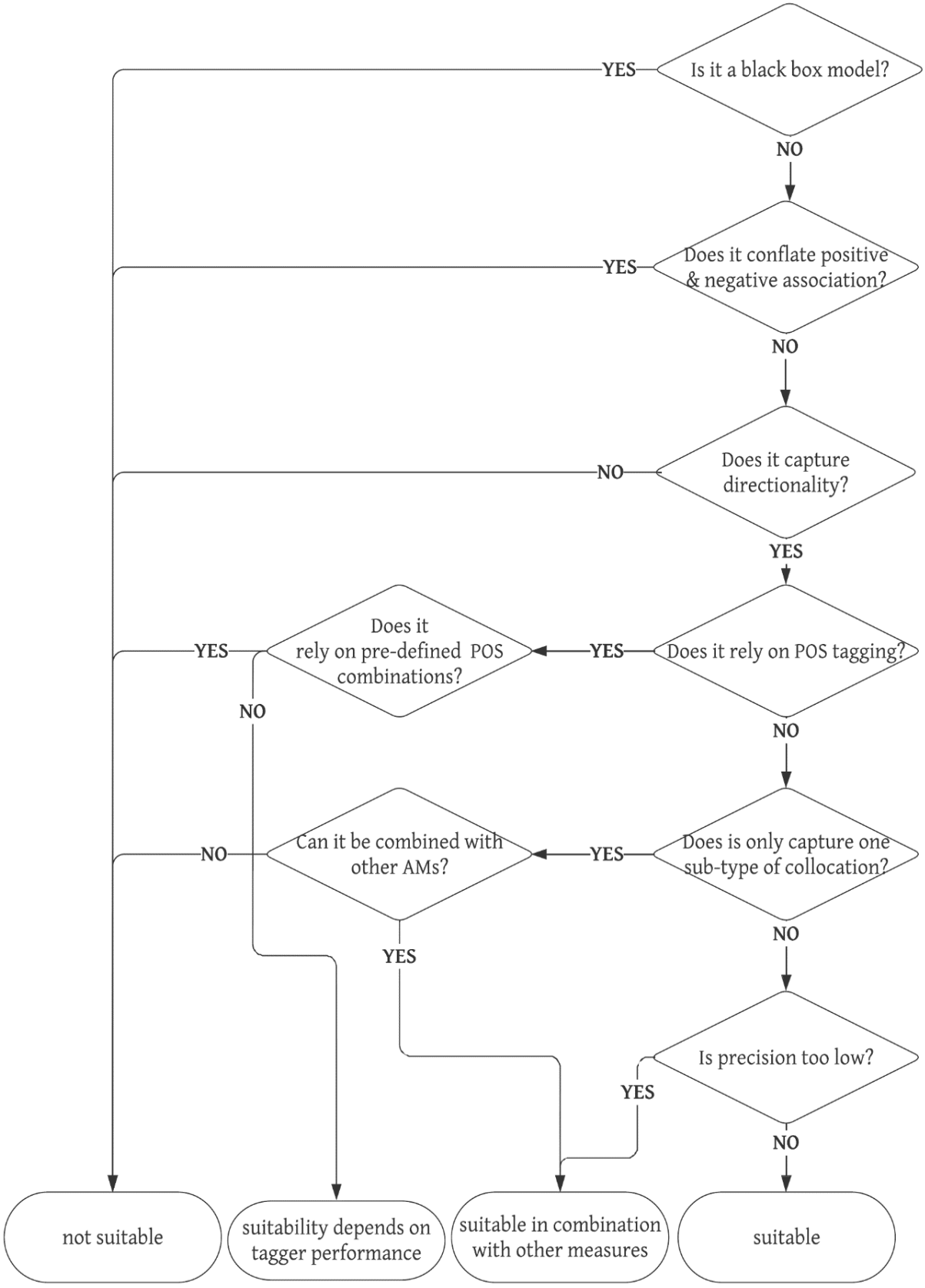


Figure 17: Flowchart for the selection of psycholinguistically valid word association approaches.

more restrictively defined types of collocation (e.g. phrasal verbs or compound nouns) simply requires it. POS filters will therefore be incorporated in the network generation pipeline, but it is essential acknowledge major limitations arising from employing such filters. Firstly, this step limits the applicability of the pipeline to low resource language since the POS tagger performance is often not adequate in these contexts. Secondly, several methods for identifying collocations rely on POS tagging as a pre-filter before an exploratory analysis is employed, this does not allow for examining latent patterns and strongly foregrounds the intuition of the researcher. This is not consistent with the larger approach at hand for this thesis and will therefore be considered unsuitable for this thesis. Despite this, these techniques warrant a mention as they may be relevant for other projects – one of these methods is LSM (Wermter, 2008, p. 109) which relies on a fixed combination of pre-defined POS constructions. This imposes restrictions on the maximum size of window spans that can be used, but it typically yields satisfactory results when examining specific pre-defined types of collocation.

If POS tagging is not essential for the method in question, a further differentiation can be made depending on whether or not the measure is strongly and systematically biased towards a specific sub-type of collocation such as lexical collocations or grammatical collocations. If so, it will be considered unsuitable unless it allows for a meaningful combination with another measure or measures with an inherent bias resulting in opposite effects. Recombination in itself, while less methodologically straightforward, may be generally desirable. While this is not a current main-stream approach in Corpus Linguistics, a number of important studies such as Bartsch (2004) combining MI and z-score, Hamilton et al. (2007) combining MI and t-score and Gabrielatos and Baker (2008) combining MI and LL. A combined approach requires normalisation and ranking strategies as well compatibility of window spans and thresholds, amongst other parameters, in order to achieve a meaningful balance between biased measures. Since this is a particularly promising avenue of methodological innovation combined AMs have been included in the network comparison evaluation in this thesis (Chapter 4).

The last layer of the proposed decision-making process concerns precision and recall rates. While selecting a method that captures all forms of word associations very broadly - such as raw frequency of occurrence approaches - has thus far seemed like a particularly effective strategy, this might not amount to the most meaningful results. Crude approaches like this benefit from their simplicity and intuitively logical nature, but the amount of noise they introduce into the final dataset which will then be graph-theoretically analysed and visualised is suboptimal. It is therefore necessary to either find measures that result in a satisfactory balance of precision and recall or to combine low

precision, high recall measures with other, secondary methods to filter out the least relevant word combinations.

Finally, when applying this framework to the pre-selection of AMs and other methods for collocation identification made here, the following broad classification emerges: Machine learning based models such as word2vec are black box models and thus considered unsuitable for the present approach until major breakthroughs regarding the explainability of the determined hyperparameters can be achieved. Linguistic rule-based approaches such as LSM are also considered to be barely suitable due to the obvious limitations of having to rely on automatic tagging as well as the reintroduction of the human and intuition-based component that the approach taken here intends to minimise. These methods will therefore only be considered as a possible final soft filtering step should other methods amount to an excessive number of word association pairs. Traditional effect-size approaches, significance-based approaches and approaches from information theory, however, emerge as the strongest candidates for identifying word association measures with psycholinguistic validity.

3.4 Conclusion: Psycholinguistically Plausible Corpus-Wide Collocation Networks

This Chapter concludes a long and detailed exploration of theories underlying the investigation of word association and collocation, the status quo of psycholinguistic research relevant in a corpus linguistic context, available graph theoretical parameters and Association Measures. All of these discussions are aimed to bring together corpus linguistic methodology, psycholinguistic theories, as well as graph theory. This results in an answer to RQ1 of this thesis which asks how large corpus-wide collocation networks can be generated using methods that are aligned with current findings from psycholinguistics. Since the networks themselves have been shown in Chapter 2.2.1 to be well suited for representing the mental lexicon due to their complex and dynamic nature, the last, and most fundamental, question remains the question of what should determine the selection of nodes and edges – in other words, the AM.

Psycholinguistic plausibility is defined as unifiability with current theories such as the cognitive commitment (Lakoff, 1991, pp. 53–55) and experimental findings from psycholinguistics, these have been explored at length in Chapter 2.5. The key findings show that a range of factors contribute to language learning, storage in the mental lexicon, and language production. Key features which are quantifiable on the basis of corpora are frequency, exclusiveness, translational probability, context, and – by way of network analysis – also connectedness. It is once again central to emphasise that this list omits features that cannot be reliably accessed via corpus-based analyses such as emotional state, social interaction between speakers, age effects, etc. The presented findings

and critical evaluations show that a connectionist approach to collocation benefits a meaningful exploration of these different layers of linguistic information via the choice and combination of relevant AMs in order to capture features such as frequency, cohesion, and translational probability.

Therefore the answer to RQ1 is that it is imperative to consider the following key factors in order to generate corpus-wide collocation networks that are aligned with current findings from psycholinguistics:

- Choice of Association Metric
- Choice of possible Filters
- Choice of Collocation Window
- Directionality
- Choice of Graph Theoretical Analytics
- Limitations to a Single Layer of Linguistic Representation

Since a major contribution of this work is aiding with the establishment of a psycholinguistically valid approach to selecting Association Metrics, Figure 18 provides an updated flowchart positioning the approaches relevant to this thesis. Employing this on the presented AMs results in the following remaining candidates (and their combinations) for further network generation and structural comparison to word associations: (log)Dice, ΔP, $\Delta P_{forward}$, $\Delta P_{backward}$, LL, $\chi 2$, Poisson, (log)Odds Ratio, and $r\varphi$.

MI-based scores, T-score, Z-score, hypergeometric likelihood, and Fisher's exact test have been excluded since they either violate basic assumptions or have an unjustifiably high computational cost in the context of the project at hand. MI-scores have been deemed unsuitable due to their reliance on the assumption that language is random as discussed in greater depth in Chapter 3.2.3. Similarly to MI-scores, T-tests are deemed unsuitable for the present study due to their methodological foundation (Evert, 2005, pp. 82–83). The fundamental assumption here is that the observed nodes and collocates are independent of one another and follow an identical distribution across all source texts – an assumption that does not hold true for the majority, if not all, corpora.

Z-Score and T-Score in particular are also problematic from a purely theoretical perspective, albeit for different reasons. The use of Z-Scores should be carefully considered due to the oversimplification when substituting a discrete (binomial) distribution with a continuous normal distribution in order to make the calculations more efficient – this approximation without corrections is at high risk for producing distorted results. T-scores on the other hand are mainly

problematic due to the fact that the basic conditions for a Student's T-test (one distribution containing n dependent normal variates) are not met.

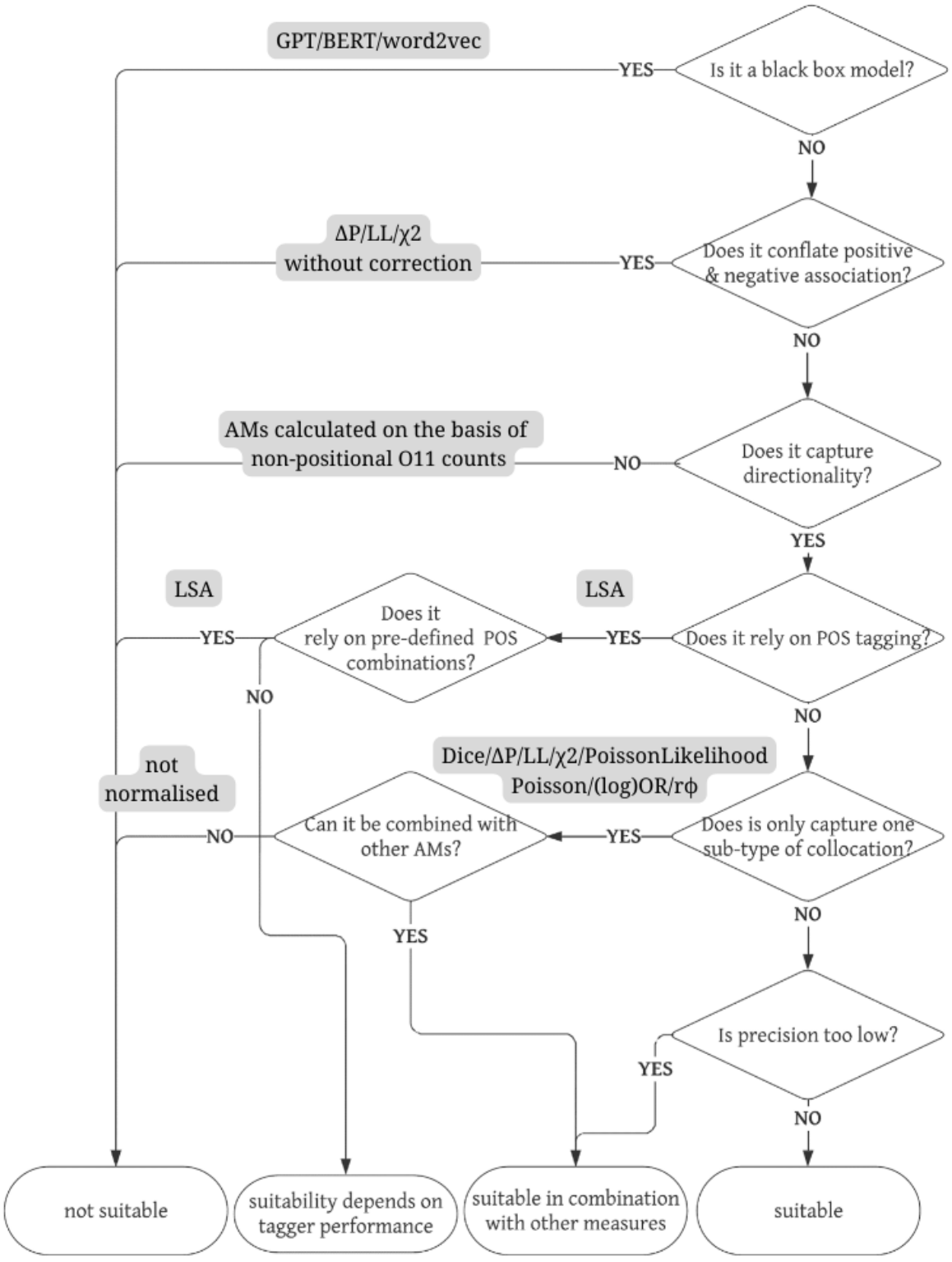


Figure 18: Flowchart for the selection of psycholinguistically valid word association approaches showing the location of measures considered for this project.

The choice to exclude MI^3, T-Score, and Z-Score has been reinforced by their subpar performance in a meta study on cognitive realism by McConnell and Blumenthal-Dramé (2019, p. 18). In this study, the authors investigate the correlation between collocations identified by various AMs and reduced reading times in psycholinguistic experiments as opposed to non-collocations. It is, however, crucial to mention that the study solely focuses on nouns and modifiers and thus does not offer insights into the applicability of AMs for examining grammatical structures or the ML in its entirety. Other non-favourably performing measures such as LL and Dice have therefore been permitted into the testing set for further exploration of their network properties due to their theoretical suitability.

4 Empirical Evaluation (RQ2 and RQ3)

Contrasting BNC 2014-based holistic collocation networks with the SWOW-EN word association network

4.1 An Introduction to Contrasting Holistic Collocation Networks with Word Association Networks

Transitioning from the comprehensive exploration of theories, methodologies, and measures in the previous chapter, this Chapter brings all these foundations together to empirically contrast holistic collocation networks with word association networks. This chapter is designed to address RQ2 and RQ3, focusing on the practical application of the theoretical foundations laid out so far.

The previous Chapter concluded by affirming that it is indeed possible to generate large corpus-wide collocation networks using methods that align with current findings from psycholinguistics. However, it also highlighted the importance of various factors such as the choice of Association Metric, possible filters, collocation window, directionality, Graph Theoretical Analytics, and the limitations to a single layer of linguistic representation. With these considerations in mind, the present Chapter aims to delve deeper into the practical aspects of the research. Fifteen corpus-based and two word-association based networks are generated and examined, comparing and contrasting their structures and properties, and evaluating the performance of different Association Measures. The goal is to identify which measures or combinations thereof lead to networks that most closely resemble word association networks (RQ2) and what shape the individual networks and their differences take (RQ3). Emerging key clusters are examined alongside words that display high linguistic availability, and potential candidates for kernel lexicon membership.

A number of key points relating to the corpus and word association component respectively are reiterated here to frame the methodology presented in this Chapter. Firstly, linguistic patterns such as collocations play a special role in the linguistic performance of an individual – their repeated usage leads to a performance increased ease of processing when compared to open constructions as evidenced in a number of psycholinguistic studies (Simpson-Vlach & Ellis, 2010, p. 489). One underlying theory of this exploration is the notion of a frequency or recency effect (Akmajian et al., 2010, pp. 433–434; Ferrer-i-Cancho & Solé, 2001, p. 2264; Forster & Chambers, 1973; Hitch et al., 2022) meaning that a high frequency of a word translates into easier availability for both linguistic comprehension and production. R. Brown and McNeill (1966), for instance, found that frequent words are more available for production, while Forster and Chambers (1973) found that frequent words are more available for comprehension. In network terms, a higher degree of a given word thus indicates an increase in ease of availability (Ferrer-i-Cancho & Solé, 2001, p. 2264). Continuing this line of thought, a certain ranking or interconnectedness of these high-frequency words or phrases would then also influence the information flow and, in turn, affect linguistic performance in terms of information search and retrieval. This falls in line with the observation that human cognitive processes show similarities to the PageRank (Borge-Holthoefer & Arenas,

2010, pp. 1290–1291), algorithm, an algorithm which, in simple terms, ranks entities according to how many important other entities link to it and perpetually readjusts these weightings. Large Linguistic Networks based on collocations are thus generated here to represent these processes.

Comparing and contrasting the emerging networks is also aimed at providing an empirical evaluation of different Association Measures used in corpus linguistics. This is necessary since there is an ever-growing list of possible statistics yet pointers towards which Association Measures can theoretically be used to infer mental associations are sparse; the first research question presented in this thesis aimed to provide these pointers, and a practical exploring of which Association Measures produce results that are more similar to experimental word association data is a core focus of the present Chapter.

Looking at the word association component, this rests on the assumption that communication between speakers is ensured when a shared or kernel lexicon exists (Ferrer-i-Cancho & Solé, 2001, p. 2261). Elements of this kernel lexicon would be expected to emerge from the word association data collected on the basis of a large number of participants. Particularly the more central and key nodes could be interpreted as candidates for kernel lexicon items. The result can then be used for comparisons with collocation networks to examine structural and qualitative similarities and differences.

Since the networks generated for this thesis provide a rich resource for investigating these phenomena, the larger aim of contrasting holistic collocation networks with word association networks has been split into three subsections pertaining to general questions, the empirical evaluation of different (combinations of) Association Measures, and the identification of lexical items displaying special functions both in the collocation network and in the word association network.

1. General Considerations
    - Are the networks intuitively interpretable?
2. Empirical evaluation of structural comparisons (Macro-level)
    - What are the properties (size, connectivity, etc.) of the resulting networks?
    - How large is the percentage overlap between unweighted nodes and edges in the different collocation networks and the word association network?
    - Which AMs or combinations thereof lead to the highest similarity to the word association networks?
3. Empirical evaluation of qualitative comparisons (Meso-level and Micro-level)
    - How similar are the emerging key clusters?

- Which words display a particularly high linguistic availability as indicated by their high degree in the respective networks?
- Which words are candidates for kernel lexicon membership?
- Do network-central hubs qualitatively differ between these networks in terms of the word class or high/low frequency of represented words (as expected on the basis of Veremyev et al. (2019, p. 3)?

It is crucial to discuss the general question concerned with intuitive interpretability in greater detail at this stage since it seemingly contradicts the overarching commitment to move away from individual introspective assessments as much as possible. This question ultimately aims to assess the quality of the observed results. While more immediately quantifiable measures such as EEG studies or reaction time studies would be desirable as an additional data source in the context of this project, financial and temporal constraints make it impossible to employ them as part of this thesis. A thorough psycholinguistic experiment testing the resulting network structures would require the recording of responses by a large number of participants using expensive specialist equipment. For now, this thesis therefore puts a strong focus on the directly measurable properties emerging from the networks themselves while structuring the data with future possibilities for cross-comparison with experimental studies in mind.

4.2 Methodology

This Chapter provides an in-depth exploration of both the methodology used to obtain large scale linguistic networks and the evaluation of the findings. The selection of datasets to represent word associations and collocations respectively are justified alongside the wider methodological approach. Providing in-depth information regarding the structure and inevitable limitations of the datasets is relevant since all further analyses and interpretations will heavily depend on participant/contributor statistics regarding sociolinguistic variables such as age, gender, L1, etc. An integral part of this thesis is furthermore the development of the computational routine that allows for generating large linguistic networks. This Chapter also contains a systematic overview and an accessible description of the subprocesses in the pipeline. The fully commented and adaptable code is available in Appendix A in the form of an interactive Jupyter lab. After a description of the methodology results obtained from contrasting holistic collocation networks with word association networks are explored and the abovementioned sub-questions are addressed and discussed.

The structure of this Chapter follows the full methodological approach from the selection of the dataset to the full implementation of the data processing pipeline for each of the two datasets. Everything pertaining to the data used in this thesis can be found in Chapters 4.2.1.1 / 4.2.1.2 and

4.2.2.1 / 4.2.2.2; an overview is provided in Table 11. The data processing procedures are described in Chapters 4.2.1.3 / 4.2.1.4 for the BNC 2014 and in Chapters 4.2.2.3 / 4.2.2.4 for SWOW-EN/SWOW-UK.

Table 11: Table summarising the properties of the datasets used in this thesis after processing.

Dataset	**Properties**
BNC 2014	99,772,275 lemmas in 88,171 texts, containing spoken language from 1,251 recordings of 668 unique speakers (Love et al., 2017, p. 320)
SWOW-EN	3,083,444 cue-response pairs from 67,355 unique participants
SWOW-UK	448,380 cue-response pairs from 9,700 unique participants

4.2.1 Collocation Networks

4.2.1.1 Rationale for Corpus Selection and Pre-Processing – The BNC 2014

The British National Corpus 2014, version 2 (BNC 2014, containing both the 90 million word written (Brezina et al., 2021) and the 10 million word spoken (Love et al., 2017) component) has been selected as the dataset for the exploration of corpus-based collocation networks for a variety of reasons. First and foremost, it presents a balanced and controlled sample of one variety of English – British English – and follows a meaningful sampling scheme. This is particularly relevant to this project since results from similarly-sized yet less meticulously collected corpora (i.e. large web corpora) are less suitable for collocation extraction tasks (Evert et al., 2017, p. 539). The BNC 2014 has been sampled to represent current written British English (Brezina et al., 2021) and contains a number of genres that are relevant to everyday life in the UK such as texts from news/magazines, fiction and TV scripts amongst others. This allows for using the BNC 2014 as a model for perceived language and thus creates a point of contact for psycholinguistic comparisons and evaluations that aim to generalise to the same population: Common users of contemporary British English. The BNC 2014 is furthermore of particular interest since it contains both spoken and written language and therefore allows for analyses of possible general differences in the structure of collocation networks on the basis of the mode in which language has been used. Another reason for using the BNC 2014 is its controlled and relatively recent sampling time from around 2010 to 2019 since this ensures that diachronic inferences are minimised. Figure 19 shows the exact figures of words per year of data collection. It becomes evident that the majority of the language in the corpus has been produced between 2014 and 2016 with all other years being represented in 5 million words or less.

The corpus design furthermore enables diachronic comparisons with the BNC 1994 in future studies since the BNC 2014 was sampled following the same overarching guidelines (the more recent version does, however, unlike the BNC 1994 also include an E-Language section comprised of Blogs, Forums, SMS, E-Mail and Review data which represents the change in domains of language use since the 1990s, for a more extensive discussion of BNC comparability see Brezina et al. (2021). Another benefit of using the BNC 2014 over most other datasets is the consistent tagging and availability of meta-data regarding each text. This allows for more detailed explorations of relevant patterns emerging from collocation networks.

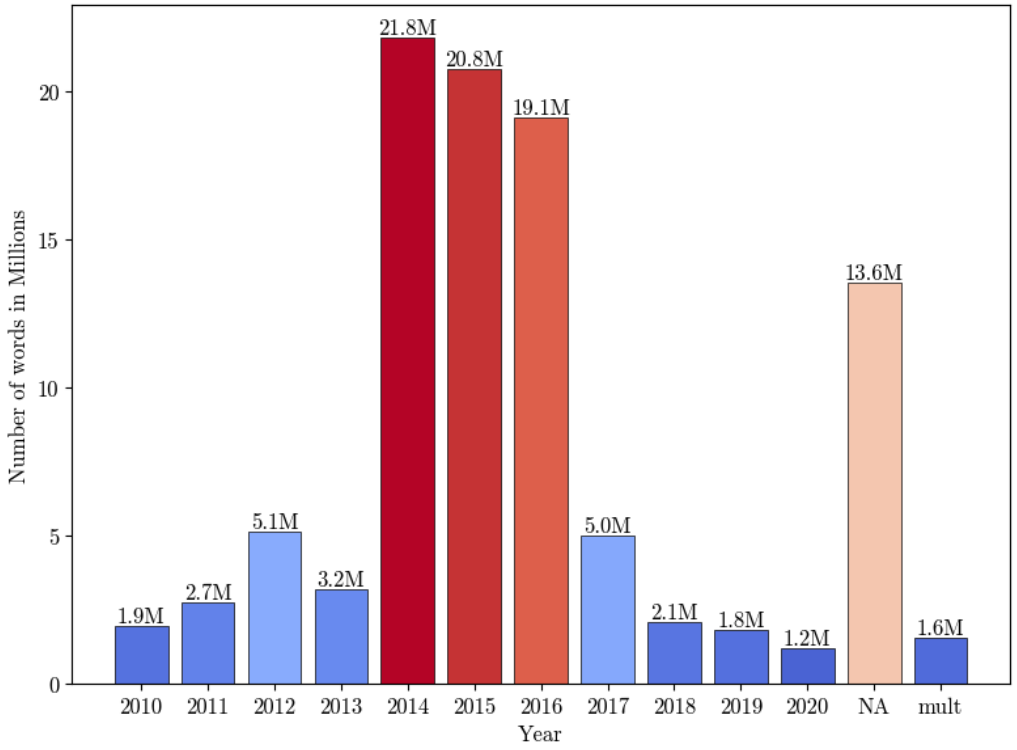


Figure 19: Words per year collected for the BNC 2014.

Figure 20 shows the genre distribution of the BNC 2014 as indicated by the embedded XML tags. The total number of words in the raw (not pre-processed) corpus is 99,949,544 in the following seven genres: Academic Writing (19,625,564 words), Electronic Language (5,171,398 words), Conversation/Spoken Language (10,317,212 words), Fiction (19,870,546 words), Magazines (14,979,505 words), Newspapers (19,996,923 words), Official Documents (6,996,882 words), and Written-To-Be-Spoken Documents (2,991,514 words). In order to understand what conclusions can and cannot be drawn on the basis of this data as well as for general transparency, the following paragraph provides a brief characterisation of all sections.

The Academic section spans scholarly writing across disciplines, including arts and humanities, medicine, natural science, politics, law, education, and technical fields; it includes both books and

academic articles. The Fiction section contains general prose, coming of age books, science fiction, fantasy, crime novels, romance and similar texts. The Newspaper section is sampled from a variety of British Newspapers and contains sports results, general news, arts and entertainment articles, as well as editorials and similar documents. Similarly, the Magazines section contains content focused on topics such as lifestyle and men's interests, TV and film, motoring, food, music, and science and technology. The Spoken section consists of participants' self-recorded, unprompted conversations; this section is completely unrestricted regarding the topic of the conversation. The E-Language section contains tweets, Facebook posts, blog entries, discussion forum interactions, emails, SMS messages, as well as online reviews. Among the smaller sections Official Documents contain reports, and meeting notes as well as Hansard transcripts, and the Written-To-Be-Spoken section contains a variety of television and drama subtitles. After pre-processing (described in greater detail in Chapter 4.2.1.1), the resulting dataset contains a total of 99,772,275 lemmas.

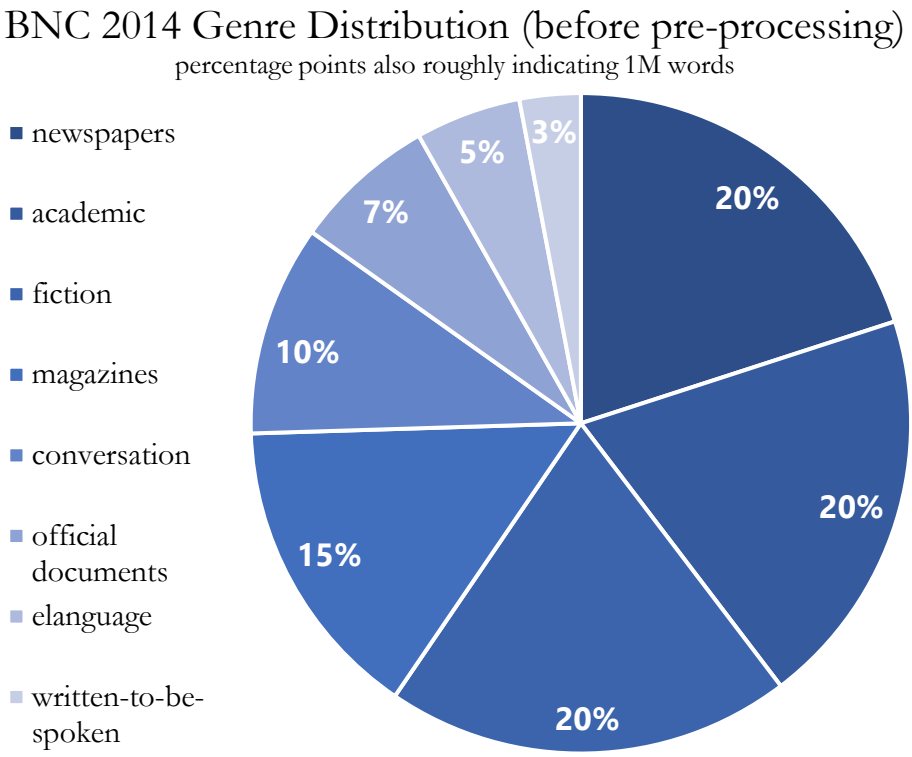


Figure 20: Genre distribution of the BNC 2014 (v2).

4.2.1.2 Dataset Evaluation

While an extensive justification as to why the BNC 2014 was used as a basis for identifying collocations has been provided, it is also important to consider the limitations of this dataset. This critical discussion of the dataset comprises two major elements: The sampling frame and its justifications as well as the quality of the collected data.

Firstly, while the BNC 2014 aims to present a balanced sample of everyday language and the sampling choices made here are systematic and well executed, some issues remain. Firstly, the majority of the corpus consists of written language while the lived reality of a non-insignificant proportion of the UK population relies primarily on spoken communication: Average literacy scores as identified by a large-scale study across multiple OECD countries indicate that the average English adult (16-65) is unable to read lengthy or dense texts and possesses an average literacy score of 273/500 (Department for Business, Innovation and Skills, 2013, p. 56). What is more, 16.4 % of adults in England score below 226/Level 2, which leads to a classification of over 7 million adults in England as functionally illiterate (ibid.). This is exacerbated by the fact that, unlike in many other countries assessed by this OECD study, the average literacy scores for England seem to be static rather than improving in England and Northern Ireland (Adult literacy Program for the International Assessment of Adult Competencies [PIAAC], 2012). Besides the low percentage of spoken language in the BNC 2014, other components of everyday language such as radio/podcast transcripts, video game material and transcribed content from social media video platforms such as YouTube, Instagram or TikTok are not represented despite the fact that they account for a sizeable part of the daily viewing times for people aged 4+ in the UK, over 40 minutes, as of 2020 (Ofcom, 2021). The minutes spent on social video platforms has not decreased after the end of UK lockdowns either and the Ofcom (2023a, p. 22) overview showing data from March 2023 highlights that language used in video material on Instagram, Snapchat, YouTube, and TikTok combined constitutes over 3 hours of exposure to the average 15-24 year old in the UK (Figure 21). It is to be expected that this trend will continue and thus desirable to include this type of data in future general-purpose datasets.

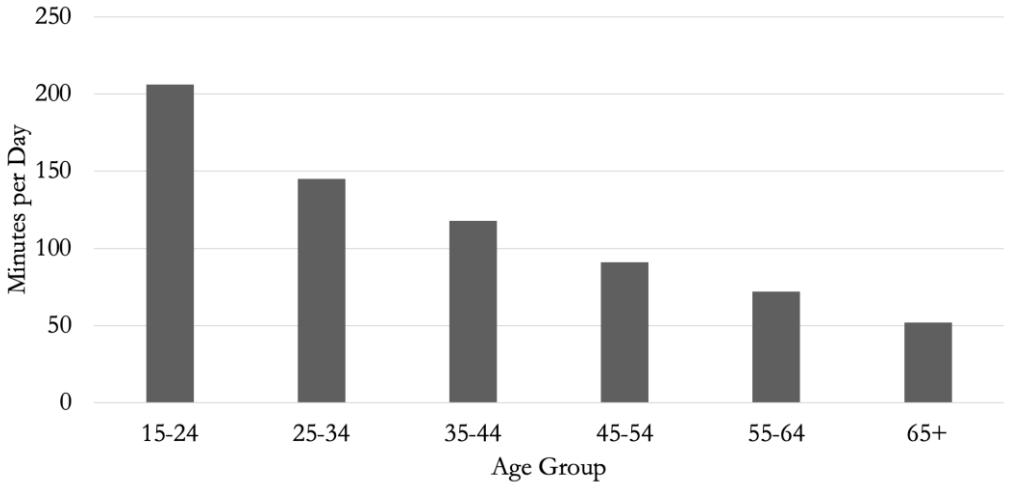


Figure 21: Average time spent per day on social video platforms (Facebook and Messenger, Instagram, Snapchat, TikTok, Twitch, and YouTube) in the UK in 2023 by different age groups (Ofcom, 2023a, p. 22).

A major difference between the BNC 1994 and the BNC 2014 is the introduction of an E-Language component. This is generally a very valuable development since this reflects the lived reality in the 21st century with 21% of people in the UK self-reporting an internet use of over 40 hours per week in 2023 (Ofcom, 2023b) better than a written section based primarily on books and print media. The E-Language section, however, also brings about specific challenges: Several types of metadata that can be reliably collected for traditional components such as author metainformation, date and place of publication are harder to obtain for online data. For blog posts and forum submissions, the anonymity of online users means that large sections will either not have a unique ID associated with them or merely a chosen pseudonym and declared age/gender with overall low credibility. This creates issues since it might result in the same author being classified as two or more different individuals on the basis of them using more than one pseudonym. Another consideration that is directly relevant to collocation extraction is the enriched context in online media that is not retained in the data collection process. A considerable part of websites and text message conversations consists of multimedia elements displayed alongside the text. These often set the scene and strongly influence the course a conversation takes; the lack of this information might distort contextual dependencies. First steps have been made towards retaining audiovisual information in corpora (Baker & Collins, 2023) which would allow the retention of this information in future releases of large-scale high-quality corpora such as the BNC 2014.

Lastly, coming back to the spoken section once again, a further limitation lies in the nature of the transcribed material. Due to the large size of this project, multiple individuals completed the transcriptions and while style guides and best practices were provided, there nevertheless remains a certain level of inconsistency in interpretation of the transcribed speech. The data naturally also does not contain sentence splits which complicates the collocation extraction process since the present methodology aims at extracting sentence-level co-occurrences. The decision to divide the spoken corpus into turns, intended as a substitute for sentence boundaries, was made because simple semi-automatic discourse element identification trials were unsuccessful, and because their development exceeds the scope of this thesis. Turns thus constitute the basis for the calculation of 'sentence-span' AM scores in the spoken section.

4.2.1.3 Pre-Processing and Tagging: General Considerations

This subsection explores universal questions that arise when working with corpora for network generation; the aim is twofold: Presenting a guide and reference points for future work on the one hand and motivating the particular choices made for this project and related linguistic network explorations on the other.

Firstly, it is important to carefully assess the quality of the available corpus in the context of the research question at hand both in terms of its appropriateness and its textual quality. A non-exhaustive list of important guiding questions contains the following points:

- Is the content of the corpus suited to answer the research question?
    - Is the size of the corpus sufficient for the study at hand?
    - Are there any over- or underrepresentation of individuals or groups?
    - Is the necessary metainformation (if any) available and accurate?
    - Are important registers and genres missing?
    - Is all of this made explicit in the methodology?
- Are the texts of adequate quality?
    - Are there duplicates?
    - Has any pre-processing been done on the dataset before it has been made available and is there sufficient documentation as to what these changes entailed?
    - Are there formatting or encoding issues, has the data been normalised?

After these initial questions have been answered, tagging needs to be considered. Important questions in this regard are the following:

- Are any types of tagging (Part-of-Speech (POS), semantic), normalisation and lemmatisation necessary or desirable in the context of the research question?
- Is there a pre-tagged version of the corpus?
- What is the accuracy of the taggers in use, are there known systematic issues that impact parts of the data more than others?

In this thesis the following decisions have been made for the BNC 2014: Since the BNC 2014 is a balanced and carefully sampled, large corpus as described above in greater detail, this is considered suitable for representing everyday British English. The available meta-information is largely available (though not evenly spread throughout genres, web-based texts widely lack this type of information) and accurate. No missing genres or registers come to mind with the exception of mixed-language scenarios (i.e. talking to non-native speakers, language mixing etc.), particular social media platforms, and child language. Duplicates have been removed from the original data of the BNC 2014, extensive POS-tagging, semantic tagging, and lemmatisation have already been carried out; no normalisation has been employed and no encoding issues have been found.

4.2.1.4 The Large Linguistic Network (LLN) Pipeline: Collocations

In this chapter, the script design for processing collocations as the underlying source of large linguistic networks is presented in the form of a skeletal overview and further guiding questions. This constitutes the heart of this thesis since it provides a major contribution towards automatic analyses and visualisations of large linguistic networks. This textual description of the interactive Python Code (henceforth referred to as LLN pipeline, see Appendix A) is provided in order to make the methodology transparent to readers with little or no coding background and to address a list of limitations that result from decisions that are necessary as part of developing this tool. The description offered here does not include any pre-processing steps that have been applied to the BNC 2014 specifically since it is aimed to be a general guide for potential future users working with different datasets. The full changes made to the raw BNC 2014 data can, however, be replicated in full via the LLN pipeline. Figure 23 provides a schematic overview of the LLN pipeline.

Aside from technical decisions such as choosing appropriate indexing strategies for storing metadata, POS-tags, and other information associated with the raw text in a given corpus the first step involves pre-processing the data and explicitly defining what will be considered possible atomic units composing collocates. It is important to draw special attention to the following points, this list – whilst certainly not exhaustive – is indicative of the most important initial considerations:

- How will hyphenated words be treated? Is "up-to-date" considered one unit and, if so, what implications does this have for "up to date"?
- Are abbreviations considered one unit or do they represent the constituent words and should be treated accordingly?
- Are numbers and special characters considered words (e.g. Will "46" be treated differently than "fourty-six" and "fourty six" (see next point)? Are both "&" and "and" valid nodes; Will they be conflated or treated as separate nodes?)
- Should variant spellings and spelling errors be normalised?
- Are co-occurrences that cross sentence boundaries valid?

In this case, numbers using digits are excluded from the dataset since they very frequently occur in lists which distorts the rank scoring of other collocations; instances of '&' and other special characters are also not counted. Hyphenated words are treated as one token, this decision mirrors the cue-association data in the SWOW dataset; the same goes for abbreviations. For the current project, no normalisation is applied to keep the data as authentic as possible, co-occurrences are not counted across sentence boundaries.

After these data- and project-dependent questions have been answered, the next steps consist of the tokenisation of individual nodes (defined as units delimited by whitespace that fit into the categories specified via the above questions). At this point, thresholds are also applied in order to make the dataset more manageable and to ensure a consistent and informative results. A minimum collocate frequency for each collocation in each subsection of the corpus has been determined in relation to the size of the subsection. The following chart contains the values for the BNC, these have been partially selected in order to minimise the rounding error since this could introduce imbalances between the validity of samples from different subsets. In practice, this means that since the academic section of the BNC 2014 is roughly 6.5 times the size of the written-to-be-spoken section, collocates need to appear roughly 6.5 times as often (i.e. 20 times) as the baseline for the smaller section to be considered approximately equally relevant. While these thresholds have been theoretically motivated, other thresholds are also plausible and might be explored in future work.

Table 12: Minimum frequency of occurrence for collocates per section of the BNC 2014. A negative rounding error means that the true threshold should be larger by the size of the rounding error; the inverse is true for positive rounding errors.

Section	Threshold	Rounding error
Newspapers	20	-0.05
Academic	20	0.32
Fiction	20	0.07
Magazines	15	-0.02
Conversation	10	-0.35
official documents	7	-0.02
electronic language	5	-0.19
written-to-be-spoken	3	0.00

The resulting ordered list containing all words in the corpus above the threshold is then used to create two counts: first, directional co-occurrence counts for the desired window size, here sentence-wide tuples (see Chapter 3.2.4 for a general discussion of this property), and directional co-occurrence counts for immediately neighbouring words (bigrams).

Since the generation of the raw collocation counts can be operationalised in a number of different ways and particularly since the creation of sentence-wide collocations is somewhat niche, the procedures used for this project are briefly explained here. For the sentence-wide tuples, the corpus has been slimmed to only contain words above the frequency threshold in each section, and sentence-split. Within each sentence, the occurring words are recombined while maintaining their

order and counted. Given the sentence 'It is late, isn't it?", the following combinations/counts are therefore obtained: (it, is):2, (it, late):1, (it, n't):1, (it, it):1, (is, late):1, (is, is):1, (is, it):2, (late, is):1, (late, n't):1, (late, it):1, (is, n't):2, (n't, it):1. Self-loops (it, it), (is, is) are counted, directionality is fully preserved, no weightings for particularly close occurrences are introduced and no corrections are made considering the different window lengths. The large number of combinations that result from this very brief example sentence already hints at the fact that the generation of these counts is computationally very expensive, especially for longer and lexically complex sentences. The number of tuples can be represented as the triangular number T_{n-1} with n being the number of words per sentence. This means that a sentence containing n words will lead to $\frac{n(n-1)}{2}$ combination tokens (note that there will likely be repetitions amongst the words in the sentence and therefore less combination types). Using the above example $6 * \frac{5}{2} = 15$ combinations are thus obtained, for a sentence consisting of 30 words this would result in 435 combinations, and for a 100-word sentence, 4950 combinations would be counted (see Figure 22).

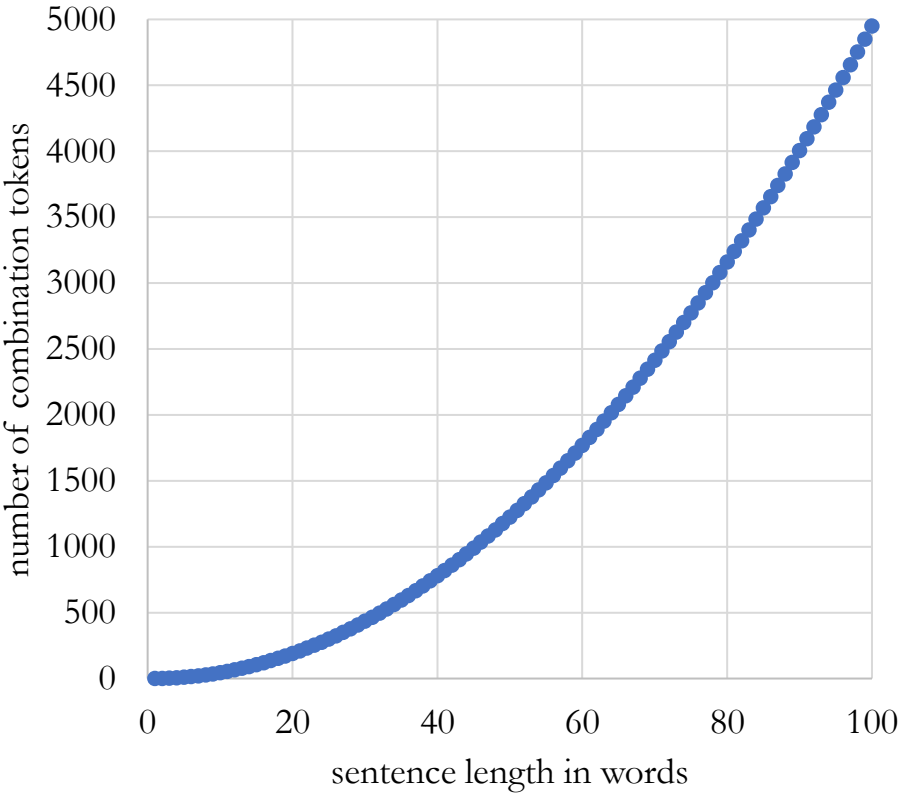


Figure 22: Number of combination tokens based on sentence length.

In order to filter out less reliable one-off combinations, the tuples generated as described above are required to appear in at least two different sentences, however there is no requirement for the sentences to occur in different files/sections.

Despite the fact that the sentence length varies considerably throughout the corpus, no window length correction has been applied. This decision has been made for three reasons, the first being that the correction itself introduces a different kind of arbitrariness since some words might have strong biases towards appearing in lengthy sentences whereas others, particularly high-frequency words such as pronouns and forms of auxiliary verbs, also commonly occur in extremely short sentences. Since it does not lie within the scope of this project to take all of these factors into consideration before every window size correction applying this would be counterproductive and lead to an alienation from the linguistic reality. The second reason for using the raw counts without a window correction is compatibility: A large number of existing corpus linguistic tools do not employ a window size correction for their calculations by default (e.g. AntConc (Anthony, 2022) and SketchEngine (Kilgarriff et al., 2014)), only LancsBox (Brezina et al., 2020) currently makes this option available to the users. Thirdly, as explained in Chapter 3.2.1, the lack of a uniform, transparent standard approach to collocation calculation in the field meant that all calculations and processing routines for this project have been written from scratch and all contingency tables do not sum to the number of words in the corpus, but instead to the total number of combinations of tokens identified in the corpus. This renders a window span correction superfluous.

In the next step, these counts are used as the basis for a selection of psycholinguistically plausible Association Measures (see Chapter 3.3); these are then calculated for every node-collocate combination. Once this process is finished, the last major methodological consideration concerns establishing a threshold. While it is theoretically possible to visualise and calculate scores for every co-occurring word pair regardless of actual strength of association, this complicates the graph theoretical analysis and might not be desirable depending on the specific research question at hand. AM options provided within the LLN pipeline are MI, MI^2, MI^3, MI^4, dice, log Dice, LL, oddsRatio, logOddsRatio, Jaccard, r_φ, χ^2, $\Delta P_{forward}$, $\Delta P_{backward}$, and ΔP. While this selection has been made on the basis that these AMs are reasonably frequently used in corpus linguistics research the script has been written in a way that allows the user to add their own calculations on the basis of the pre-calculated contingency table in order to allow for the incorporation of more specialised AMs should they be required for a project. The association calculation step concludes the traditional corpus linguistic processing of the data.

After this, the LLN pipeline proposes two systematic ways of establishing thresholds: Option one is selecting a set threshold, this is recommended when comparability with other research projects is desirable or the AM in question has a theoretically motivated threshold, e.g. a LL of 6.63 for $p < 0.01$ or a LL of 3.84 for $p < 0.05$. Option two, the option that has been used for network comparison in this thesis, is a quantile-based approach where the user can elect to only retain

collocations that lie above a certain percentile threshold, i.e. above the 9th percentile. The pipeline will then output the value of the selected AM that this percentile threshold corresponds to. A major benefit of this approach is that it helps normalise the effect the distributional structure of the AM has on the network. Beyond this, the selection of multiple AMs and thresholds is also supported in the LLN pipeline and encouraged for obtaining robust results and combatting some of the many limitations discussed in Chapter 3. The approach to remedy issues arising from the use of significance-based AMs by also simultaneously employing effect-size based AMs has been taken in seminal research or papers such a Bartsch (2004) who combine MI and z-score, Hamilton et al. (2007) who combine MI and t-score and Gabrielatos and Baker (2008) who combine MI and LL. The LLN pipeline thus allows for recursive selection and combination of the 15 AMs listed above.

The next and final component of the workflow involves the graph theoretical analysis of the obtained node-collocate pairs and their associated AM values. For the purpose of this the choice of a suitable network generation tool becomes pivotal. One such tool, Cytoscape (Shannon et al., 2003); a free and open-source network analysis software commonly used primarily in biology and chemistry, can effectively handle the loading, visualisation, and processing of networks, including those generated by the LLN pipeline. While Cytoscape offers a user-friendly interface and integrated graph theory tools, its limitations include performance speed, the need for visualisation throughout the filtering stages, and challenges in efficient chaining of commands, e.g. to filter out certain part of speech tags, words that are part of certain constructions or contain special characters.

An alternative option, an equally free and open-source Python-based network pipeline written for this thesis and provided in the Appendix alongside the AM generation script, provides flexibility for customisation, and offers benefits such as faster processing, increased algorithmic control, and a pre-written export function to use for presenting networks as dynamic web apps on websites. Another key feature of this approach is full control over when visualisations are being generated. Cytoscape defaults to a visual representation of networks whereas the LLN pipeline allows for complex filtering or cluster identification before running costly visualisations. The two core limitations of this approach are the command-based user interface when compared to the more user-friendly GUI of Cytoscape, and, despite the overall more efficient approach, a hard upper limit for graph theoretical analyses of extremely large networks at around 10 million edges.

The decision between Cytoscape or other pre-existing software tools and the LLN pipeline should be made with careful consideration of trade-offs related to user-friendliness, adaptability, and computational efficiency. For the purposes of comparing the word association and corpus-based

networks in this thesis the LLN pipeline is employed for the computation of all graph-theoretical parameters due to the large network size and available features.

An overview graph theoretical parameters of particular interest to linguistic research is provided in Chapter 2.7.1. Due to the availability of a large number of computationally costly graph theoretical parameters it is recommended to pre-define theoretically motivated relevant parameters for the exploration at hand rather than carrying out full calculations. This significantly increases the computational efficiency, particularly when dealing with large datasets such as the BNC 2014. Technological advances will likely render this step unnecessary in the future.

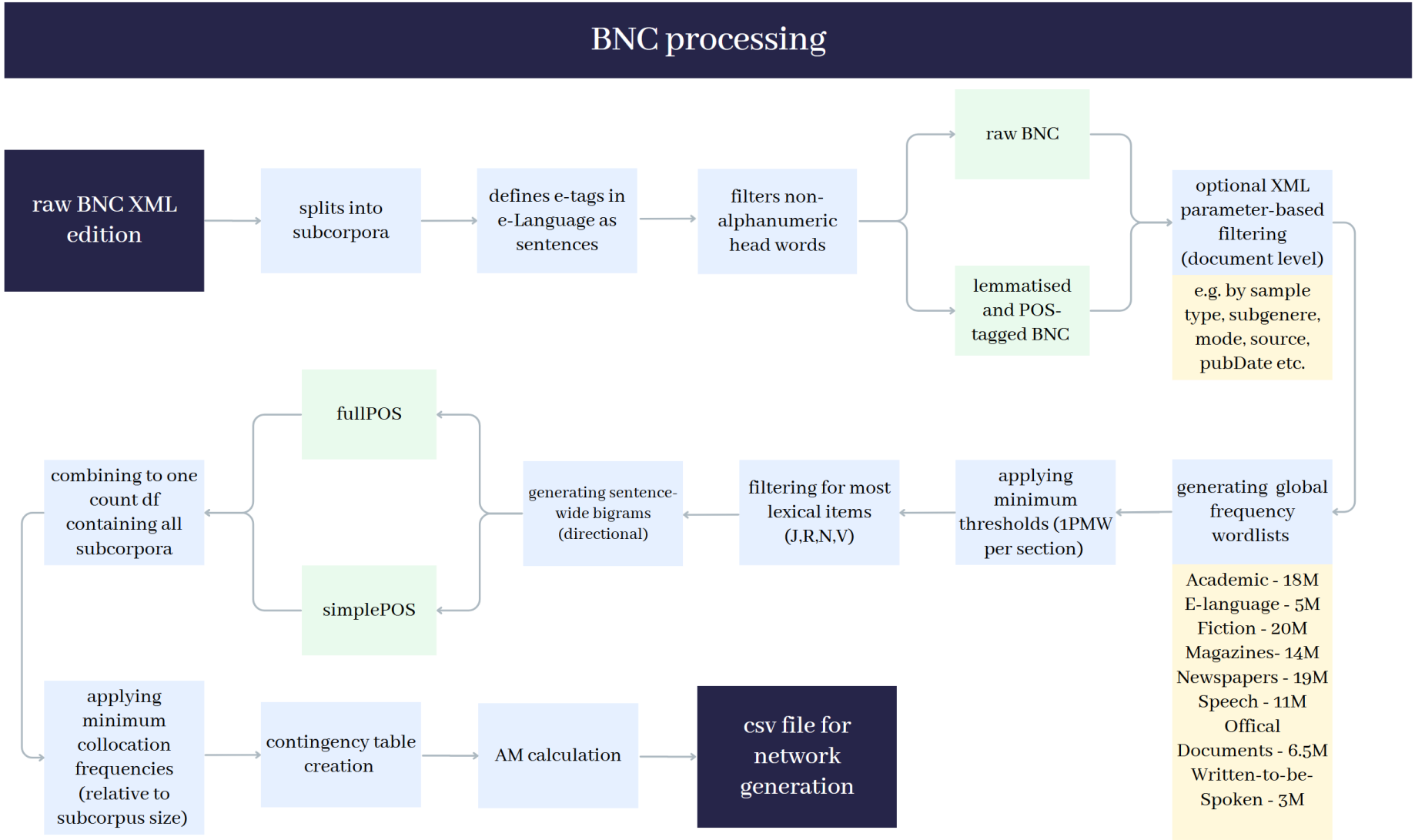


Figure 23: Schematic overview of the BNC 2014 processing component in the LLN pipeline.

4.2.2 Word Association Network

4.2.2.1 Rationale for Word Association Database Selection and Pre-Processing – SWOW

Word association tasks have been instrumental in exploring the structure and content of the mental lexicon and patterns of thought for several centuries; see Locke (1700) who discusses that there are ideas where

> *"[…], the one no sooner at any time comes into the Understanding but its Associate appears with it; and if they are more than two which are thus united, the whole gang always inseparable shew themselves together."*

or Galton (1879, p. 154)[14] who published a study examining and categorising his own associations.

While the approaches mentioned above are largely conceptually oriented and interpret associations as signs of general behaviours or, in Locke's case, as failures in reasoning, the applications of word association studies are much broader and more varied nowadays. Word associations are commonly employed to investigate both the psycholinguistic underpinnings of word storage and retrieval generally (Deyne & Storms, 2015; Simmons et al., 2008, p. 106) and knowledge of vocabulary items or 'stereotypy' in language learning research (Fitzpatrick & Izura, 2011; Meara, 2009). While this thesis does not allow for a full exploration of past research into word associations Fitzpatrick and Thwaites (2020) provide a more comprehensive critical overview.

This Chapter sets out the foundations of the word association network component of this thesis and largely mirrors the structure of the above Chapter on the selection of the BNC 2014, its metadata, and the processing pipeline used to generate corpus-wide collocation networks. Initially, this Chapter provides an overview of psycholinguistic databases that are already available and function as potential candidates to be the basis of word association networks which could model the mental lexicon. The databases that will be presented in greater detail here are WordNet (Fellbaum, 2006), the Semantic Atlas (SA) (Ploux et al., 2010, p. 356), University of South Florida Free Association Norms (USF Norms) (Nelson et al., 2004) and the English component of the Small World Of Words (SWOW-EN) (Deyne et al., 2019, pp. 998–999).

The databases can be classified into two types which roughly correspond to the opposite ends of the methodological spectrum ranging from rationalist linguistic introspection to empirical investigation (Grieve, 2021; Krug et al., 2013). Partially introspective datasets that were collected with a focus on psycholinguistic and cognitive theory but ultimately heavily rely on researchers' intuitions for classification such as WordNet (Fellbaum, 2006, p. 668), and datasets that were collected through experiments with a large number of participant-generated nodes such as the USF

[14] This work was undertaken before Galton applied himself to 'studying' eugenics.

Norms and SWOW-EN on the other. Both types of datasets require discussion in order to depict the diverse approaches employed in researching the mental lexicon and motivate the selection of one over the other for this project. The introspective approach involves reflected, meaningful, and labelled connections curated by lexicographers. In contrast, the experimental approach results in spontaneous and unlabelled connections established by a significant number of non-expert participants.

The first resource, WordNet, is a semantic database of English nouns, verbs, adjectives, and adverbs. Since these sections are hardly interconnected, one could view WordNet as consisting of four distinct subnetworks. The nodes, i.e. specific meanings of lexical items, are connected through edges that signify lexical and conceptual relationships and thus form un-ordered groups of connected words, so-called "synsets". WordNet currently spans 117,000 of these structures. It is mainly based on hyponymy but meronymy, antonymy and synonymy are also encoded. The data for the project was originally sourced in 1985 using the Brown Corpus and edges were created using existing knowledge from a variety of different thesauruses and a team of lexicographers.

The second notable word association database, the Semantic Atlas was designed to display synonym groupings, so called cliques, that are identified through correspondence factor analysis. Similarly to WordNet, the SA also operates on the basis of distinct word meanings rather than words as a broader concept. The dataset assembled for the SA, like WordNet, draws from various dictionaries and thesauri. However, the key distinction lies in the organisation of the meanings it contains. Unlike WordNet, which relies heavily on the hierarchical structure and researchers' judgments, the SA adopts a geometrical structure. This structure is primarily based on vectors derived directly from the data itself, rather than individual interpretations (Ploux et al., 2010, p. 356). While this positions the SA closer to more empirical approaches, it is still ultimately derived from the judgements made by individual lexicographers compiling the source thesauri and dictionaries.

The third dataset, the University of South Florida free association, rhyme, and word fragment norms, belongs to the empirical type of psycholinguistic databases: They were collected via free association tasks carried out by more than 6,000 participants. Construction of these norms started in the US in 1973; participants were prompted to note down the first **meaningfully related** or strongly associated word that came to mind when presented with a list of about 100 to 120 words that had been randomly chosen from a total pool of about 5,000 cue words. The participants were furthermore instructed to fill one (and only one) blank next to the cue word in a paper booklet which makes this approach inherently directional. As described in greater detail in Chapter 3.2.4, due to the inherent directional nature of certain word associations an overrepresentation of right-

predictive pairs is to be expected. There is, however, a considerable overlap of cue words and words that have been used as responses by participants which means that observations regarding the symmetry of right-predictive relationships are possible. While this dataset is biased towards representing nouns, adjectives, and verbs it also includes other parts of speech. The cue words were chosen on the basis of previous experiments, individual research projects (i.e. rhyme words), and specific interests of the researchers working on the project (Nelson et al., 2004, p. 403). This presents a major methodological issue when using this dataset as the basis for word association networks since the lack of a systematic approach to cue word selection will distort the overall structure of the resulting network.

SWOW-EN represents a similar but more methodologically refined, modern, and larger-scale association database. The project is ongoing, but the data used in this thesis has been collected from 2011 to 2018. This timespan matches the data collection time of the BNC 2014 thus furthering the ease of comparability. The purpose of the SWOW-EN project is embarking on a higher-resolution exploration of the mental lexicon through a stimulus-response based semantic network and the version used in this thesis comprises a total of about 12,000 cue words that have been presented to over 90,000 participants (Deyne et al., 2019, p. 987). The list of cue words was determined using snowball sampling and lists from other stimulus-response datasets (such as the USF norms) to include both high frequency and low frequency cues. For each of these cue words, about 100 association judgements have been collected in the form of a primary response (R1; the first response that has been given) and two further responses (which combined make up R123). When exploring the data, the authors found that the R123 dataset does not only include more responses, it also includes more diverse responses as indicated by a greater type variety (Deyne et al., 2019, p. 991). The higher resolution picture of the mental lexicon that can be captured through these different types of stimulus-response groups is an essential and unique feature of the dataset. In combination with the recency of the data collection and the exceptionally large size of the collected data, this makes SWOW-EN a perfectly suitable underlying database for the generation of the psycholinguistic network in this thesis.

On a more general level datasets of the second type are more suitable as source material for direct psycholinguistic comparisons. The psychological reality of empirical datasets such as SWOW-EN is directly quantifiable through memory performance measurements such as reaction times and recall (Nelson et al., 2000; Nelson et al., 2004, p. 403), and a dataset less influenced by the proxy layer introduced via a researcher's intuition is preferable. There are, however, also limitations. Since the language learning modes of the large number of participants that contributed to the SWOW cue-response dataset are completely opaque except for the general reading and writing abilities

necessary to respond to the cues, all observations made in this thesis can only be considered results of combined long-term effects of different types of Statistical Learning, memorisation, and language production. No claims pertaining to separate visual and auditory learning mechanisms are therefore made on the basis of this data.

Lastly, when considering possible datasets for the comparison of corpus-based graphs with word-association-based graphs it would also have been an option to collect an entire large-scale dataset for this particular project from scratch. The decision to rely on SWOW-EN instead has been made for a variety of reasons: Firstly, the main focus of this work is not on the experimental design and data collection for the psycholinguistic component only but has a larger scope of graph-theoretically analysis and visualisation of different networks. Secondly, the excellent quality - and quantity far beyond what the resources for this project's resources could have permitted - eliminates the necessity for separate data collection. Lastly, the data collection phase of the existing database perfectly overlaps with the data collection span for the BNC 2014 which means that unwanted inferences, i.e. by the large-scale global discourse around COVID-19 that took place in parallel with the work on this thesis, are avoided.

4.2.2.2 Dataset Evaluation

After having motivated the selection of SWOW-EN (2018 version, Deyne et al. (2019)) as the underlying dataset this section explores the metadata associated with two subsets of SWOW, SWOW-EN and SWOW-UK. Chapter 4.2.2.3 explores the full processing pipeline used to generate the subsets in greater detail.

The UK-only dataset contains 448,380 cue-response pairs[15], while the dataset comprised of responses from all native English[16] speakers contains 3,083,444 cue-response pairs. While participants were prompted to provide a response set of up to three responses per cue, a cue-response pair is here defined as each of these possible three responses for compatibility with single-response experiments. The metadata displayed in Figure 24 and Figure 25 describe the number of responses per year, region, gender, education, self-reported first language, and age range. This data collectively provides a comprehensive overview of the demographic distribution and response types in the word association database in order to assess the diversity and representativeness of the data. The motivation for calculating and discussing the metadata at length is that this kind of data has often been neglected in word association research leading to a mismatch between participants and the norming group (Fitzpatrick et al., 2015, p. 45) and thus considerable limitations when it

[15] When considering each individual response word as one cue-response pair. Cue-response pairs in the metadata Figures below count multi-word responses as a single response.
[16] Please note that Irish is included as well, this is further justified in this Chapter.

comes to generalisability. The data collection phase spans roughly 8 years from 2011 to 2017. 2012 and 2014 respectively are by far the strongest years for response collection in the UK subset with over 140,000 cue-association pairs collected in either year, all other years lie below 50,000. This corresponds very well to the BNC 2014 data collection time span which also exhibits a peak around 2014 (see Figure 19).

The regional graph further shows that the vast majority of responses have been collected from people in the UK at the time of participation in the experiment. Less than one percent of responses have been collected from participants in a (former) European Crown Dependency such as Jersey, Malta, or the Isle of Man and about 4.5% of responses were collected from participants in Ireland. The decision to retain Irish responses was made since the UK dataset is overall sparse, and isoglosses do not generally follow the political border (Kallen, 2000). The gender distribution shows that the data is skewed towards participants identifying as female with just over 60% of word associations contributed by them.

The majority of responses do not have an educational background associated with them which is partially due to the fact that this metric has only been collected from 2013 onwards (Deyne et al., 2019, p. 988), but the available data shows a favour towards categories 3 (*High School*), 4 (*College or University Bachelor*), and 5 (*College or University Master*). This causes reason to suspect the dataset may be skewed towards associations from people with a higher educational background than the average population. While this generally needs acknowledging as a limitation it also mirrors the metadata found in the BNC 2014 thus enabling a more immediate comparison between the datasets in this thesis.

First language reported by respondents largely mirrors the regional information described above, about 5.6% of participants have declared their native language to be Irish; it is important to note here that all responses were given in English. Looking at the age distribution a significant skew towards responses from younger people can be observed in both SWOW-UK and SWOW-EN. Looking at the target demographic as delineated by the UK population data from 2018 (Office for National Statistics, 2018), see Figure 25, cutting associations made by participants under the age of 36 might seem desirable. Since the primary goal of this thesis is a comparison between collocation networks and word association networks it is preferable to strike a balance between matching the SWOW dataset with the BNC 2014 as well as the general population both datasets aim to represent. A cursory glance at the E-Language and Spoken sections of the BNC 2014 reveals a heavy skew towards the 26-30 age range. No other sub-corpora could be considered since age information regarding authors in any written section except for parts of E-Language is not available, and it can be speculated that authors in sections such as newspapers and official documents are generally over

the age of 30, balancing out the extreme picture emerging from the BNC 2014, E-Language and Spoken sections. While the SWOW dataset generally displays a less pronounced age discrepancy between participants and the target population than is typically observed in similar studies, it is important to note that any incongruity between the sociodemographic characteristics of the target population and the respondents in cue-association experiments poses a methodological challenge that warrants further investigation (Fitzpatrick et al., 2015, p. 45), and this is a major limitation of this study.

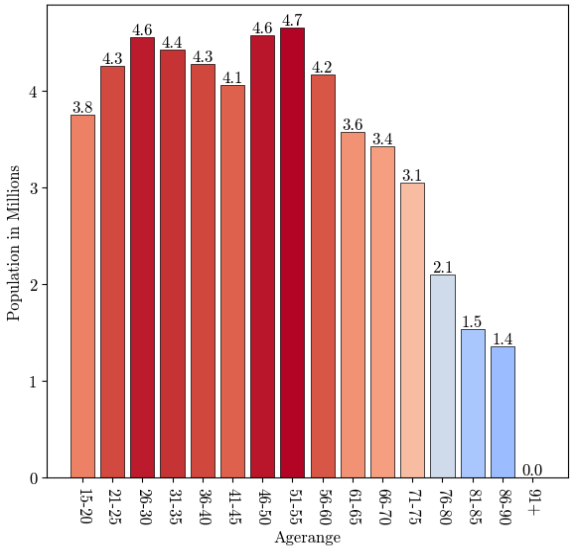


Figure 25: UK population in mid-2018 (Office for National Statistics, 2018) by age displayed in the same age bins as the SWOW datasets for comparison with Figure 24.

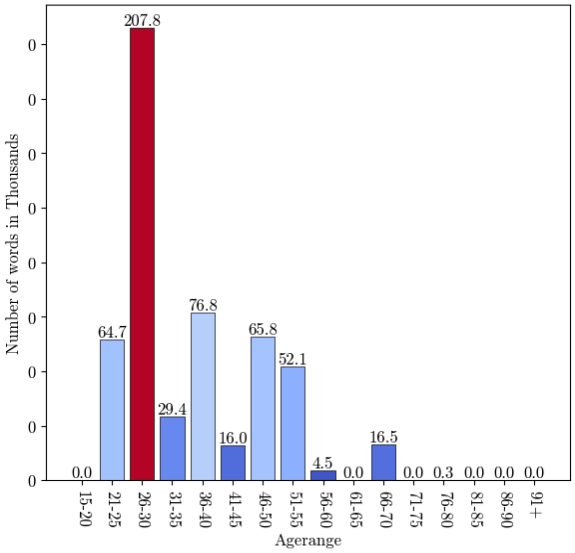


Figure 24: Number of words in Thousands by age in the BNC 2014 E-Language section. A similar picture emerges in the Spoken section (CASS, 2018, pp. 20–21), no age metadata is available for the other subcorpora.

The much larger native speaker dataset (see Figure 27) generally shows a very similar picture in terms of the gender distribution with the same bias towards female respondents alongside a very similar educational background skewed towards participants having completed a High School equivalent at least. Notable differences between the datasets are the sampling time with the native speaker dataset peaking more strongly in 2012, the age range where the UK dataset underrepresentation in the age bracket of 36-50 is slightly less severe, and the regional and L1 backgrounds. The majority (63%) of responses in this dataset have been given by respondents who reported their native language as US-American English, followed by much smaller shares for the British English (15%), Canadian English (13%), and Australian English (6%). Contributions from

native speakers of New Zealand English and Irish English respectively make up less than 2% of the total responses. As is to be expected, a similar picture emerges when exploring the regional distribution. 67% of responses were given from participants currently in the US, 15% from the UK, 9% from Canada, 7% from Australia, and less that 2% from South Africa or a European Crown Dependency.

In conclusion, the exploration of the SWOW metadata, specifically the SWOW-EN and SWOW-UK subsets created for this thesis, reveals a compromise between actual population representation and comparability to the BNC 2014. The limitations of the metadata such as a skew towards female participants, highly educated participants, and younger participants need to be made transparent throughout the project since they impact generalisability (Fitzpatrick et al., 2015, p. 45). Future word-association experiments and corpus compilations aiming for generalisability must address these methodological challenges. However, in the current comparative network study, the use of these datasets signifies a balance between accurately representing the population and ensuring comparability between datasets in the absence of more suitable alternatives.

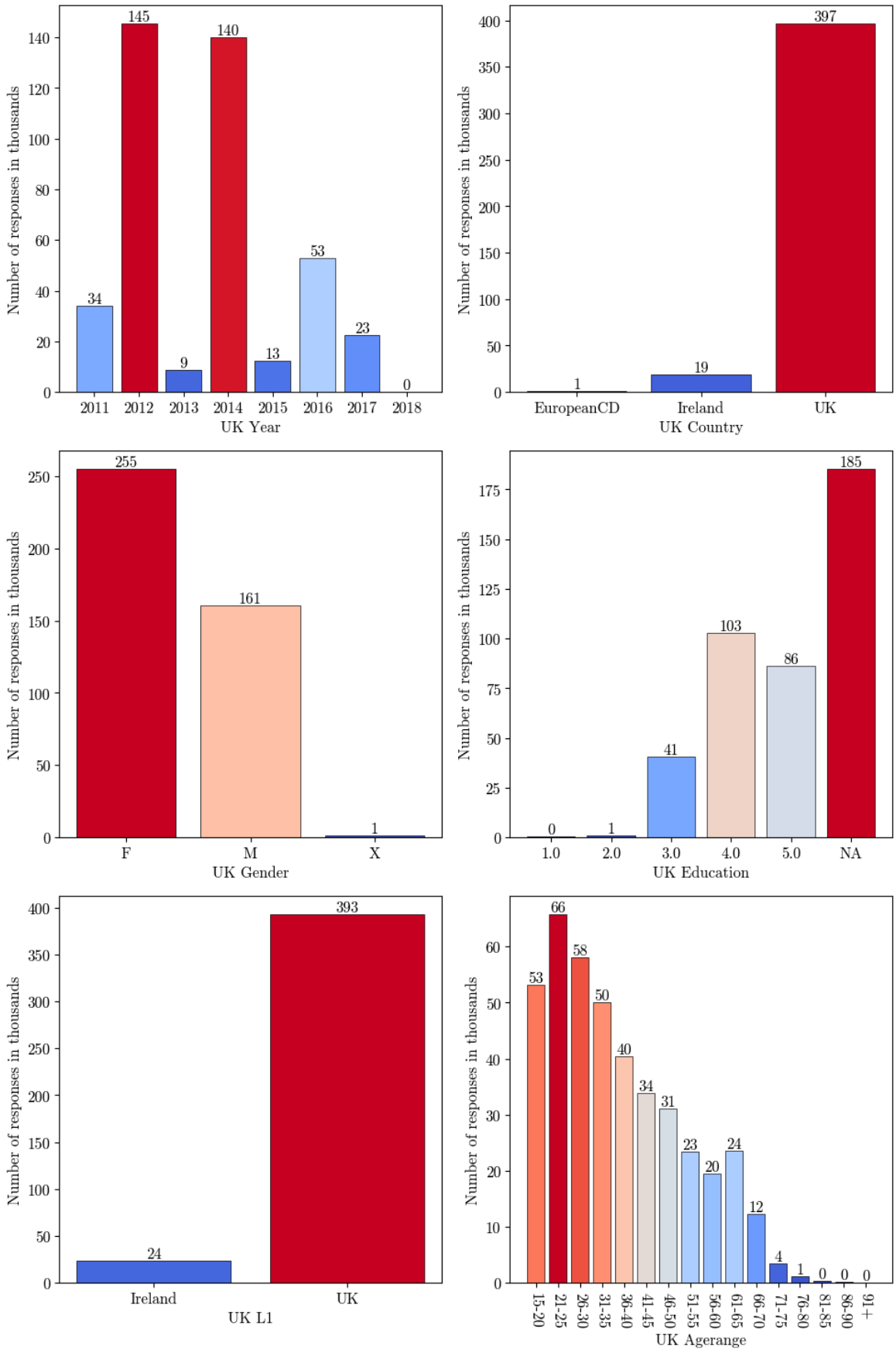


Figure 26: Metadata description of socio-economic metadata available for SWOW-UK.

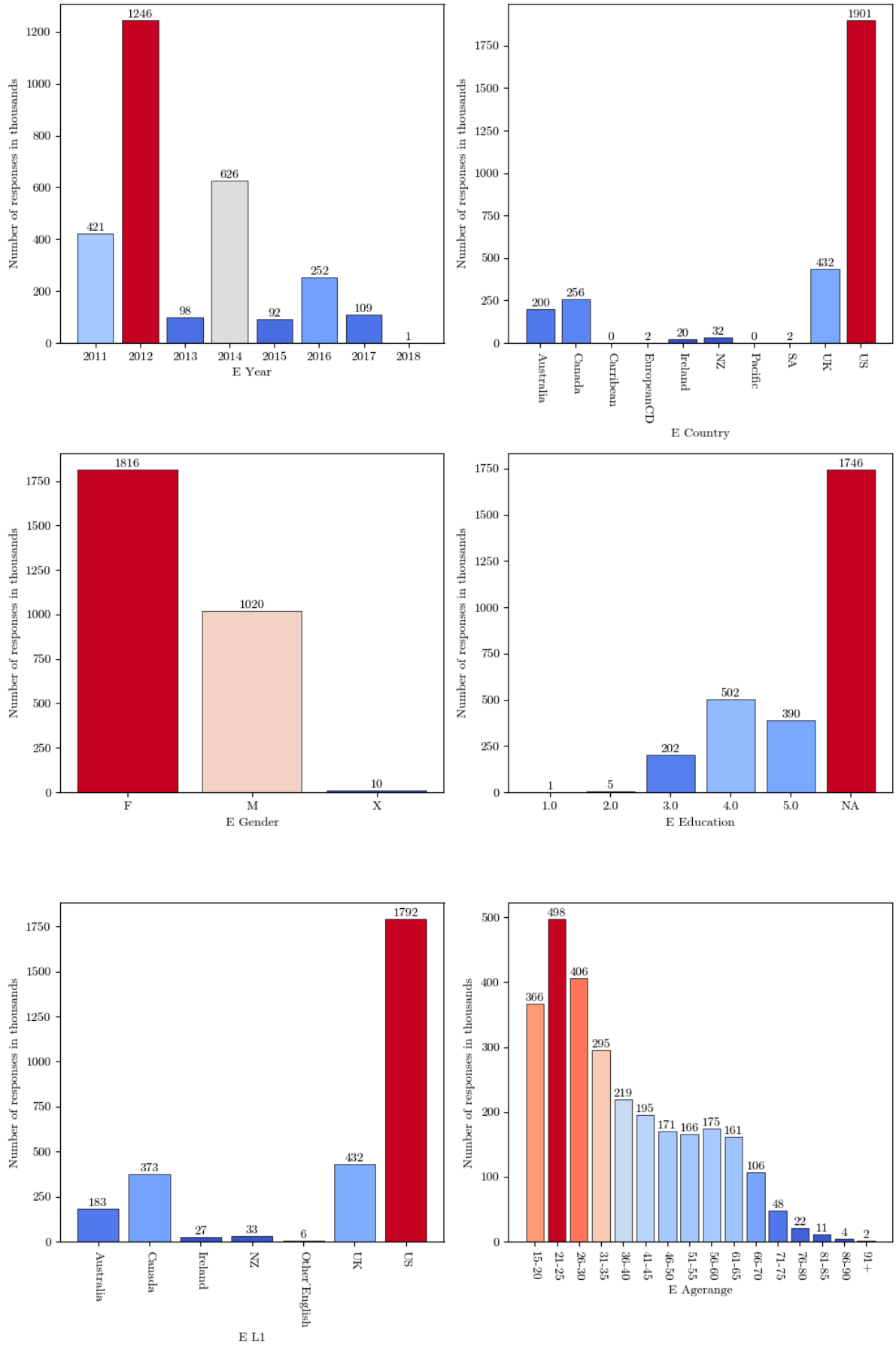


Figure 27: Metadata description of socio-economic metadata available for SWOW-EN.

4.2.2.3 Pre-Processing and Tagging: General Considerations

This subsection loosely mirrors Chapter 4.2.1.3 in that it also provides an overview and conceptual exploration of steps required to extract word association data in a shape that can be used as the basis for a weighted directed network. It, again, aims to present both a resource for future work on word association networks and to motivate methodological choices made as part of this project specifically.

The following list of questions, while certainly non-exhaustive considering the varied nature word association studies can take, is designed to be indicative of the most important initial considerations when working with word association data for network generation:

- What metadata is available (and is it sufficient)?
- Should a subset of the data be selected to better fit the desired target population?
- Did individual participants contribute to a significant percentage of responses?
- What level of normalisation is desirable for cues/responses?
- How many responses have been given per cue?
- Should there be a weighting and if so, how does this impact multi-word and non-primary responses?
- Do the weights require further adjustments e.g. on the basis of an overrepresentation or underrepresentation of certain cues as part of the experiment?
- Should a subset of the data be selected on the basis of the suitability of the cues to the research question at hand?

The data processed in the present study is the 2018 version of the SWOW-EN dataset, a full documentation of the individual transformations is available via the Appendix A in the format of an interactive Jupyter Notebook. The following Chapter addresses the principled decisions described in the overview above taken as part of this particular project and provides insights into the data processing that do not require computational knowledge.

4.2.2.4 The Large Linguistic Network (LLN) Pipeline: Word Associations

Firstly, the decision was made not to use the already pre-processed data by Deyne et al. (2019) since the edited version, while generally useful in a variety of contexts, would hinder a direct comparison to the BNC 2014 in a number of ways. This is the case since large-scale spell checking, normalisation, Americanisation and the removal of surplus responses if cues already have been responded to a certain number of times has been carried out as part of the data cleaning. Since the BNC 2014 is designed to represent British English specifically, which clashes with American

spellings, and has not been cleaned using the same normalisation and regularisation principles, the decision was made to retain the raw responses in the SWOW dataset to mirror this.

A first exploration of the dataset found that significant normalisation steps regarding e.g. participant-reported country names are necessary due to significant spelling differences and partial use of non-Latin alphabets. In a second step, the complexity of the data was reduced by filtering out information that goes beyond the level of detail present in the majority of the BNC 2014 such as exact times of day for particular responses, as well as city information, retaining just the raw and unedited cues, speaker IDs for latter metadata analyses, and primary, secondary and tertiary responses. The data was subsequently split into a metadata set and a cue information set to minimise file sizes and reduce computational demand for the network generation. For a broad overview of metadata categories and statistics see Chapter 4.2.1.2.

After normalising and grouping the metadata two subsets are created from SWOW-EN: SWOW-EN and SWOW-UK. The SWOW-EN dataset is a copy of the full SWOW-EN data that has been filtered for self-reported native speaker status as documented in Chapter 4.2.2.2 and participants currently living in an English-speaking country. SWOW-UK is a subset of SWOW-EN and has been filtered further to only contain data from participants with an L1 originating in the UK or Ireland and living in the British Isles or (former) Crown Dependencies. These steps have been undertaken in order to ensure that only the target population, in this case native English speakers/British English speakers, is represented in the data. An exploration of the share of responses per participant showed that all participants were presented with between 14 and 18 cues, which means that no individual has skewed the dataset significantly.

Moving on from processing the data on a meta-level it is essential to explore the shape of the individual data points. The data is subsequently split into a metadata set and a cue information set to minimise file sizes and reduce computational demand for the network generation. Example 1 and Example 2 illustrate the shape the cue dataset takes.

	cue[17]	**R1**	**R2**	**R3**
Example 1	afternoon	nap	pm	delight
Example 2	a	indefinite article	first letter of alphabet	No more responses

[17] The scheme of presenting the cue first followed by R1-3 will be used consistently from this point onwards when examples from the dataset are presented. Emphases always added by the author.

In the present study, cue-response pairs where participants indicated unfamiliarity with the cue word were filtered out. Additionally, partial responses – those missing a tertiary or a secondary and tertiary response (as exemplified by Example 2) – were retained. This filtering decision was made to keep the dataset as large and therefore representative as possible. Subsequently, all responses were transformed to lowercase and lemmatised for compatibility with the BNC 2014 dataset. This preprocessing step is particularly relevant when constructing networks for comparison with other linguistic data, as a shared node unit is essential. Other alternatives are preserving raw responses, stemming them, or assigning them POS tags. Since this thesis aims to contrast holistic word association and collocation networks, POS information could not be included. While the use of lemmatised words alongside their POS as nodes would be valuable, reliably determining POS category membership for isolated word associations is impossible[18]. For instance, in Example 1, the term "nap" could refer to either a verb, *nap_V*, or a noun, *nap_N*, and the lack of further context results in the absence of a ground truth. Given these inherent ambiguities, POS information has thus been disregarded to ensure comparability across datasets.

Example 3	weaken	lose strength	fading	**not as** healthy

Another property of the data requires special consideration: Manual exploration of samples of the data shows that negations are very common in multi-word responses (see Example 3). This is expected given the nature of cue-response tasks since it is common to associate and define a word via its antonyms, both on a general level since "binary opposition is one of the most important principles governing the structure of languages" (Lyons, 1977, p. 271) and in a word association context (Fitzpatrick et al., 2015, p. 40). Including negations in the network is undoubtedly a meaningful addition, due to the weighting conventions explained earlier in this chapter negations and structures expressing comparisons such as "not", "no", "unlike", "as", "more", etc. are, however, expected to exhibit an exceptionally large cumulative weight. This will be accounted for in Chapter 4.3.4.

Lastly, in some cases participants responded with a multi-word expression or brief description rather than individual words (see Example 2). Since snowball sampling was used to expand the set of cues, this issue also affected a number of cues given to participants. Cue-response pairs containing multiple words are meaningful and valuable but present a challenge since they cannot be preserved as one node unit in the network without violating the convention that every word represents a node. This problem has therefore been solved by introducing proportional weights by

[18] For other projects, researchers might want to consider a Unigram tagger such as the one available via NLTK Bird et al. (2009). This is not recommended for comparison with regularly POS-tagged corpora since Unigram tagging is entirely based on tag probability of a training corpus and thus blanket assigns the most frequent POS tag to *all* word forms.

number of words presented as cues and responses, described in greater detail towards the end of this Chapter.

The existence of (partially) incomplete responses means that the number of responses per cue is not fixed; this motivates using cue-response pairs (of which there can be a maximum of three per given cue) over response sets as the basis for the metadata visualisation. Distributional analyses have been carried out both on the full SWOW-EN and SWOW-UK to ensure there is a sufficient minimal and average number of responses per cue with a fairly even distribution.

The results have been found to be satisfactory with a mean of 82 responses per cue (σ = 9.0, min = 36) for SWOW-EN, but problematic for SWOW-UK with a mean of 12 responses per cue (σ = 5.1, min = 1), see Figure 28. Since the comparative nature of the present study requires a British English dataset SWOW-UK is nevertheless used as the basis for the comparative networks, but a lower weight threshold is employed. In future studies it might also be worthwhile to investigate a model based on the larger SWOW-EN dataset where cue-response pairs from the UK subset are assigned double weights in order to mitigate distortions.

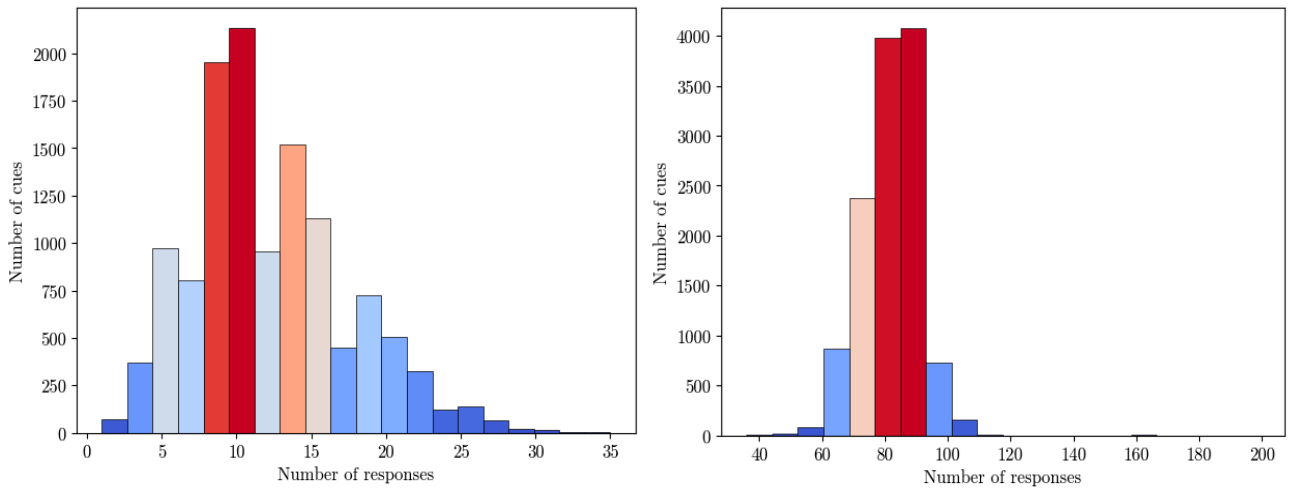


Figure 28: Histograms showing number of cues by number of responses for SWOW-UK (left) and SWOW-EN (right).

The following weighting system has been introduced in order to correctly represent the differences in importance of responses and to combat the overrepresentation of secondary and tertiary responses that are a result of priming through earlier responses: All primary responses are assigned a total weight of 1, all secondary responses are assigned a total weight of 0.6 and all tertiary responses are assigned a total weight of 0.4. Within this weighting system, further splits are made for multi-word responses. In Example 2, the total weight assigned to all responses given would be 1.6; the full distribution would look as follows: a-indefinite: 0.5; a-article 0.5 (which leads to a combined value of 1 for the primary response), a-first 0.15; a-letter 0.15; a-of 0.15; a-alphabet 0.15 (which leads to a combined value of 0.6 for the secondary response). The introduction of a weights

system furthermore limits the potential overrepresentation of participants that consistently gave the full set of three responses as opposed to participants that gave one or two responses to a factor of 1:2.

When examining the data, heavy tails were observed in the distribution of weights, partially caused by the uneven distribution of cues/responses. A normalisation of responses is not carried out since this difference in weight caused by the difference in valid responses per cue does reflect a real-world phenomenon: Responses where participants could not associate anything with a given cue word indicate the absence of a direct connection in the ML. In order to retain this information while combatting a disproportionate distortion of the observed weights a different path is taken: The weights for the 25th percentile of the cue-response distribution (i.e. the 25% of cues that received the least responses) are inflated to the value they would have reached at the first quartile threshold. In this case this means that they will be treated as if they had at least 9 responses per cue in SWOW-UK, and at least 76 responses in SWOW-EN. The inverse procedure is repeated for the 75th percentile (i.e. the 25% of cues that received the most responses), artificially compressing the weights to the value they would have reached at the 75th percentile threshold. In this case this means that they will be treated as if they only had 15 responses per cue in SWOW-UK and only 88 responses in SWOW-EN. This procedure has been adopted since it tackles outliers on both ends of the spectrum unlike more common methods such as a log transformation. It is important to note that the individual cue-response pair frequencies have only been affected indirectly by this since the distributions of how many sets of responses per cue are present in the dataset have been used as the basis for this transformation.

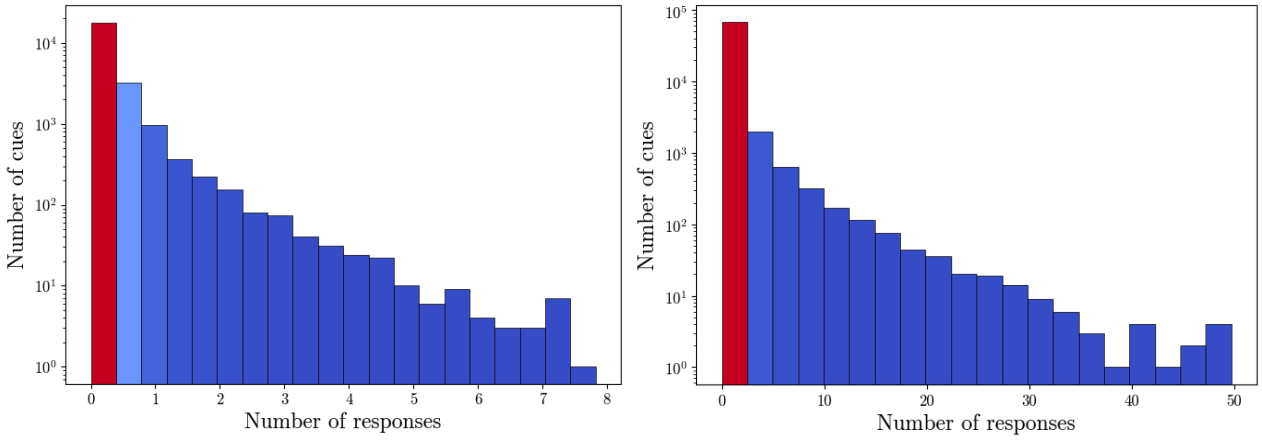


Figure 29: Logarithmic histograms showing the number of individual cue-response pairs in bins based on their weight scores for SWOW-UK (left) and SWOW-EN (right).

The weight distributions that have been achieved after weight-adjustment for both datasets are presented in two logarithmic histograms in Figure 29. These figures show that, as intuitively expected, the present datasets contain a very large number of low weight cue-response pairs (i.e.

infrequent combinations not shared by a large number of other participants) and progressively fewer high weight cue-response pairs.

The maximum weight lies at 41.1 for the combination *engine, google* in SWOW-EN; the theoretical maximal value assuming that every primary response given for the cue *engine* is *google* would be 63.3. The maximum for SWOW-UK lies at a weight of 7.3 for the combination *hang, wait* ; the theoretical maximal value assuming that every primary response given for the cue *hang* is *wait* would be 11.3[19]. A discussion of high-weight cue-response pairs is provided in the results section in Chapter 4.3.3.

Having completed the last major formative step of the pre-processing pipeline, the last topic to discuss is the introduction of a weight cut off value. While a strict threshold would be advisable when dealing with noisy data, the decision was made to only cut off the 10% lowest-weight associations for SWOW-UK due to its high quality and small size in comparison with the number of collocations emerging from the BNC 2014. A higher threshold of 50% was employed for the SWOW-EN since its larger size affords this.

While this is not put in practise in the present study, the pre-processing pipeline shows an optional step where researchers could filter for particular items of interest or delete items from pre-defined stop-word lists. The code written for this project explicitly allows for the implementation of such steps by other researchers due to its modular and interactive nature.

Having employed the pre-processing pipeline described in this chapter, the word association data now perfectly mirrors the output of the pre-processing script employed for the BNC 2014. This is crucial for enabling future graph theoretical comparisons. The CSV file export and the decision of carrying out initial explorations using LLN or other software (e.g. Cytoscape (Shannon et al., 2003) are identical to the considerations described in Chapter 4.2.1.4.

[19] Kindly note that the fractions here stem from split cues. This also explains *hang, wait* since *wait* is a very common response to the multi-word cue *hang on* (*wait* being the primary response to *hang on* in 8/14 cases), but also gets associated in other contexts such as *hanging*, *waiting*.

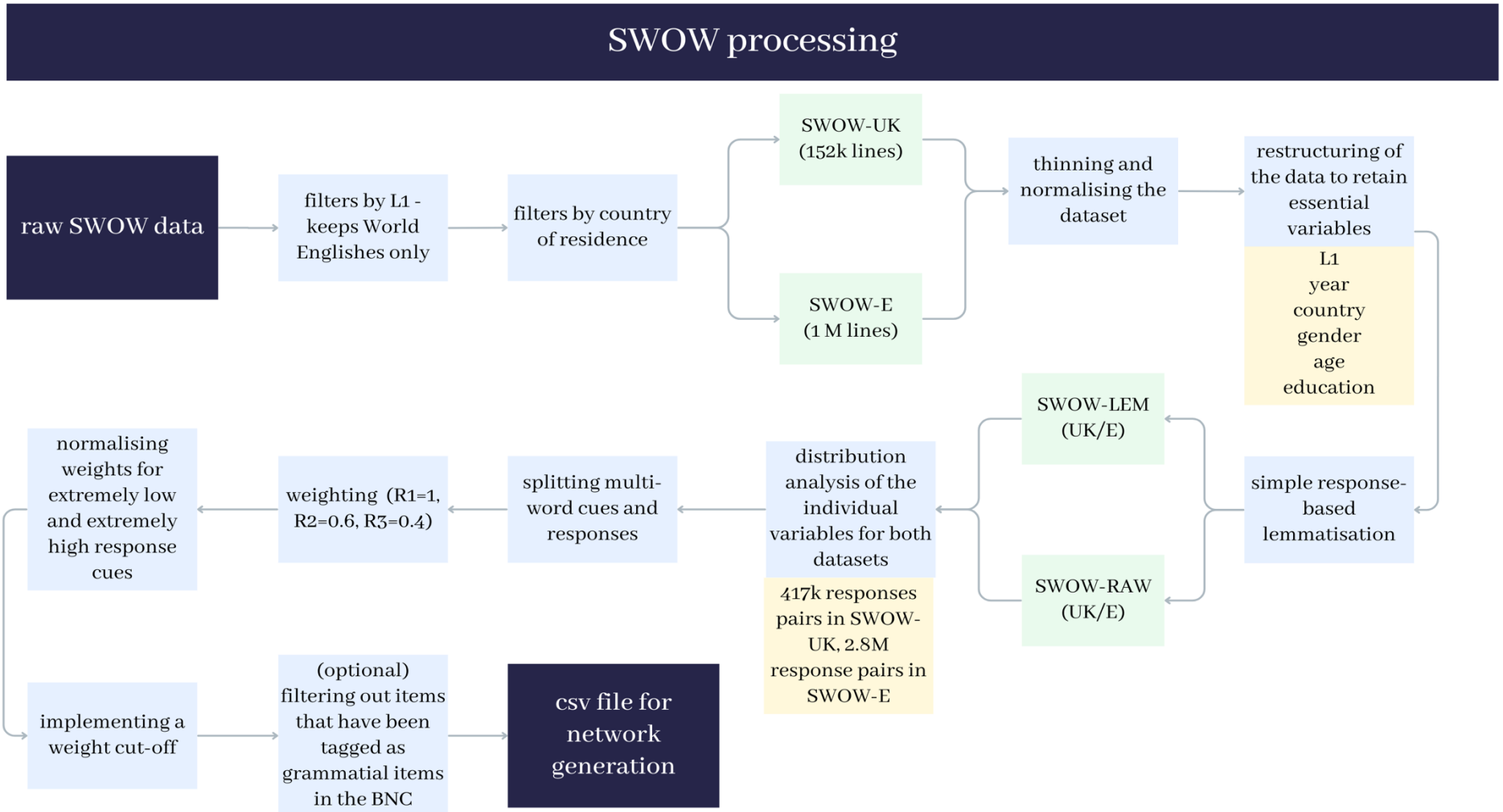


Figure 30: Schematic overview of the SWOW processing component in the LLN pipeline.

4.2.3 Systematic Exploration and Comparison of Large Linguistic Networks

After an exploration of the prerequisites for generating large collocation and word association networks and a detailed description of the steps taken for this project, this final methodological section formulates and justifies the approach to comparing the resulting networks.

A key question this thesis aims to explore is the extent to which different types of collocation networks (based on the same corpus) resemble word association networks. The selection of the Association Measure configurations is directly based on the outcomes of Chapter 3.3 concerning the selection of psycholinguistically plausible options since the aim is a meaningful comparison to a network based on a psycholinguistic dataset.

Firstly, it is essential to ensure that the 'atomic unit' of the networks is identical or as similar as possible. In this study, the unit used to constitute nodes is the lemmatised version of the raw association/word in the corpus. Secondly, all networks that will be compared against each other need to be loaded or generated. In the present study this is networks based on the output of fourteen different Association Measures, χ^2, log Dice, $\Delta P_{forward}$, r_{φ}, LL, $\Delta P_{backward}$, OddsRatio, and combinations of log Dice and $\Delta P_{forward}$, χ^2 and LL, log Dice and χ^2, $\Delta P_{forward}$ and χ^2, a three AM combination consisting of log Dice, χ^2 and LL, as well as a network based on SWOW-UK above first percentile, and a network based on full SWOW-EN above the fifth percentile respectively. In the empirical evaluation Chapters, whenever networks are referred to by the AM used to generate the underlying data e.g. 'the LL network' this refers to the network based on the top 1% highest scoring LL collocations. For a full description of percentiles and thresholds kindly consult the LLN pipeline in Appendix A.

Before embarking on the comparison stage of the project, a thorough inspection of the emerging networks and their properties is strongly encouraged. Besides the obvious knowledge gain from rooting the network findings in the corpus/word association data, and insights a partially qualitative manual analysis of particularly interesting graph theoretical features, this also serves to ensure the interpretability and validity of the larger-scale comparisons to follow. In practice, when examining the resulting corpus-based networks in particular with regards to key clusters a core difference is the consistently high degree of grammatical elements such as personal and possessive pronouns, determiners, and prepositions. The decision was made not to manually remove grammatical elements using e.g. a stop word list since the aim of this evaluation is a comparison of existing approaches to collocation identification and word association.

A large number of available approaches to network comparison - see Akoglu et al. (2015) for a general overview – is not suitable for the analysis of large linguistic networks. This is the case for

two reasons: Firstly, the unfavourable time-complexity renders any evaluation inaccessible to researchers unable to utilise high-power compute. Secondly, certain approaches are optimised for non-linguistic data and therefore make assumptions regarding frequency profiles and network properties that do not apply to frequency profiles found in linguistic networks. Based on the findings of a large-scale evaluative study by Wills and Meyer (2020), the decision was made to use NetSimile (Berlingerio et al., 2012) as well as adjacency spectral distance (Wilson & Zhu, 2008) for LLN analysis. Alongside this, the raw percentage overlap between the edges in the different networks is calculated for each pair of BNC 2014-based network and SWOW-based network. This does not only allow for an exploration of the shared items on a word level, rooting them in the text/association data directly, as well as enabling an effective comparison of different AM values against the 'gold standard' of the word association network.

Moving to the meso-level of analysis, MCODE (Bader & Hogue, 2003) clusters are extracted from the corpus-based networks in order to explore which highly interconnected groups of words emerge from the respective networks. Since this requires extensive analyses of a large number of emerging clusters, networks which show a favourable percentage overlap, NetSimile or adjacency spectral distance values when compared to SWOW-UK or represent common existing corpus linguistic approaches are prioritised and analysed in this manner. Since the full output of anchored clustering can result in large number of very small sub networks containing only two or three nodes, a filter was applied which only retains clusters that are both above the 0.8 percentile in terms of their size as measured per the number of nodes and contain at least five nodes.

Lastly, on the level of individual nodes the nodes with the highest betweenness centrality, eigenvector centrality, degree centrality, and clustering coefficient are extracted for each network and concordance lines contextualising their frequency of use and special structural position are presented.

4.2.4 Visualising Large Linguistic Networks

In the application of network methods to linguistic research, a pivotal stage is the selection of visualisation options. McCosker and Wilken (2014, p. 157) point out that many existing research outputs discussing data visualisation emphasise the *beauty* of the data as a core feature. While an aesthetically pleasing output is desirable, a good visualisation ultimately constitutes an analytical tool in its own right. This is particularly relevant in a linguistic context since the presentation of results can influence the pattern recognition of the researcher and thus prompt an in-depth analysis of certain features, concordance lines, and statistics over others and thus significantly impacting the research output given that research time is finite. This subchapter thus describes the difference

between standard tabular analyses and network approaches in the context of collocation analysis. This is, of course, not the only analytical application of visualisation to linguistic research, but it is the one that lies at the heart of this thesis. This chapter further presents force-directed graphs as one recommended option for large linguistic network visualisation, and posits an extension to collocation parameter notation, eCPN, in order to make visualisation choices as transparent as statistical choices.

Phillips' (1985) seminal early work on lexical networks already describes how textual discourse is structured into lexical patterns that can be depicted as networks of collocating words. Since then, the field of collocation network research has made significant progress, both practical and theoretical. The practical advances are tied to the evolution of computational technology, which now allows for the processing of extremely large datasets (comprising millions or billions of words) rendering the visualisation of holistic, corpus-wide collocation networks possible. The theoretical advances are borne from graph theory and its applications to linguistics as explored in this thesis.

The graph in Figure 31 illustrates the analytical advantage of using a network view for collocation analysis over the traditional tabular approach to collocation analysis. Both Figure 31 and Table 13 display a first- and second-order collocation network derived from the third-order collocates of *alcoholic* as a noun in the BNC 2014[20]. The table, sorted by the highest scoring log Dice values, provides Part-of-Speech information for each node and collocate. However, it is nearly impossible to simultaneously comprehend the various layers of information contained in the table (directionality, frequency of an element's occurrence as a collocate, Association Measure (AM) strength, POS group membership) when solely inspecting the table.

The view provided in Figure 31 on the other hand allows for a more immediate interpretation of all these factors since it is possible to directly grasp the rough distribution of colours (POS group membership), along with the position of individual words (centrality and frequency of occurrence of an element as a collocate), thickness and roughly also length of the edge (AM strength), and arrow direction (directionality). This mode of visualisation additionally highlights that *peanut_N* , *smoker_N* , *wine_N* etc. collocate with themselves. A further benefit of the network approach is the customisability given different research questions. Assuming that other features such as positive or negative collocation, semantic category, concreteness rating etc. are of relevance to a project the visualisation options can be changed i.e. to colour-code edges or nodes according to other

[20] Atomic unit: lemma_POS, AM: log Dice, Threshold: 6.73, Sentence-span, min collocation frequency: 10, min collocate frequency: 10 | (2-dimensional, Edge length: AM, Colour-coding: POS, Layout type: Edge-weighted spring embedded) – Visualisation Software: Cytoscape Shannon et al. (2003)

properties, change the size of nodes, their opacity, add edge labels, or even add further layers of edges.

Table 13: Customary tabular representation of third order collocates of *alcohol_N* in the BNC 2014 v.1.

node	*collocate*	*logDice*	*node*	*collocate*	*logDice*	*node*	*collocate*	*logDice*
peanut_N	peanut_N	9.67	tea_N	tea_N	7.93	electricity_N	consumption_N	7.32
bottle_N	wine_N	9.15	electricity_N	electricity_N	7.90	cigarette_N	tobacco_N	7.32
dietary_J	intake_N	8.78	drink_V	coffee_N	7.85	beer_N	bottle_N	7.30
dietary_J	magnesium_N	8.58	smoker_N	smoking_N	7.84	smoking_N	smoking_J	7.29
smoker_N	smoker_N	8.54	wine_N	bottle_N	7.76	alcohol_N	tobacco_N	7.27
calorie_N	intake_N	8.43	imperial_J	tobacco_N	7.74	dietary_J	dietary_J	7.26
tobacco_N	tobacco_N	8.33	drink_V	alcohol_N	7.73	substance_N	substance_N	7.23
magnesium_N	intake_N	8.25	drink_V	drink_V	7.69	smoking_J	cigarette_N	7.22
wine_N	wine_N	8.20	energy_N	consumption_N	7.66	intake_N	dietary_J	7.22
drink_V	beer_N	8.13	fuel_N	fuel_N	7.64	cigarette_N	smoking_N	7.15
drink_V	wine_N	8.09	bottle_N	bottle_N	7.61	alcohol_N	intake_N	7.11
fuel_N	consumption_N	8.08	coffee_N	coffee_N	7.60	tobacco_N	smoking_J	7.11
substance_N	misuse_N	8.07	tobacco_N	smoking_N	7.59	smoker_N	cigarette_N	7.09
energy_N	energy_N	8.06	alcohol_N	alcohol_N	7.50	intake_N	breath_N	7.07
drink_V	tea_N	8.03	alcohol_N	consumption_N	7.49	eat_V	drink_V	7.04
eat_V	eat_V	8.02	smoking_J	smoking_J	7.45	tobacco_N	cigarette_N	7.01
smoker_N	smoking_J	8.01	bottle_N	beer_N	7.44	alcohol_N	misuse_N	7.01
intake_N	intake_N	8.01	drink_V	bottle_N	7.44	imperial_J	imperial_J	7.00
beer_N	beer_N	7.95	intake_N	peanut_N	7.40			
tea_N	coffee_N	7.93	tobacco_N	smoker_N	7.40			

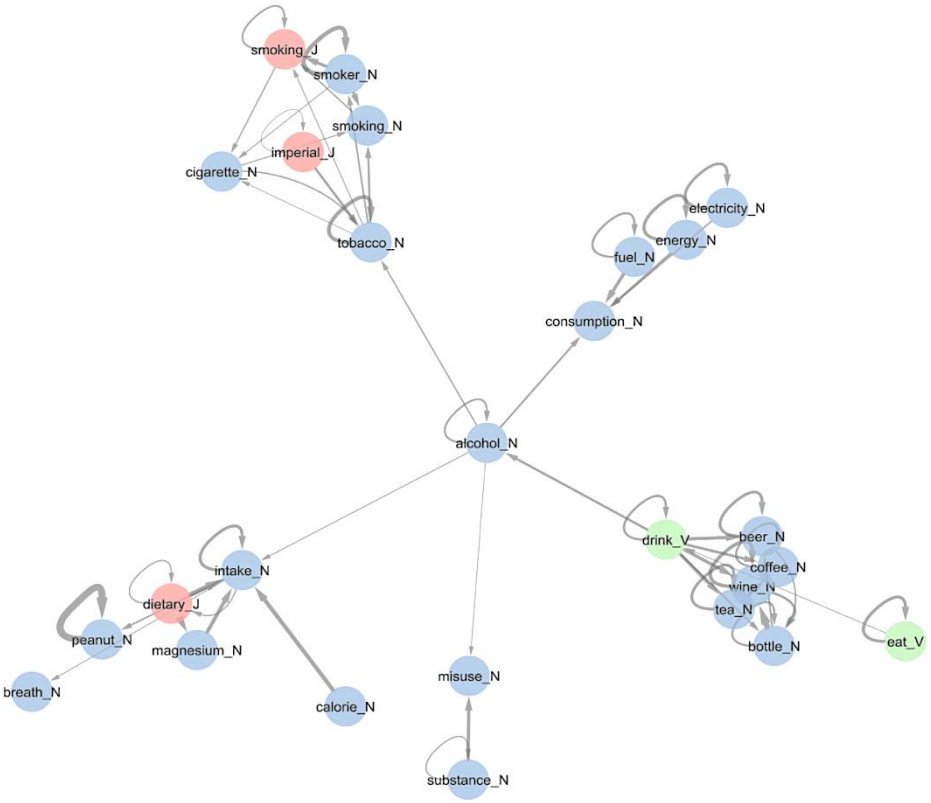


Figure 31: Visual representation of third order collocates of *alcohol_N* in the BNC 2014.

The network depicted in Figure 32 serves to illustrate this: It is identical to Figure 31 in its structure, but the colour coding denotes the ratio of male to female speakers using the respective lemma in the BNC 2014. Visualisations such as this are particularly useful since the nature of the colour-coding as a spectrum effectively represents the continuous nature of many sociolinguistic features such as gender ratio data, age, or social class.

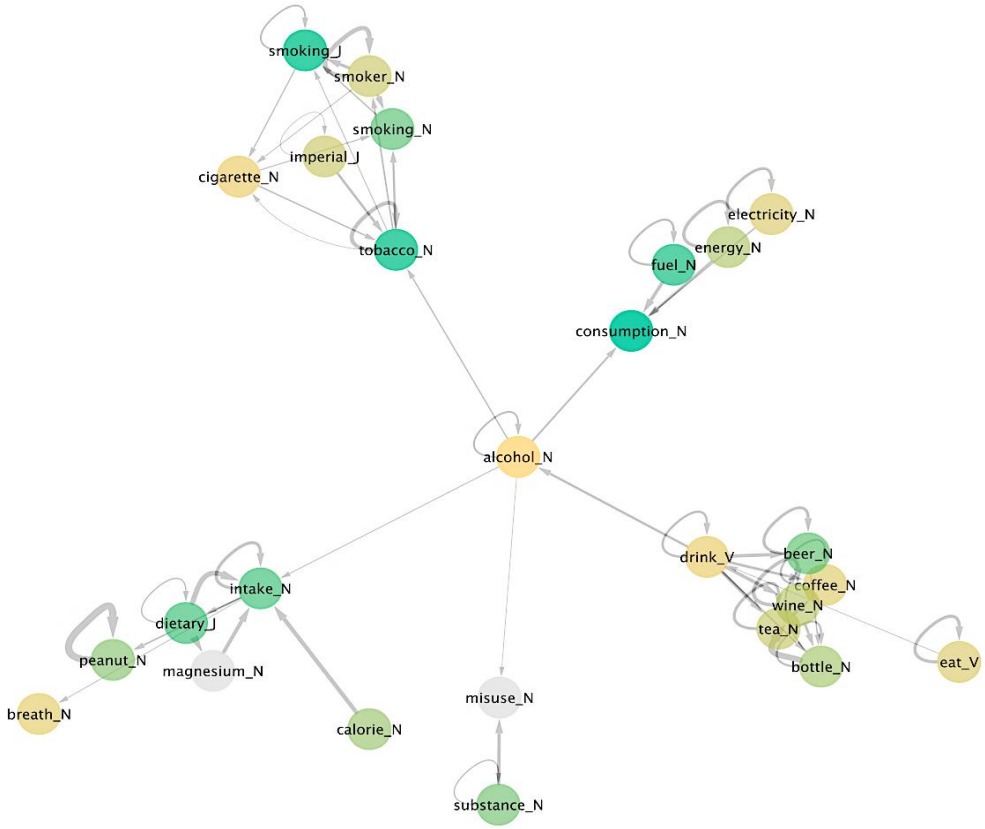


Figure 32: Visual representation of third order collocates of *alcohol_N* in the BNC 2014. Colour coding represents male to female ratio (yellow = more female, green = more male) based on relative frequencies of use for each word. Gender information not available for grey nodes.

Properties that emerge from tables at a glance are the total number of collocations, the rank order of collocations by the highest collocation score, and the exact individual association measure scores. While some research questions do not require any further properties and can be sufficiently answered using a short collocation table, a large number of research projects, e.g. in the field of discourse analysis, immediately benefit from properties emerging from a network approach. These properties, as demonstrated above, are the number of unique collocates (total number of nodes in the network), the number of collocations (total number of edges in the network), centrality/embeddedness (position of a word within the network), number of distinct discourses (number of unconnected sub-networks), and, crucially, metadata such as POS-group membership

which can be indicated via colour-coding/shapes of nodes. Additionally, a convenient way of making metadata visible serves to ease interpretation and pattern identification which is especially relevant in sociolinguistic studies. In the *alcohol_N* example it can immediately be observed that the majority of the words in the network are nouns, and the verbs *drink* and *eat* collocate in the same sub-area of the network. A strength of this approach is that any metadata can be plotted depending on the research question at hand (see Figure 32 for a sociolinguistic example).

Even in the non-clustered network above, different contexts of the word *alcohol_N* emerge from a network analysis due to the strong connectivity between mutually interconnected collocations, e.g. in the *eat_V|drink_V* branch which is populated with terms to do with general food/drink intake such as *bottle_N* and *coffee_N*, whereas the connection to *tobacco_N* leads to a cigarette-focused subsection of the network, and the *intake_N* branch leads to specialist terms describing surrounding a more formal discussion of dietary choices. Obtaining this information from a table alone is much more difficult as compared to a network view. Memorising the table would put a very high mental load on the researcher and require permanent and full accessibility of node-specific knowledge, i.e. how often is the word part of a collocation in the table, what is the gender ratio or mean age of speakers using this particular word etc. Keeping these factors at the forefront of the researcher's mind, particularly when dealing with corpus-wide or other large-scale collocation networks is not practicable.

After establishing the motivation for using a visual output for linguistic network analyses, the focus now lies on a description of recommended visualisation parameters for large linguistic networks. Clear communication about the capabilities and limitations of the visualisation is of utmost importance. In this research context, it is crucial to note that there is no singular "network" – the visualisations presented are merely one possible representation of the data.

Arranging large graphs with small world properties, such as the collocation networks analysed in this thesis, in Euclidean space presents substantial challenges (Hu, 2012, p. 527). Minimising overlap and stress of edges exhibiting large edge-length variance is particularly problematic in a linguistic use case. Maintaining readability and allowing for dynamic comparisons of unfiltered, base graphs with filtered graphs in a print format is also challenging.

Workarounds for these problems exist, such as printing detailed views of graphs or series consisting of the base graph and its filtered counterpart. However, these solutions split the reader's attention since they require repeatedly flipping back and forth. More dynamic representations that allow for zooming in and out of graphs, fading unwanted nodes out in the same view, and highlighting elements of interest are therefore more suitable for analyses of large-scale linguistic networks. One

particularly helpful approach in this context is 3-dimensional visualisations of graphs since this representation helps avoid confusion due to edge-overlap and can almost present an immersive experience centred around presenting the researcher with a maximum of easily interpretable information, see Figure 33 for an example.

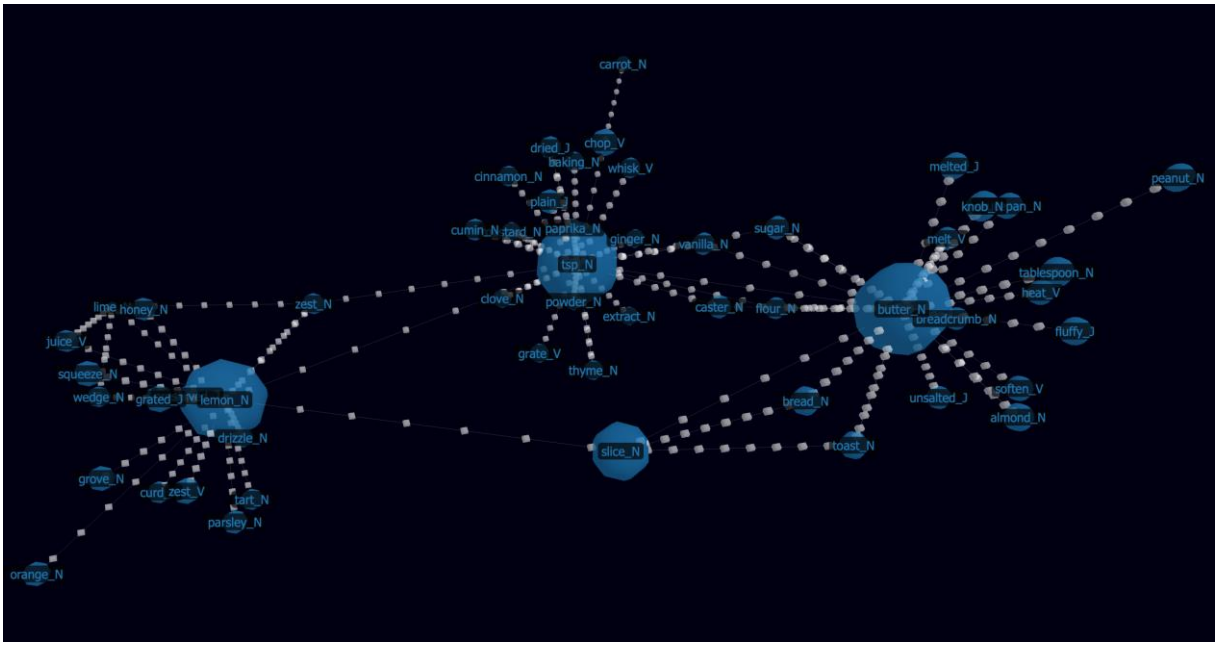


Figure 33: Screenshot of a three-dimensional cluster representation of a cluster based on substance abuse/food in the BNC 2014, dynamic visualisation available in Appendix A (Schmück, 2024).

The most intuitive type of network representation for large linguistics is force-directed graphs. These graphs aim to position the nodes of a graph by assigning them a weight on the basis of different properties; force-directed graphs are popular display option for a variety of research applications in different fields (Lu & Si, 2020, p. 9655). In the example in Figure 33 and throughout this thesis force-directed layouts are used to visualise networks. In Figure 32 one such layout, specifically the Kamada and Kawai (1989, p. 8) edge-weighted spring directed layout, is used. This algorithm represents nodes (here words that form part of a collocation) as floating connectors whose positions are determined by the force which the connected edges exert. This force is calculated by modelling the edges as physical springs between nodes which aim to bring words with a shorter shortest path between them closer to each other. Since a multitude of connected nodes are 'dragging' on each node, the positioning is not perfectly determined. Instead, the algorithm is designed determine how closely the nodes are placed to their theoretically optimal distance by assigning the placement options a metric called energy and minimising the overall layout energy. Due to this approach, they do not directly represent collocational strengths and rather display the graph theoretical distance between individual nodes. This is particularly valuable in contexts where information flow is relevant, such as the mapping of word associations in the mental lexicon.

When looking at entire large linguistic networks containing thousands of nodes (in the case of this thesis even tens and hundreds of thousands) the overall shape of networks this large tends to be circular regardless of considerable differences in visualisation parameters. The following example depicted in Figure 34 illustrates this. This view is created from MI^2 based collocations in the Spoken BNC 2014 and depicts the largest connected component only. It is generated in the open-source software Cytoscape using an edge-weighted spring directed layout (Kamada & Kawai, 1989).

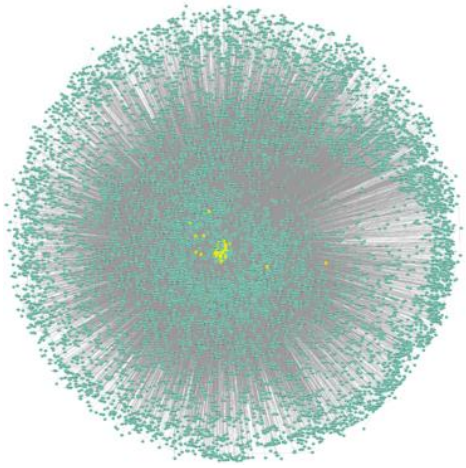


Figure 34: Edge-weighted spring-directed layout of a network representing the largest connected component of the Spoken BNC 2014 based on MI^2 scores[3] ≥ 10. Words connected to the term *normal* marked in yellow.

This thesis follows the principle that, when it comes to visualising large-scale networks, the different visualisation options and parameters have to be chosen with a clear research question and readability for a specific purpose in mind. Since it is not possible to take in the full information value contained in a network as large as this, it is therefore recommended to explore statistics alone on a holistic level while visualising networks on the meso-scale, e.g. via cluster analyses on n-th order collocations/word associations. A remedy for visual explorability of small to medium sized networks or clusters of networks is edge bundling (Holten & van Wijk, 2009, p. 983), see Figure 35. Where possible, edge bundling is applied in this thesis to reduce complexity and ease interpretability.

In order to provide a systematic way of documenting the use of visualisation options for collocation networks, an expansion of collocation parameter notation (CPN), a standardised way of reporting parameters for the identification of collocations, as proposed by Brezina (2018, p. 65) is introduced. CPN parameters include i) a unique Statistic ID linked to an equation, ii) Statistic name, iii) Statistic cut-off value, iv) Left and Right span, v) Minimum collocate freq., vi) Minimum collocation frequency and vii) any Filter applied to the data. Reporting of these parameters is essential for the transparency and replicability (Brezina, 2018, p. 75; Gablasova et al., 2017; Sinclair

et al., 2004) of the results. eCPN, an extended version of CPN, was therefore specifically designed for visual representation of collocations.

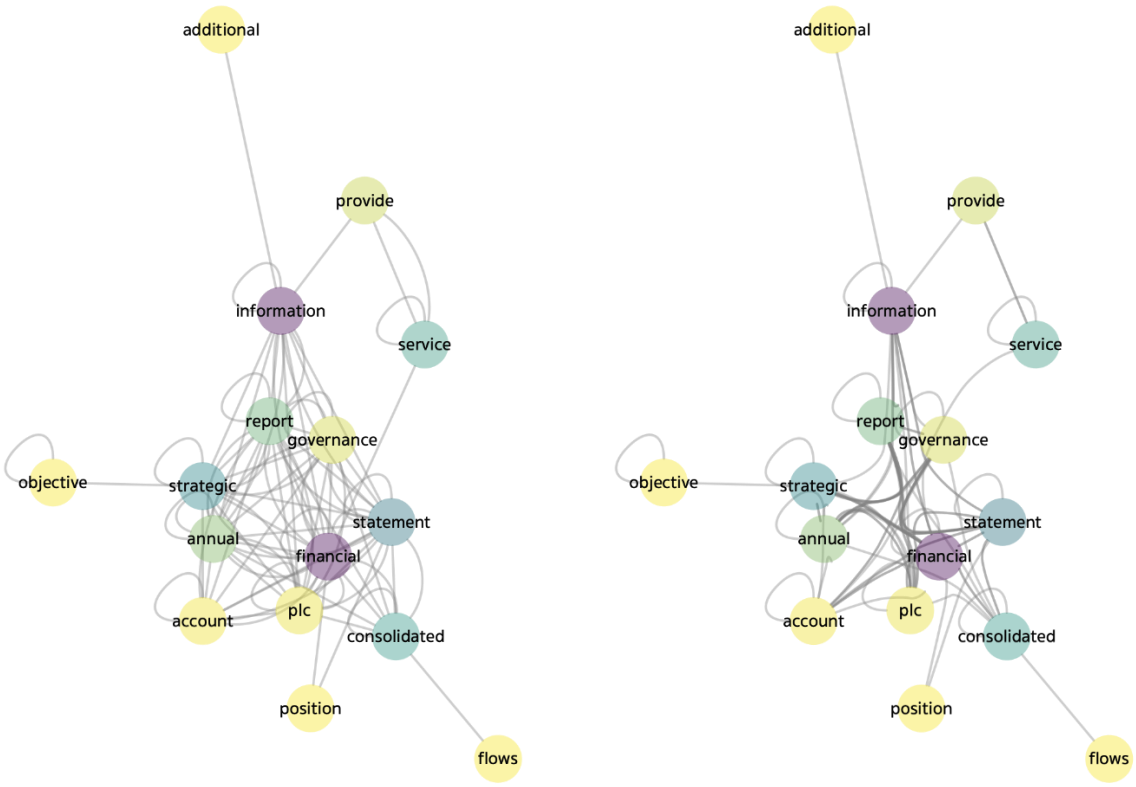


Figure 35: The effect of edge bundling (right) on a subcluster of the BNC 2014 centred around *financial information*[21] (left).

The collocation visualisation parameters listed here are commonly used, but not all parameters will be actively used concurrently to convey information in a single network. It is therefore recommended to state a collocation visualisation notation, e.g. as "default shapes, colour = % of female speakers, size = lemma frequency, directional, no filter, static, edge layout with edge length representing AM strength". It is further recommended to keep parameters the same or similar within a single piece of research to ease comparability; in these contexts the eCPN only needs to be defined once per document.

To conclude, linguistic network visualisations can act as a useful analytical tool, for example when exploring collocation networks given that weightings, displayed dimensions, filters etc. are carefully considered. Different sets of customisation options are available in different software packages, the visualisations in this thesis are based on Cytoscape (Shannon et al., 2003) and networkX (Hagberg et al., 2008), but notable alternatives are iGraph (Csardi & Nepusz, 2005), Gephi (Bastian et al., 2009). A conventionally publishable, two-dimensional, and static representation of the

[21] eCPN: lemma, log Dice, 5.42, sentence-span, 1 per million words | node colour: betweenness centrality (purple: max), directional, no filter, static, Kamada-Kawai layout

network necessitates further decisions about the visualisation, resulting in significant information loss and should therefore be accompanied by a dynamic version of the network wherever possible.

Table 14: eCPN notation

Collocation identification	Atomic unit (lemma, word, etc.) Including POS	Statistic	Statistic cut-off value	L – R Span	Minimum frequency
Collocation visualisation	Edge & node shapes	Colour coding	Node size	Directionality	Filter
	Edge Layout (e.g. attribute based/AM-based static /dynamic)				

The pipeline developed for this thesis is intended to be published and used by other researchers. Therefore, working towards an accessible and transparent interface and emphasizing ease of usability is important (Burghardt, 2018, p. 327). For this reason, full Jupyter notebooks containing extensively commented code can be found in Appendix A alongside network files that can be read using NetworkX (Hagberg et al., 2008) as described in the code, or via open source software Cytoscape (Shannon et al., 2003).

4.3 Results

This Chapter presents the empirical results related to RQs 2 and 3. This is structured into general questions prefacing and framing the answers to specific RQs, and results from the macro-, meso-, and macro-levels of the networks. Since the research questions at hand aim to explore the usability of the method and structural differences between word-association-based networks and collocation networks, several graph theoretical measures and individual nodes with graph theoretically relevant properties such as high centrality are of particular interest; these are explored in Chapter 4.3.4. These meta-parameters can be seen as a means to quantify differences and similarities of the two networks, and they aid an exploration of the nature of the datasets based purely on their intrinsic properties rather than the researcher's intuition. The following sub-sections explore general questions pertaining to the usability of large networks for linguistic analyses, provide an empirical evaluation of structural comparisons as well as a qualitative exploration of similarities and differences between the networks thus reiterating and systematically answering RQs 2 and 3.

4.3.1 General Questions

At a macro level, the initial question revolves around the intuitiveness of large linguistic network interpretations. The exploration of seventeen networks, fifteen of which are collocation networks based on the BNC 2014 and two of which are word association networks based on SWOW, shows

that the resulting networks are large and thus difficult to visualise, yet easily accessible for analysis and in-depth exploration. A normalisation of all AMs and associative weights to a scale from 0.01 to 1 allows for easy structural comparisons and the provided pipeline (Appendix A) allows for extensive searches of individual terms in all of these networks, highlighting if they are contained in the respective network, and how well interconnected they are.

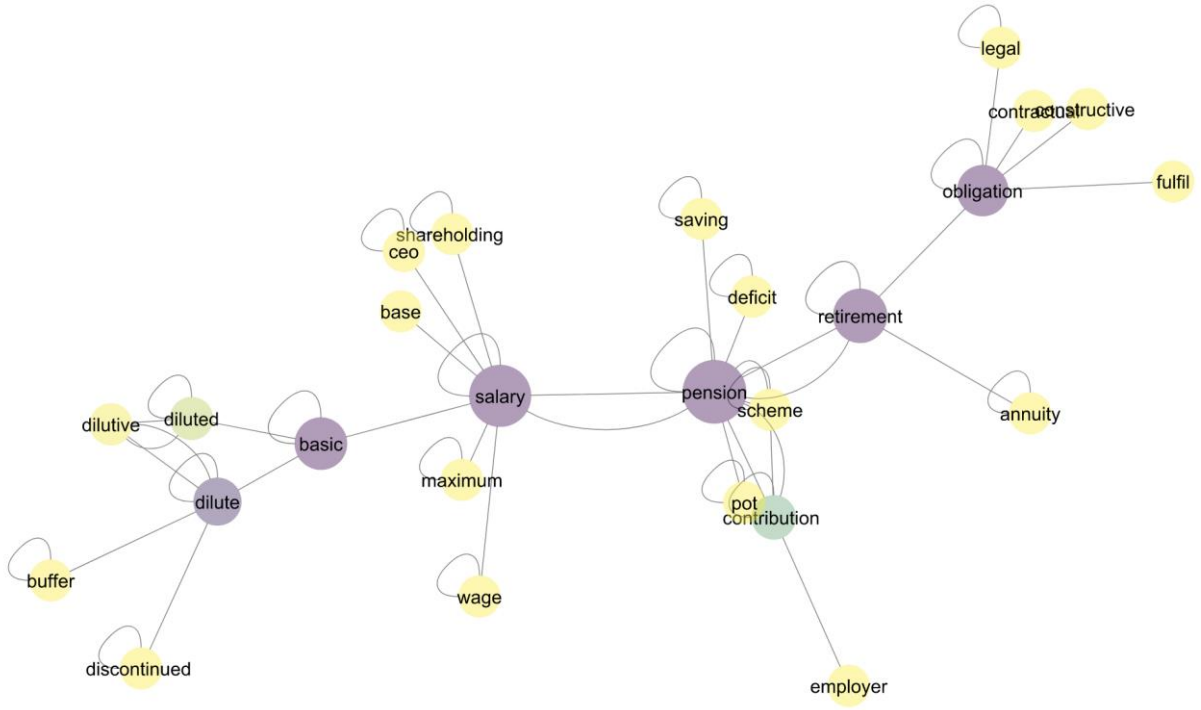


Figure 36: Pension/Salary key cluster extracted from the BNC 2014 - log Dice. Node colour represents betweenness centrality (purple: max, yellow: min).

Emerging key clusters in particular lend themselves to analyses and are presented here in order to illustrate the interpretability of LLNs. The first cluster (Figure 36) is the pension/salary key cluster extracted from the BNC 2014 using log Dice. The chain of connected terms — *dilute* > *basic* (semantically ambiguous) > *salary* > *pension* > *retirement* > *obligation* — emerges organically from the data without requiring a pre-defined 6-gram and illustrates the dynamic nature of the networks. Along the path, each term branches out to further related terms; *pension*, for example, is connected to *pot*, *saving*, *deficit*, *scheme* and *contribution* and the latter two are, again, interconnected. The visual representation effectively showcases the interpretability of complex interrelations among these terms. A further analytical step, omitted here for reasons of brevity but carried out for high degree centrality/betweenness centrality/eigenvector centrality nodes in Chapter 4.3.4, is exploring the associated concordance lines.

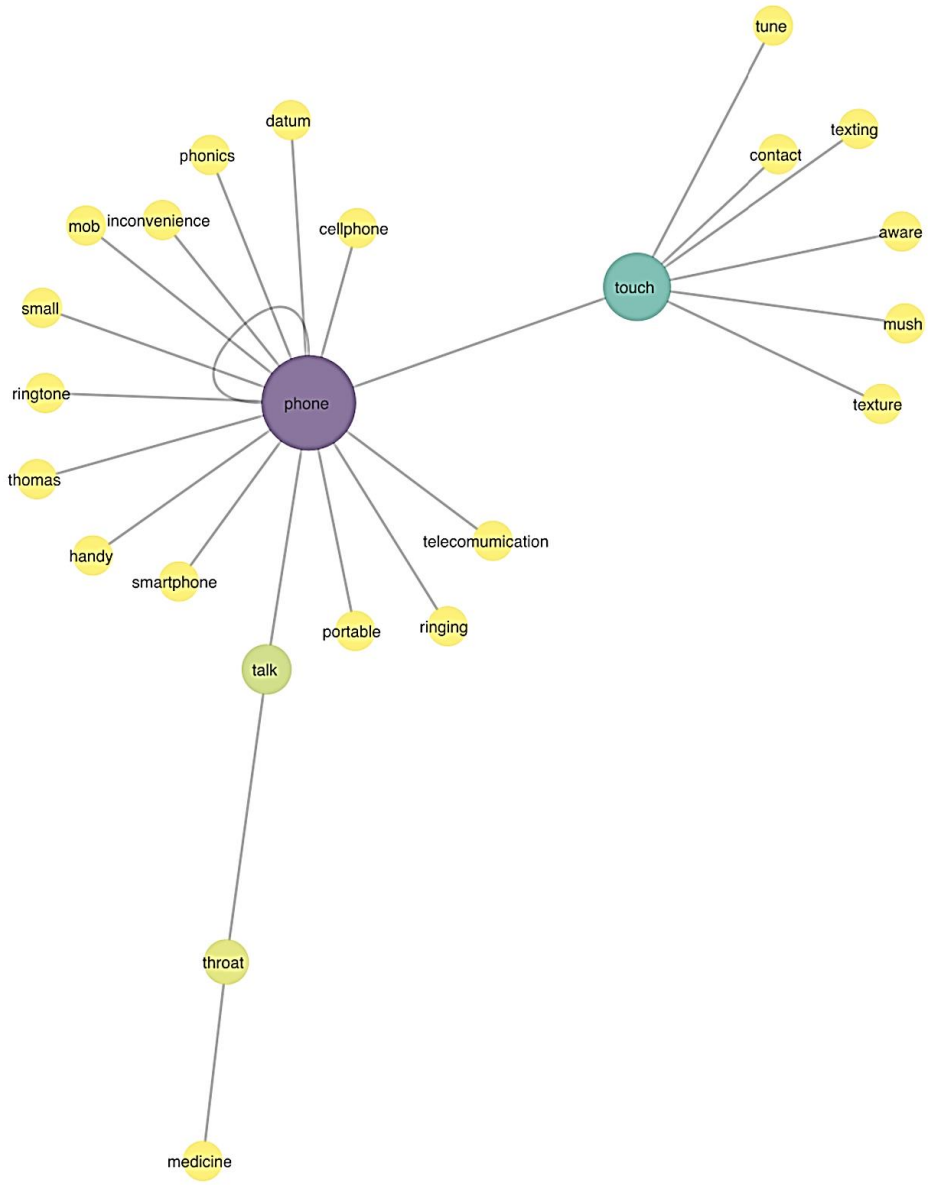


Figure 37: Phone/Touch key cluster extracted from SWOW-UK. Node colour represents betweenness centrality (purple: max, yellow: min).

The second example, depicted in Figure 37, is the phone/touch cluster extracted from SWOW-UK. This network shows a strong connection between the terms *phone* and *touch*, and, as is typical for word association clusters, near perfect star sub-networks centred almost perfectly around a single term. The shape immediately highlights structural differences between word associations and collocation-based networks with much higher clustering around select terms and less interconnectivity in the SWOW network. Common associative flows, such as *medicine>throat>talk* or *talk>phone>touch*, can also be identified and interpreted very easily using this visualisation. Looking qualitatively at the associations contained in the star network structure surrounding *phone* a large number of near-synonyms like *smartphone* and *cellphone*, as well as *handy*, the German word

for cellphone, emerge alongside related terms like *ringing*, *ringtone*, and *portable*. Interestingly, value judgements such as *inconvenience* also appear in the phone cluster.

Overall, linguistic analysis of networks, especially key clusters, reveals structural connections without requiring exhaustive re-reading of tabular statistics on collocation or word association. Since the remaining subchapters explore network analyses in depth on all three levels no further examples are provided here, but on a general level it can be observed that networks, and in particular key clusters, lend themselves easily to linguistic analysis and the structural connections become apparent without requiring intensive re-reading of collocation or word association statistics presented in a tabular format.

4.3.2 Macro-Level Empirical Evaluation of Structural Comparisons

This Chapter presents the results from the macro level exploration of the seventeen networks generated on the basis of collocations found in the BNC 2014 and word associations in SWOW. The first question to be explored here is what the graph theoretical properties of the resulting networks are. Table 15 shows a full overview of the results.

Examining the statistics on the overall shape of the networks in terms of the number of nodes and edges shows there is a large variation from as little nodes as just under 2,000 for the χ^2 log Dice network to almost 27,000 notes for the backwards translational probability network. It is particularly interesting to note that the number of edges does not positively correlate with the number of nodes with low node networks such as the log Dice LL χ^2 network exhibiting the highest number of edges at about 42,000. Naturally the SWOW-UK network is of particular interest here since it is the source for all later comparisons. Both the number of nodes and the number of edges for this dataset are close to overall average values at about 5,600 nodes and 21,000 edges.

An examination of the number of nodes and edges, however, does not tell the full story which becomes apparent when comparing the network with the most similar number of nodes and edges, $\Delta P_{forward}$ χ^2. The pattern observed in this network in almost all other macro level metrics is substantially different to the metrics observed for the word association networks. A further observation worth highlighting is that the log Dice LL χ^2 network is an outlier in many respects. It does not only contain the highest number of edges but also the highest mean degree centrality, the highest average degree, and the highest density as well as the lowest number of strongly connected components. Other features of this network including qualitative information on which words contribute most to the mean degree centrality, and average degree are presented in Chapter 4.3.4.

	SWOW UK	SWOW E	X2 ll	logD ll X2	X2	ll	X2 logDice	deltaPf X2
number of nodes	5635	9608	4462	2896	4772	7997	1950	9554
number of edges	20785	35634	33400	42057	32405	35997	12670	26603
density (*100)	0.07	0.04	0.17	0.50	0.14	0.06	0.33	0.03
average clustering coefficient	0.26	0.22	0.28	0.21	0.19	0.16	0.15	0.12
characteristic path length of largest component	3.71	3.79	2.86	3.24	3.15	3.77	4.29	3.35
mean degree centrality (*100)	0.13	0.08	0.34	1.00	0.28	0.11	0.67	0.06
mean eigenvector centrality (*100)	0.76	0.48	0.59	0.67	0.53	0.31	0.59	0.05
average degree	3.69	3.71	7.49	14.52	6.79	4.50	6.50	2.78
number of selfloops	180	274	1030	1237	1112	1673	982	265
strongly conn. comp.	5233	9169	1969	802	2080	5252	943	9407
number of comp.	1	1	1	1	1	796	48	2
transitivity	0.01	0.02	0.07	0.24	0.09	0.08	0.32	0.24

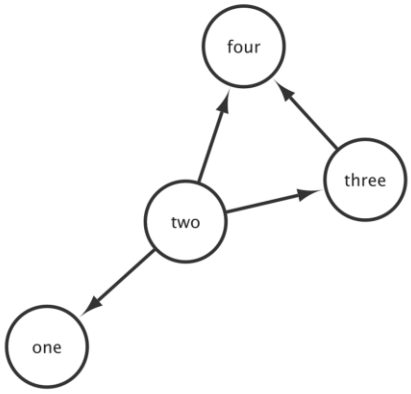
	logD	ll logD	pearsonsrho	OddsR	logD deltaPf	deltaPf	deltaPb	poisson
number of nodes	16519	8829	20183	21344	12487	25807	27982	11630
number of edges	35997	23819	35997	35997	12780	35997	35993	35995
density (*100)	0.01	0.03	0.01	0.01	0.01	0.01	0.00	0.03
average clustering coefficient	0.08	0.08	0.07	0.06	0.04	0.01	0.01	0.00
characteristic path length of largest component	8.18	6.14	7.95	8.48	11.57	3.43	3.23	6.19
mean degree centrality (*100)	0.03	0.06	0.02	0.02	0.02	0.01	0.01	0.05
mean eigenvector centrality (*100)	0.06	0.13	0.05	0.08	0.03	0.01	0.01	0.40
average degree	2.18	2.70	1.78	1.69	1.02	1.39	1.29	3.10
number of selfloops	4119	2161	2766	2263	247	30	20	0
strongly conn. comp.	13030	6908	16272	17032	12346	25797	27977	6568
number of comp.	1873	1361	1529	1531	2381	948	785	1
transitivity	0.48	0.32	0.14	0.08	0.26	0.02	0.00	0.00

Table 15: Macro-level graph theoretical parameters in BNC 2014 based collocation networks and SWOW-UK/SWOW-EN networks. Blue = highest value in column, white = lowest value in column.

An interesting point from the perspective of characterising the outputs of different association measures is the number of self-loops, i.e. words that collocate with themselves, produced by each statistic. The largest number of self-loops is found for the log Dice networks at about 4,000, the Poisson network, on the other hand, exhibits none. SWOW based networks generally do produce self-loops, but a reasonably low number of them at 180 for SWOW-UK and 274 for SWOW-EN. Examining self-loops is particularly interesting since existing small-scale collocation networks do not generally have the capability to display them leading to their dismissal in analyses.

Beyond this, it is also interesting to note that the log Dice $\Delta P_{forward}$ as well as OddsRatio, r_φ, LL log Dice, and log Dice networks produce graphs that 'fall apart', i.e. have thousands of unconnected subcomponents, whereas the six other networks, one of which is the SWOW-UK network, are fully connected.

Before turning the attention to qualitative differences, the following section examines structural similarities and differences between the networks generated for this thesis. This represents an intermediary step that takes in both large-scale statistical information such as the information presented in Table 15 into consideration, but also goes a step further in allowing for contrasting the structural similarities and differences between the graphs. In order to achieve this, two different graph theoretical methods for network similarity measurement are harnessed: NetSimile (Berlingerio et al., 2012), and Adjacency Spectral Distance (ASD; Wilson and Zhu (2008)). While NetSimile is based on the Canberra distance between statistics similar to the ones presented in Table 15, ASD takes a different approach. In order to measure the level of similarity between two graphs using ASD, the graphs are transformed into an adjacency matrix representation of themselves. This takes the shape of a symmetrical matrix where the length represents the number of nodes and the values within the matrix are populated with the respective edge weights. An example of this representation is provided in Figure 38.



	one	two	three	four
one	0	1	0	0
two	0	0	0	0
three	0	1	0	0
four	0	1	1	0

Figure 38: Force-directed layout (left) and matrix representation (right) of a directed sample network.

Creating adjacency matrices for each of the networks then enables a comparison of the similarity and difference between the edges present in the respective graphs; see Jurman et al. (2011) for a detailed description of this methodological approach. Figure 39 shows the results emerging from NetSimile in a heat map representation, smaller values represent more similar networks as determined using this method. Figure 40 on the other hand presents the results emerging from ASD, again plotted with blue values representing more similar networks. Since both of these figures show a comprehensive overview of all pairwise comparisons between the generated networks but the present focus lies on similarity specifically in relation to SWOW-UK, Table 16 and Table 17 contain just the SWOW-UK similarities when measured against all other networks.

Table 16: NetSimile values between SWOW-UK and all other networks, smaller values are more similar.

network	netsimile
SWOW E	6.17
X2 logDice	14.86
X2	15.48
logD ll X2	15.49
ll logD	16.49
ll	17.41
poisson	17.46
X2 ll	18.40
ll deltaPf	18.88
logD	18.92
OddsR	19.05
pearsonsrho	19.63
deltaPf X2	21.30
logD deltaPf	23.17
deltaPf	28.25
deltaPb	28.95

The theoretical minimum of the Canberra distance measurement lies at zero, the upper bound is open-ended (e.g. for an infinitely large reference network with no overlap). An exploration of the similarity rating presented in Table 16 shows that, again, SWOW-EN shows the largest similarity with a reasonably low Canberra distance of 6.17. At roughly twice this distance (14.86) χ^2 log Dice is the most structurally similar collocation-based network, followed by χ^2 (15.48) and, with a near identical difference (15.49) log Dice LL χ^2. Most other networks lie at a distance of between 15 and 25. The largest Canberra distance to SWOW-UK is measured for $\Delta P_{backward}$ and $\Delta P_{forward}$ at over 28 for both networks, meaning that these networks are by far the least similar to SWOW-UK. Looking at the intra collocational relationships further it becomes apparent that several AMs produce networks that are structurally highly similar, e.g. r_φ and OddsRatio. To a lesser degree this high similarity is also measured between r_φ/OddsRatio and log Dice $\Delta P_{forward}$, log Dice, and LL log Dice respectively. A further point worth mentioning is the high similarity between $\Delta P_{backward}$ and $\Delta P_{forward}$ in conjunction with the strong dissimilarity of these two networks when compared to all others. This pattern prompts the investigation into the key qualitative differences between $\Delta P_{backward}$ / $\Delta P_{forward}$ and log Dice LL χ^2 in Chapter 4.3.4.

Before examining the ASD results It is important to note that the absolute values representing the adjacency distance are not directly comparable to the NetSimile values. In contrast to the NetSimile approach, results showing the adjacency matrix difference between each network pair show that the highest similarity to SWOW-UK is not displayed by SWOW-EN (28.24) but rather by the log Dice LL χ^2 network at 15.91.

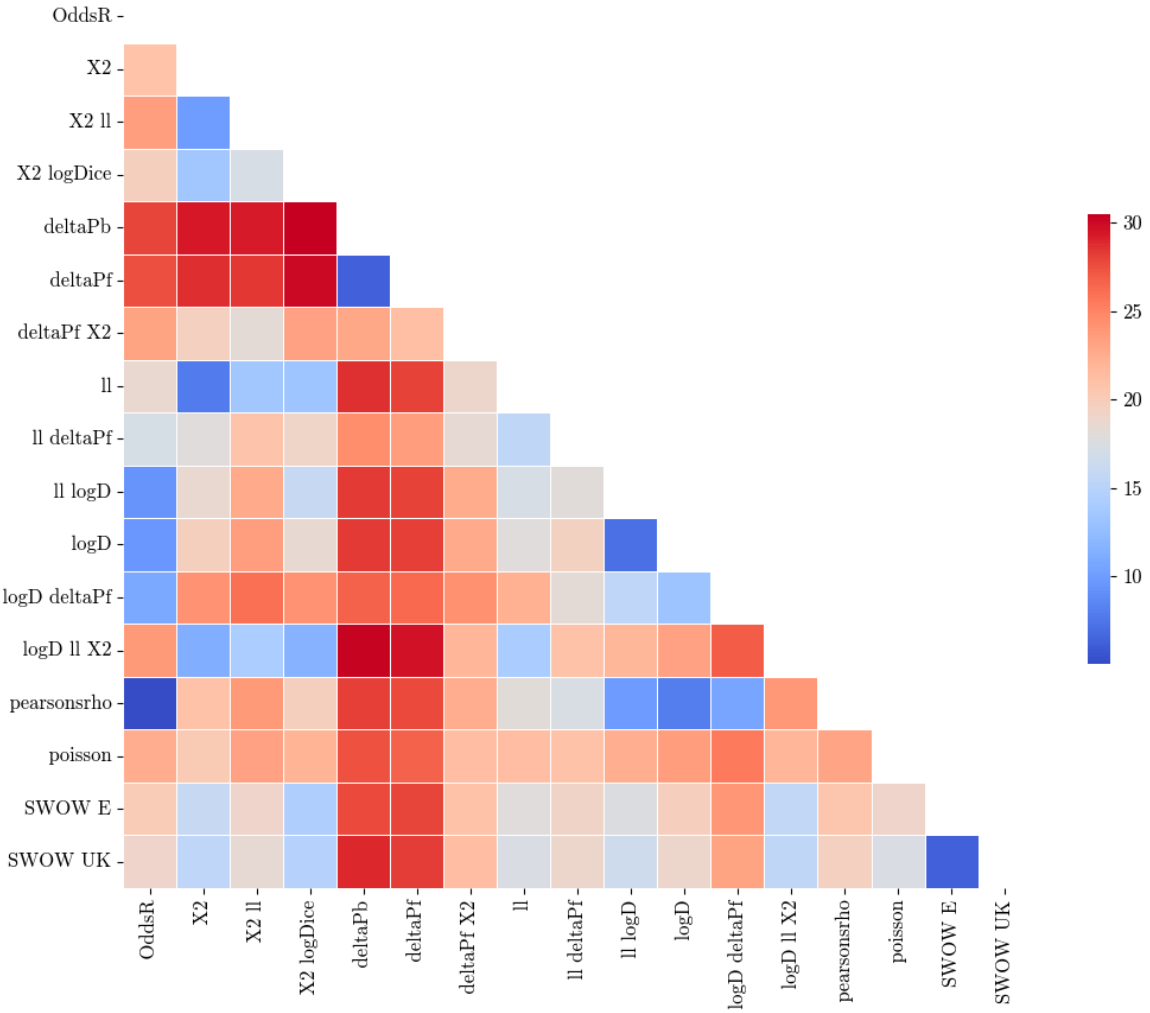


Figure 39: Heatmap showing inter-network similarity measured via NetSimile. Smaller (blue) values represent more similar networks.

The next most similar networks all display a sizable drop in similarity when compared to the log Dice LL χ^2 with χ^2 log Dice at 18.54, χ^2 LL at 20.98 and χ^2 at 21.19. In terms of the performance of the previous outlier networks $\Delta P_{forward}$ and $\Delta P_{backward}$ ASD also shows large differences to SWOW-UK at 28.36 and 29.71 respectively. OddsRatio and r_φ and show the least similarity at 70.56 and 75.49 respectively. Turning the attention to the full matrix depicted in Figure 40 it becomes apparent that SWOW-UK is generally dissimilar from all other networks and that SWOW-EN shows significantly more similarities with collocation-based networks; this may due to the size-effect emerging from an ASD measurement between differently sized networks. The intra collocation network similarities are highest between $\Delta P_{forward}$ χ^2 and LL $\Delta P_{forward}$, and further large similarities, again, exist between $\Delta P_{forward}$ and $\Delta P_{backward}$. Interestingly, OddsRatio and r_φ strongly differ from one another and not only from SWOW-UK as noted above, but also from all other collocation-based networks.

Table 17: Adjacency Spectral Distance values between SWOW-UK and all other networks, smaller values are more similar.

network	adj dist
logD ll X2	15.91
X2 logDice	18.54
X2 ll	20.98
X2	21.19
ll logD	23.45
logD	24.49
ll	24.82
poisson	26.05
logD deltaPf	26.66
SWOW E	28.24
deltaPf	28.36
deltaPb	29.71
deltaPf X2	30.36
ll deltaPf	30.63
logDdeltaP	55.69
OddsR	70.56
pearsonsrho	75.49

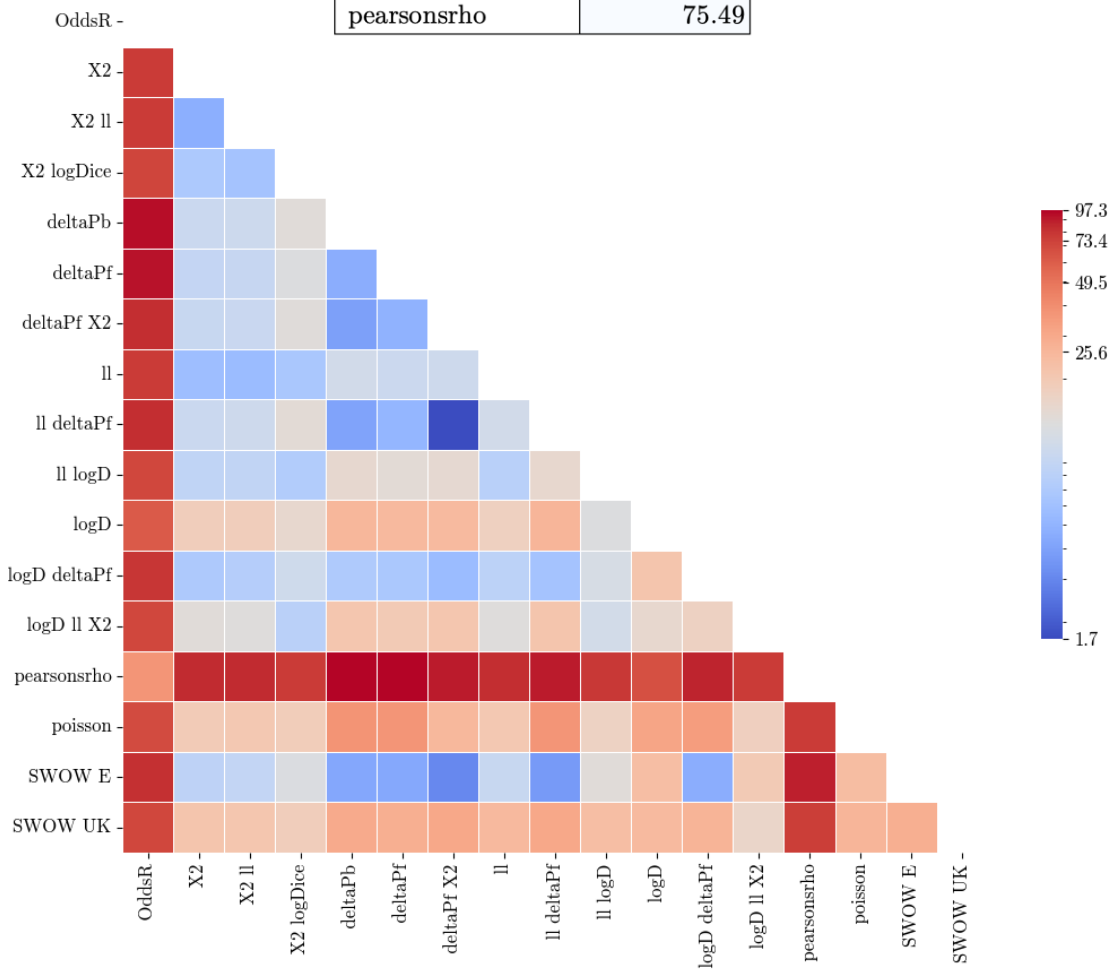


Figure 40: Heatmap showing inter-network similarity measured via Adjacency Spectral Distance on a logarithmic scale. Smaller (blue) values represent more similar networks.

After having looked at quantitative and structural features describing the shape of the networks, it is now time to focus on the qualitative similarities and differences via examining the percentage overlap of the exact edges which represent either collocations or word associations. Table 18 shows the percentage overlap between SWOW-UK and all other networks when taking all edges or just the strongest 100, 500, or 1000 edges into consideration.

Table 18: Percentage overlap between SWOW-UK and all other networks for all edges, the top 100, 500, and 1000.

	all	top 100	top 500	top 1000
SWOW E	25.06	100	100	100
logD ll X2	1.50	4	4	4.4
X2 ll	1.40	3	2.8	2.6
X2 logDice	1.39	0	0.8	1.3
X2	1.39	4	3.4	3.4
ll	1.32	2	3.2	3
ll logD	1.29	2	2.8	2.8
logD	0.92	2	2.6	2.4
deltaPf X2	0.49	0	1.6	1.1
pearsonsrho	0.46	3	2.2	1.8
ll deltaPf	0.19	1	0.2	0.4
logD deltaPf	0.18	1	0.2	0.3
OddsR	0.11	2	0.4	0.6
poisson	0.04	0	0	0.1
deltaPb	0.03	0	0	0.1
deltaPf	0.03	0	0	0.1

The different cut offs are considered since not all networks represent the same number of nodes as SWOW-UK and the overall percentage overlap is therefore swayed by whether or not the network has the same dimensions. Since all networks are larger than 1000 edges results from the strongest 100, 500, or 1000 edges are unaffected by this skew. Unsurprisingly the highest overlap emerges from the SWOW-EN network which has been used as a grounding for a realistic upper bound. An extremely high overlap was expected since SWOW-UK is a subset of SWOW-EN, and this is indeed the case with 100% overlap for the top 100, 500, and 1000 edges and 25% overlap overall. The colour coding in Table 18 acts as a visual aid to convey that not all subsections show the same best candidate ranking. Aside from SWOW-EN, the highest top 100 overlap is found in the log Dice LL χ^2 and χ^2 networks with 4% respectively, followed by the χ^2 LL and r_φ networks at 3%. Poisson, $\Delta P_{backward}$, and $\Delta P_{forward}$ share none of the 100 strongest edges with SWOW-UK. This picture changes when considering the top500 edges. The log Dice LL χ^2 network still shares 4% of the nodes with SWOW-UK, but χ^2, LL, LL log Dice, χ^2 LL, and log Dice now sit at around 3%. r_φ has dropped to 2.2% and Poisson, $\Delta P_{backward}$, and $\Delta P_{forward}$ still share none of the edges. Considering

the top 1000 log Dice LL χ^2 still leads the table of collocation-based networks at 4.4% overlap and χ^2, LL, LL log Dice, and χ^2 LL still show about 3% overlap each but their internal order has slightly changed with χ^2 LL falling from 2.8% overlap to 2.6%. r_φ has further dropped to 1.8%.

To sum up, the best performing association measures in terms of their percentage overlap with the SWOW-UK word association network are a combination of log Dice, LL, and χ^2, as well as a combination of χ^2 and LL, followed by χ^2. In terms of the best performing networks by NetSimile values, the combination of combination of χ^2 and log Dice performs best, followed by χ^2 on its own, and the same triple AM network based on log Dice, LL, and χ^2, that performed best in terms of the overall percentage overlap. Looking at Adjacency Spectral Distance, this triple combination, again, performs best, followed by a combination of χ^2 and log Dice, and, lastly, χ^2 and LL. Systematic structural differences from both SWOW-UK and all other networks are observed for r_φ, Poisson, $\Delta P_{backward}$, and $\Delta P_{forward}$.

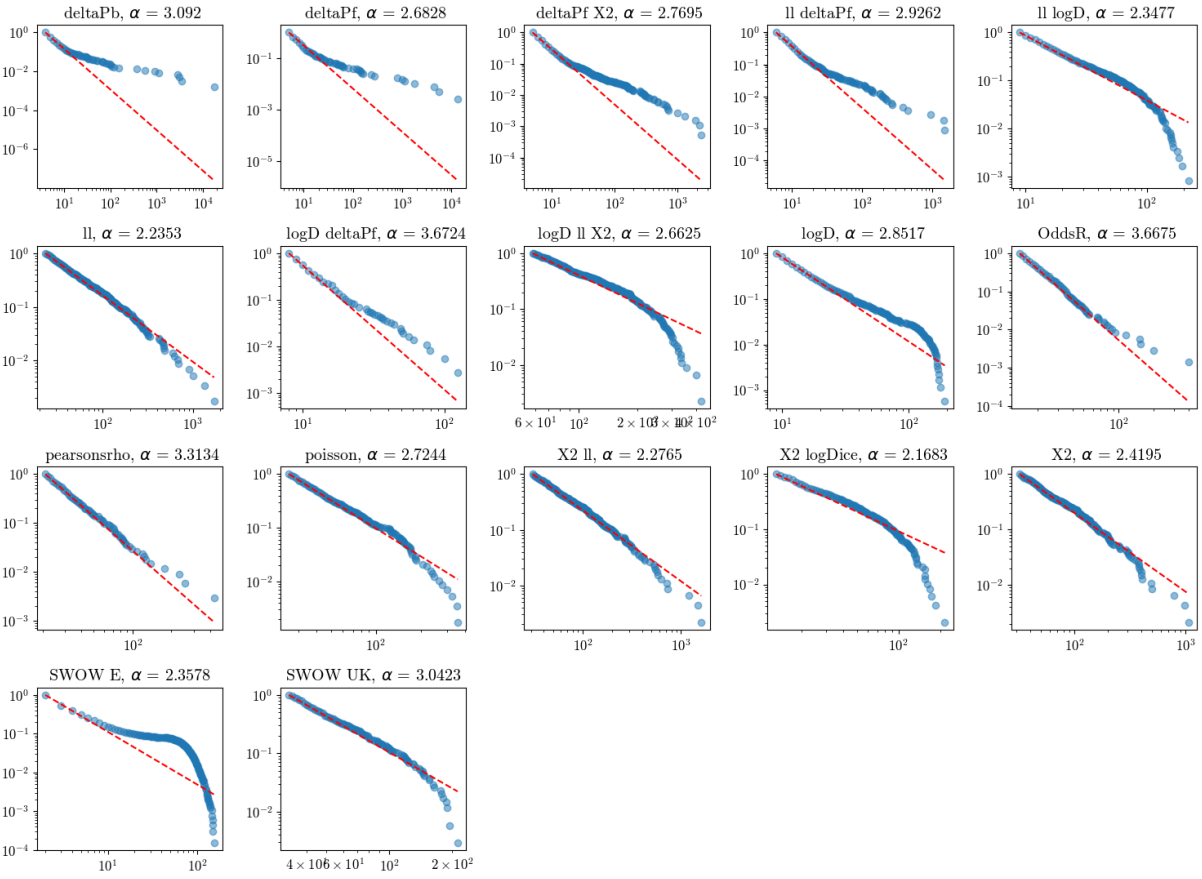


Figure 41: CCDFs P(x) and their maximum likelihood power-law fits all examined networks plotted on a log-log scale. For better readability please see Appendix B.

Lastly, Figure 41 shows the complementary cumulative distribution function (CCDF) P(x) which denotes the fraction of edges that have degree x or greater. The general reader might prefer to skip this brief section since it is more graph theoretical than linguistic in nature. An exploration of the

power law fits shown in this plot indicates that all $\alpha > 2$ which makes all networks scale-free. Most networks further fall into the $k^{-\alpha}$ with $2 \leq \alpha \leq 3$ (Siew et al., 2019, p. 6). A degree distribution of this shape – specifically with an α of 2 or higher – consequently proposes that the distribution stretches towards infinity (or is *scale-free*).

4.3.3 Meso-Level Empirical Evaluation of Structural Comparisons

Having explored the macro-level properties of the networks at hand, this Chapter allows for a closer look at the meso-level of a number of particularly relevant networks as identified in 4.3.2 namely the SWOW-UK, log Dice LL χ^2, and log Dice networks. Unfortunately, a thorough examination of clusters emerging from all seventeen networks far exceeds the spatial limitations of this thesis and the analyses presented here are therefore limited to the three abovementioned networks, networks showing promise in the previous Chapter representing common psycholinguistically plausible approaches used in Corpus Linguistics have therefore been selected and prioritised.

Exploring linguistic networks via clusters allows for more traditional linguistic analyses including discussions of semantic domains, cohesion, and, in the case of collocations, the type of collocation (see Chapter 2.4.2) at hand. This is particularly of interest considering that previous research indicates a repeated use of the same word class for word associations in adults (Aitchison, 2008, p. 86). A qualitative examination of clusters is further relevant in order to determine whether or not the networks capture primarily paradigmatic relationships since this feature is particularly relevant in a word association/ML context (Fitzpatrick & Thwaites, 2020). These points are examined after a contrastive analysis of the different clusters and evaluation of how topically similar the emerging key clusters are.

The evaluation itself is carried out using the methodology described in Chapter 4.3.2 which generates MCODE (Bader & Hogue, 2003) clusters on the basis of each network. This analysis can be reproduced using the code provided in Appendix A. For each of the networks, key collocational or word association clusters are extracted and visualised in this manner before a manual analysis of their respective high betweenness centrality nodes allows for grouping them into semantic clusters. Full visualisations of these clusters suffer from a common issue in linguistic network analyses, a lack of space. To remedy this, dynamic zoomable versions of these clusters are provided in Appendix B, Figure 42 is used as a stand-in for the emerging clusters to provide the reader with an illustrative example. Tables showing the respective clusters and their manual classification into the domains such as the ones provided in Figure 43, Figure 44, Figure 46 and along the spectra of collocations established in Chapter 2.4.2 can also be accessed in Appendix B. All manual analytical

steps in this thesis such as the classification of clusters into key domains, the classification of closed class items into word classes, and the classification into types of collocation have been carried out by the author. Interpretative and therefore subjective work is necessary in order to obtain the respective classifications which means that diverging results can be expected when re-coding these examples. The full classification lists and tables are therefore included in the Appendix B and double-coding by other researchers is explicitly invited.

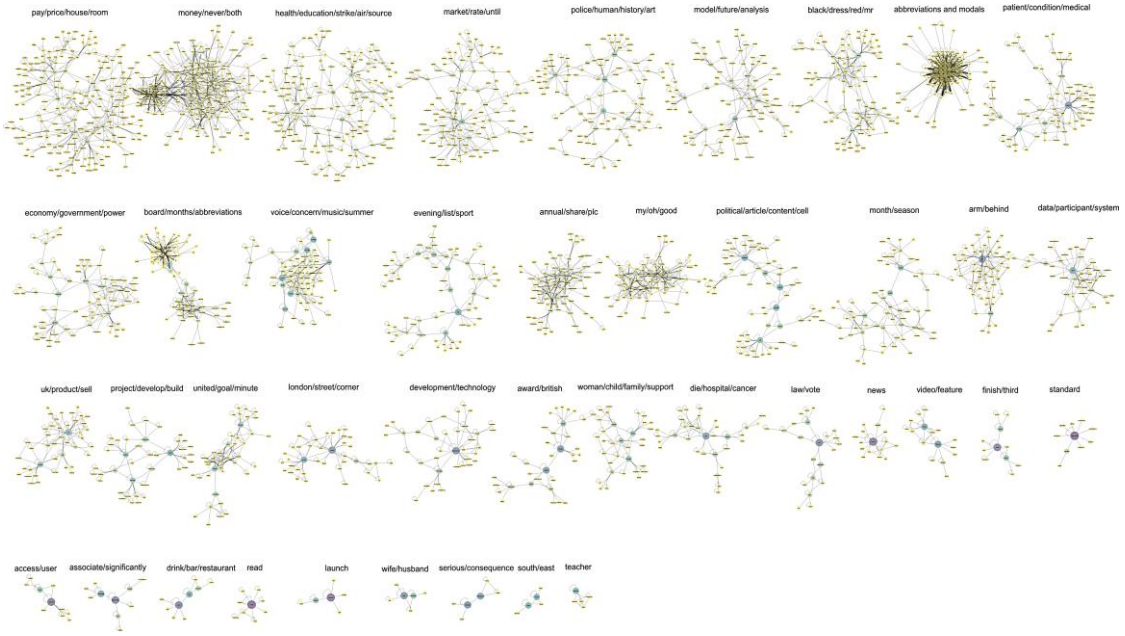


Figure 42: Visualisation of the annotated clusters emerging from the log Dice LL χ^2 network.

The cluster groups identified across all three networks range from hardly tangible AbstractConcept clusters such as the cluster surrounding *Luck* in SWOW-UK or the *Standard* cluster in log Dice LL χ^2 over clusters describing time, places, people, several scientific subdomains as well as media, religion, and entertainment. Beyond this, completely different types of clusters containing primarily grammatical (closed-class) items, groups of terms expressing judgement or positioning of the language user as well as register markers are also emerging from the clusters. For a contextualisation of the results, it is important to note that in this classification clusters surrounding individual people's names have been grouped under Individual, whereas clusters describing people in general, such as people carrying out different jobs or fulfilling certain roles within a family or other social construct as Person. Figure 43 displays the percentage of clusters belonging to the respective semantic groups for the clusters emerging from the SWOW-UK, log Dice LL χ^2, and log Dice networks. Percentages rather than the absolute frequencies are used here, since not all networks produce an equal number of cluster groups. The results show that SWOW-UK is overall the most action-focused with a quarter of all SWOW-UK clusters representing actions such as the *Leave|Take|Off* cluster and the *Search|Explore* cluster. Other cluster groups represented in SWOW-

UK are places, general terms, grammar, medical terms, expressions of judgement, technology, policing, people, abstract concepts, finance, food, the human body, clothes, nature, music, and religion. Notable similarities between SWOW-UK and log Dice LL χ^2 clusters constitute themselves in the focus on places and people. All SWOW-UK categories are also represented in at least one of the two collocation-based networks. When examining the log Dice LL χ^2 network, the cluster group picture is overall similar aside from a weaker focus on actions and a stronger focus on abstract concepts, time, sport, politics, and register markers. Around 12% of log Dice LL χ^2 clusters are centred around actions, followed by clusters representing time (11%), abstract concepts, technology, and place (each 10%). It is particularly important to note that the log Dice LL χ^2 focus on time is almost unique, with only log Dice also representing this category, albeit much less strongly. Lastly, the analysis of log Dice shows a strong focus on groups from specific registers and domains such as financial terms (17%), automotive terms (10%), names of individual people (9%), sport (9%), technology (8%), and law, education, and medical terms (8%). Many of the groups identified in log Dice are also unique groups which do not occur in any other clusters. Exemplary for this are the aforementioned automotive clusters, clusters describing individual people, foreign terms, law, abbreviations, chemistry, education, and leisure clusters.

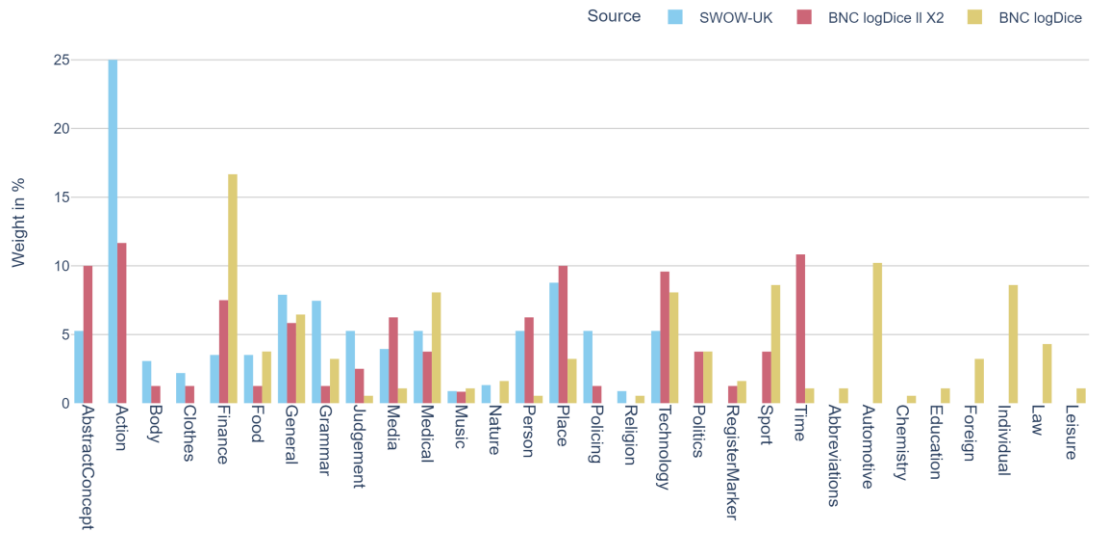


Figure 43: Cluster Group distribution for clusters emerging from the SWOW-UK, log Dice LL χ^2, and log Dice networks.

After having looked at the differences emerging from the cluster groups, it is also essential to look at the commonalities and overlap between word association clusters and collocational clusters. Overall, log Dice LL χ^2 shows a better fit when taking the word association network as the reference point than log Dice. This is the case since actions and abstract concepts as well as people, places, and media play an important role in both sets of clusters. The core groups similarly

represented in all networks analysed here are centred around general terms, medical and technological terms, as well as food.

Examining the presented analytical approach to network clustering in terms of their practicability, the classification task itself is quite simple since high BC items clearly indicate the overall theme of a cluster thus easing the load on the researcher to produce descriptors themselves as is the case in Multidimensional Analysis. It is, however, noticeable that different association measures and network generation pipelines lead to clusters with differing conciseness. log Dice clusters, for instance, tend to centre around a single theme whereas the larger clusters in log Dice LL χ^2 and SWOW-UK often contain more than one domain and were therefore tagged as multiple domains. An example for this is the *Tub|Warm|Change|Washing* SWOW-UK cluster which has been tagged as *General* (weight 0.33*)*, *Clothes* (weight 0.33*)*, and *Action* (weight 0.33).

Having examined the topics emerging from the respective clusters it is time to shift the focus towards the POS membership of the terms contained within the clusters as well as a classification of the strongest collocational/associative links into types of collocation/association. Before the results are presented, it is important to highlight a limitation of this analysis: the absence of any meaningful possibility to POS-tag word associations. Reliable tagging is impossible since there is no context from which the usage of a word could be derived due to participants providing individual, largely isolated words as responses. The lack of any further context prohibits interpretation as to the exact word sense intended by the participant, and thus often makes it impossible to reliably and robustly derive POS information. This can be illustrated when looking at a high betweenness centrality term in a Cluster emerging from the SWOW-UK network, the word *work*. This term could represent a noun or a verb, and even a closer look at the words associated with this term such as *tiring*, *arduous*, *remuneration*, and *cardio*, amongst others, only allow for a manual word sense disambiguation into a sense related of *work* [out] (again, interpretable as a noun or verb) and work as indicating wage labour. Given this one-word-association, determining whether *work* relates to the action of working or the concept of work itself is impossible. This limitation means that a nuanced examination of the hypothesis that the same word classes will be associated with one another is not possible on the basis of this dataset. Nevertheless, remarks are made regarding the predominance of closed versus open class words and specific types of closed-class words contained in the individual clusters in the following sections. In practice, the volume of words awaiting manual classification can be prohibitive as well. For this reason, only nodes with a non-zero betweenness centrality and a non-zero clustering coefficient as well as a degree of at least two have been analysed here. In the case of the present analysis this, for example, reduced the number of words to analyse from 3,674 to 985 for log Dice LL χ^2.

The manual classification into closed-class vs. open-class items shows a very similar tendency across all three networks: around 8.2% of the terms contained in the log Dice clusters are closed-class items, for log Dice LL χ^2 this figure sits at 8.3%, and SWOW-UK shows the lowest share of closed-class items at 7.1%. Beyond this, Figure 44 shows which types of closed-class items are contained in each respective group of clusters. Here, again log Dice LL χ^2 and log Dice show a rather similar picture containing all types of closed-class items. Across all networks, prepositions are most strongly represented (47% of closed class items in SWOW-UK; 26% in log Dice LL χ^2; 22% in log Dice), followed by adverbs (20% of closed class items in SWOW-UK; 18% in log Dice LL χ^2; 21% in log Dice). For the collocation-based networks, the next most frequent groups are number terms (15% in log Dice LL χ^2; 17% in log Dice), and possessive pronouns (7% in both). Neither of these are found in the SWOW-UK clusters. Wh-adverbs and personal pronouns, and conjunctions are overall less frequent but also represented across all three networks. Negation markers, determiners, demonstrative pronouns, and the possessive marker *'s* are unique to collocation networks. Since there are fewer items in closed class groups such as determiners and negations than in groups such as adverbs or prepositions, this low number is intuitive. In the case of *'s* this being a unique feature of log Dice LL χ^2 and log Dice is to be expected due to the splitting of the possessive marker from nouns as part of the pre-processing a corpus undergoes; this is naturally not the case for word associations.

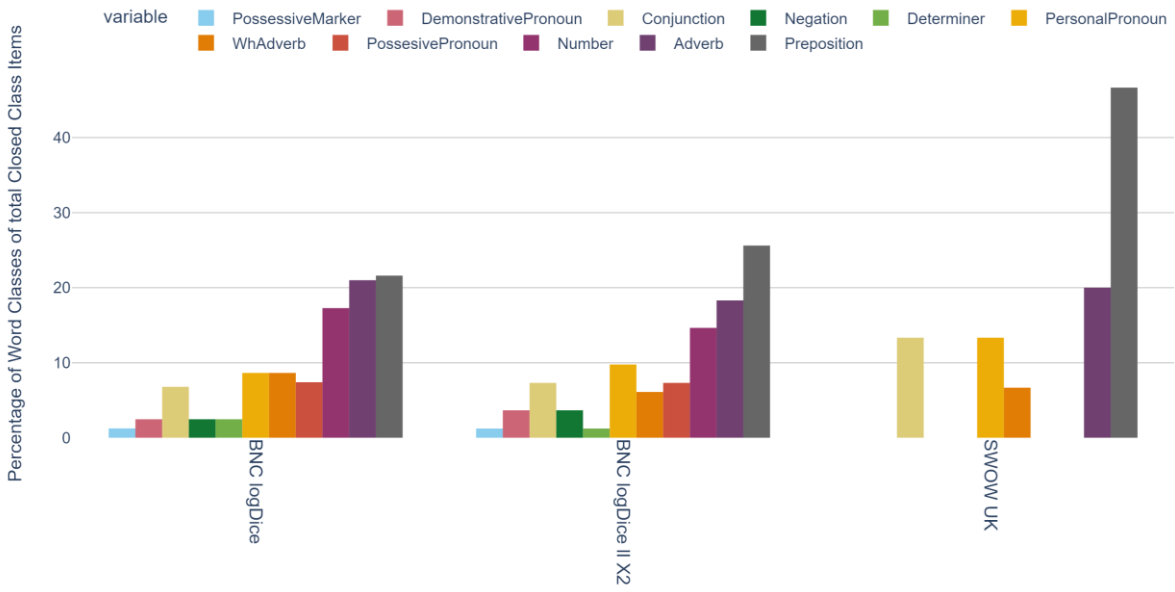


Figure 44: Percentage of word classes represented in closed class items for all key clusters emerging from each of the three networks.

Beyond strictly closed-class items, a breakdown of the word classes found in central hubs of the SWOW-UK clusters shows a high share of compound verbs (*give up*, *wake up*, *think over*, *make together*) and even a cluster consisting mainly of prepositions and compound verbs interacting with one

another, see Figure 45. Looking at log Dice LL χ^2 the breakdown of word classes found in central hubs exhibits a less pronounced focus on prepositions, but a similar share of adverbs and wh-adverbs when compared to SWOW-UK. The word classes found in central hubs for log Dice are overall very similar to the one observed for log Dice LL χ^2.

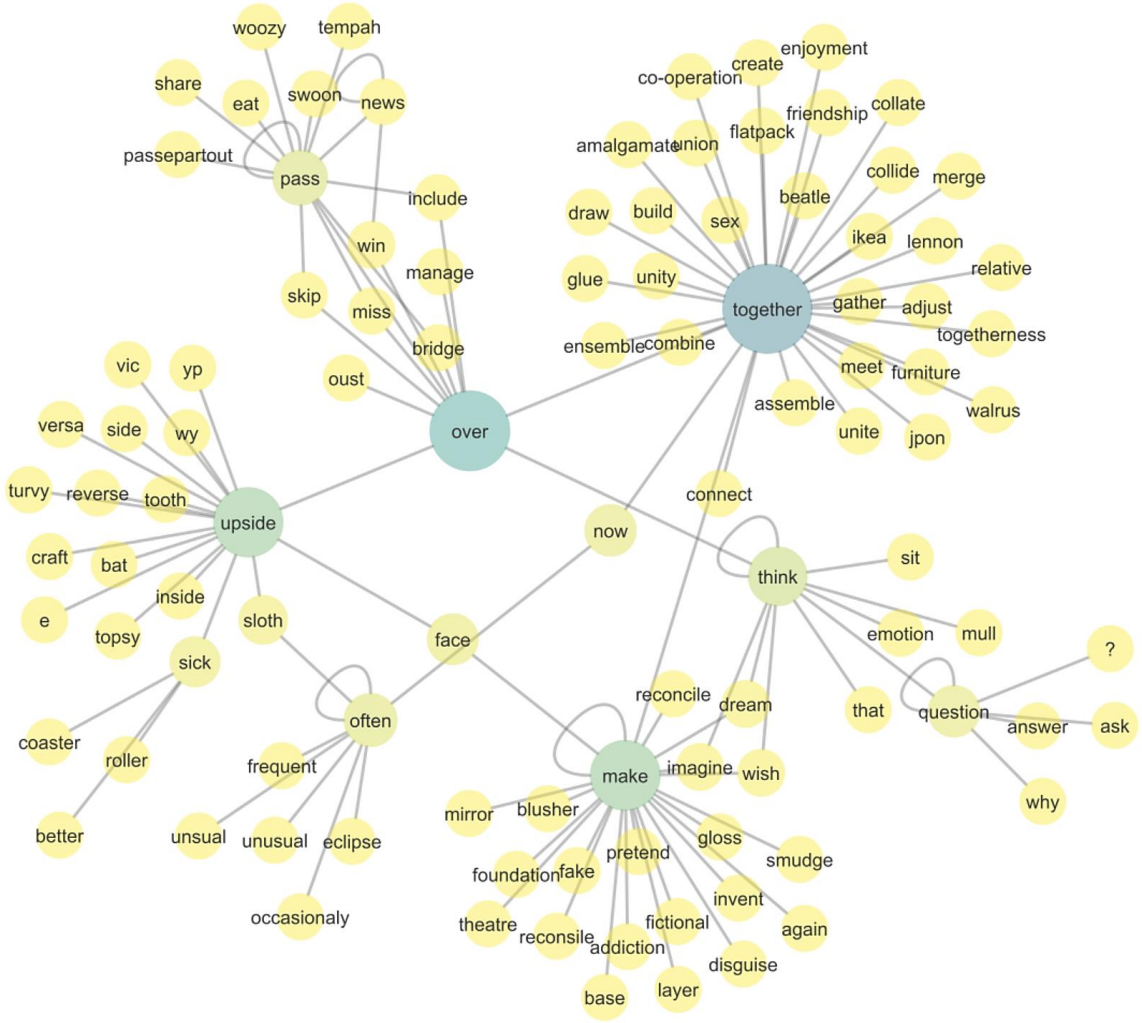


Figure 45: Make/Preposition SWOW-UK cluster.

The last component of the cluster analysis concerns itself with categorising key collocations and word associations making up these clusters into the different types of collocation identified in Chapter 2.4.2. Practically speaking only a subset of the strongest edges could be analysed and the top 100 collocations/associations with the strongest AM values or associative weight and an edge betweenness of at least 100 have been considered. These are taken to represent the strongest links holding together key clusters. The analysis has been carried out with the classification into spectra of collocation such as grammatical <> lexical, paradigmatic <> syntagmatic, etc., as presented in Chapter 2.4.2 in mind. Since only a subsample has been analysed, no classification as to the symmetric/asymmetric nature of the edges has been made. Predictability could further not be

coded for since this only applies to a subset of collocations and is strongly context-dependent. An analysis of predictability without an in-depth analysis of all concordance lines or associative contexts is not viable within the constraints of this thesis. Distinctions into grammatical <> lexical, paradigmatic <> syntagmatic, and more nuanced types of connection (compounds, phraseological, hyponymy etc.) have been made.

In the SWOW-UK network, this analysis shows that around 5% of the strongest word associations are grammatical, and 95% are lexical in nature. Across the Paradigmaticity spectrum, nearly three quarters of the analysed associations (72%) are paradigmatic, and 27% syntagmatic. This presents a key difference which sets the highly weighted word associations apart from strong collocations; these show the inverse pattern. In terms of subtypes of association, SWOW-UK further shows a strong emphasis on encyclopaedic connections such as *moon|wolf* and *tea|boston*. Synonyms, e.g. *next|beside* or *officer|cop* also occur more frequently than in the collocation-based clusters at hand. Further strongly represented categories are compound nouns, hyponyms and named entities. It is of particular interest to note that there are no categories represented in SWOW-UK which are not also represented in collocation-based clusters.

A look at log Dice LL χ^2 shows that the strongest collocations in this subset are markedly more grammatical than the ones presented in SWOW-UK with 34% grammatical collocations and 61% lexical collocations. Further, as previously alluded to, log Dice LL χ^2 shows the inverse behaviour on the paradigmaticity when compared to SWOW-UK with 80% syntagmatic collocations and 15% paradigmatic collocations. Looking at the subtypes of collocation, log Dice LL χ^2 is dominated by compound nouns (35%) and is the only subcluster that contains fixed expressions such as *fair|enough* and *never|before* (19%). Apart from these types of collocation, Verb/Object structures (e.g. *door|open*), and phrasal verbs (e.g. *draw|attention*) are also strongly represented in log Dice LL χ^2 clusters at 9% and 6% respectively. Numerical collocations are also unique to log Dice LL χ^2 and make up 2% of key collocations in these clusters.

Lastly, the strongest collocations found in log Dice clusters almost as strong a focus on lexical collocations as SWOW-UK (94%) alongside the lowest share of grammatical collocations (3%), thus markedly different from log Dice LL χ^2 which exhibits a stronger grammatical focus. In contrast to this, log Dice shows more similarity to log Dice LL χ^2 on the syntagmatic/paradigmatic spectrum with 70% of the analysed collocations being syntagmatic and 27% paradigmatic.

Figure 46 contains a contrastive overview of collocation type distribution. Looking at the more nuanced types of collocations presented in the log Dice clusters it becomes evident that this AM also identifies a large number of compound nouns (25.5%) as collocations and results in more

dominant collocations representing named entities (14.5%, e.g. *atlanta*|*georgia* or *tdi*|*quattro*) than all other investigated networks. This is often paired with a focus on financial terminology with a strong emphasis on company names and products. Beyond this, collocations which contain abbreviations and are emerging from lists more generally such as *taurus*|*apr*, *aquarius*|*jan* etc. are common in key clusters in the log Dice network and neither list-based nor abbreviated collocations are represented in the other two networks analysed here. In contrast to log Dice LL χ^2 and in line with SWOW-UK, the analysed log Dice collocations do contain encyclopaedic collocations (e.g. *ant*|*dec*, a collocation representing the British TV presenter duo Ant & Dec), collocations based on a shared stem (e.g. *auditing*|*auditors*), and collocations relating to place names such as *atlanta*|*georgia* and *papua*|*guinea*.

An observation of the visual properties of the networks shows differing levels of centralisation around high BC nodes. The log Dice clusters, and to a lesser extent also the log Dice LL χ^2 clusters are often chained, whereas SWOW-UK clusters are generally more radial as can be seen in Figure 47.

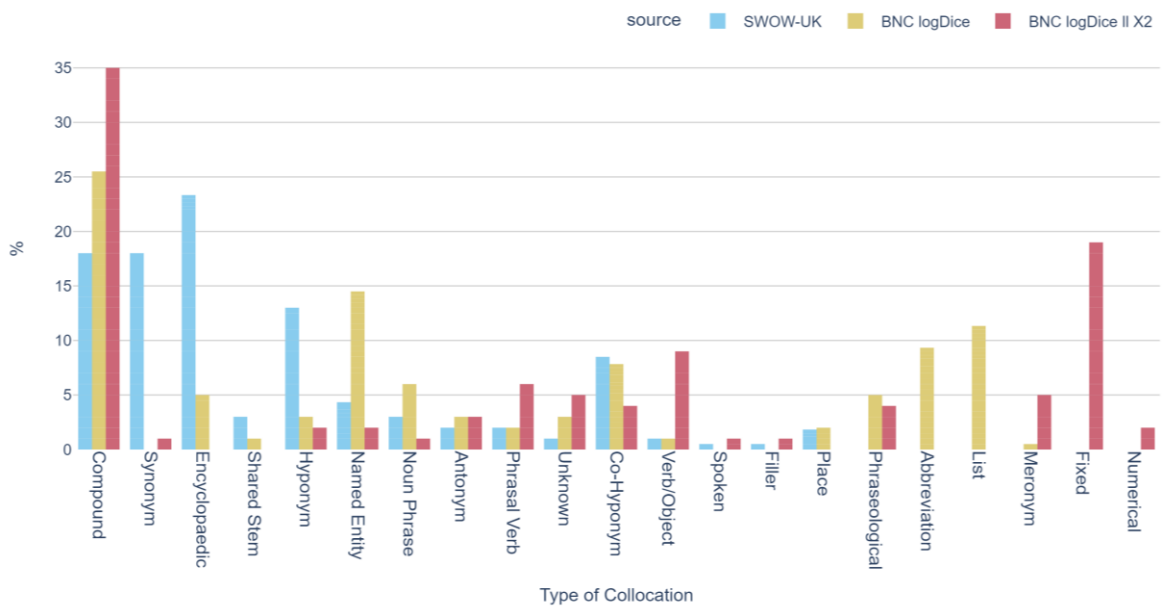


Figure 46: Collocation type distribution for highest edge betweenness associations/collocations from the SWOW-UK, log Dice LL χ^2, and log Dice clusters. See Appendix B for comprehensive tables showing the respective clusters and their manual classification into the respective types.

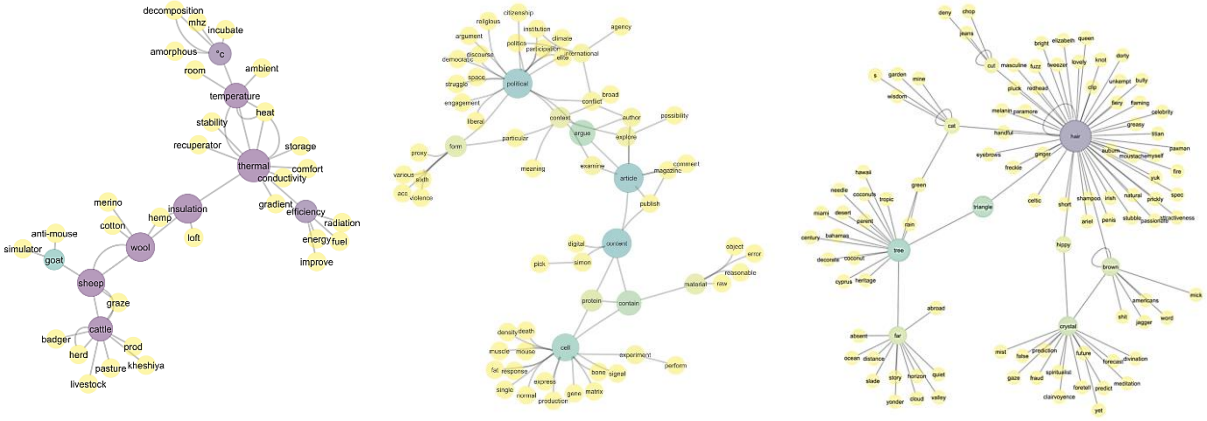


Figure 47: Differences in typical cluster shape. Generally most chained: log Dice (left), both chained and radial: log Dice LL χ^2 (centre), most radial: SWOW-UK (right).

4.3.4 Micro-Level Empirical Evaluation of Structural Comparisons

The third and last subsection of the Results chapter concerns itself with an empirical evaluation of network comparisons on the micro level. This last analytical step. Allows for an explanation of words with special functions in the respective networks such as words with a high betweenness centrality, a high eigenvector centrality, or a high degree centrality.

These concepts are explained in great detail in Chapter 2.7.1.1, their function and linguistic relevancy are also briefly reiterated in this section. The first graph theoretical property investigated here is degree centrality, which is measured by combining the in-degree and out-degree of a node (H. Chen et al., 2018, p. 8). Given the seventeen networks available for exploration in this thesis it is of high linguistic interest to examine which words are overall the most interconnected in each of them since this measurement can be taken as an indication for associative richness (in the case of cue-responses) or collocational versatility. In previous literature, degree centrality, specifically a high out degree, has been shown to have a facilitation effect on semantic processing (Pexman et al., 2003). More generally, degree centrality has been described to influence mental processing (Nelson et al., 1987) and serve as the basis for preferential attachment models (Mak & Twitchell, 2020, p. 1067). Aside from this, high betweenness centrality words are also analysed across all networks. BC measures how strongly interactions between other words depend on the word in question, thus identifying cluster-central 'long-range nodes' (Bordag, 2003, p. 330) or hubs (Veremyev et al., 2019, p. 5) which are crucial for semantic salience in collostructional analyses (Dekalo & Hampe, 2017, p. 165). These nodes act as shortcuts between clusters, often represented by common verbs, articles, and function words (Bordag, 2003, p. 330), and are key in connecting different contexts within a corpus or psycholinguistic network. Betweenness centrality differs from

closeness centrality in that it focuses on how immediately important a word is to the remaining network, whereas closeness centrality evaluates the importance of a word for easing information spread through the network. Since the existence of a shortest path cannot be interpreted as true and proportional information flow for modelling mental processes, closeness centrality is not examined in this Chapter. Instead, the two remaining parameters of interest are eigenvector centrality and the clustering coefficient. Eigenvector centrality (Bonacich, 1972; Pradhan et al., 2020) also measures the influence of a word in the network, but the score is assigned by weighing connections to other high-scoring nodes which contribute more highly than connections to low-scoring nodes. The long-term collocational influence of a word is expressed via its EC value, which means that it is also closely interlinked with the concept of preferential attachment (Castro & Siew, 2020, p. 16; Mak & Twitchell, 2020, p. 1059; Sheridan & Onodera, 2018, p. 1). Finally, high ranking clustering coefficient values are obtained for all networks. In simple terms, the clustering coefficient (ClCoef) indicates the probability of words being connected to their collocates or associative neighbours (Borge-Holthoefer & Arenas, 2010, p. 1268; Steyvers & Tenenbaum, 2005, p. 46; Watts & Strogatz, 1998, p. 441). In linguistic data, a high ClCoef denotes strongly clustered collocates in a tight-knit context, while a low ClCoef suggests distinct contextual embeddings. Unlike in the Meso-Analysis Chapter, this chapter presents results from all sub-networks. This is possible since the manual analysis of the 25 top scoring terms and the overlap between the top 200 words in each network is far less time-consuming and bound by spatial constraints than full cluster analyses.

For each of the properties examined here, the general analytical approach is the same. On the one hand, the top 25 highest scoring linguistic items are qualitatively examined in order to assess similarities and differences between the types of words found in these lists and their possible functions. Then, matrices similar to the ones presented in Chapter 4.3.2, are examined. These are based on the overlap found in the top 200 highest scoring items for each property in each of the seventeen networks. This allows for a broader examination of similarities and differences, and for discussions regarding which association measures, albeit using markedly different theoretical approaches, lead to overall similar results. Tables containing the top 25 items for each network are not printed here due to spatial constraints but can be accessed in Appendix C and more extensive lists can be generated using the interactive code provided alongside this thesis. The same is true for all other micro-level analyses.

The first property to be explored is eigenvector centrality. An initial look at the 25 highest scoring EC items in SWOW-UK and SWOW-EN shows that the networks share a large number of the top EC items: Fifteen of them are identical, examples for these are *go*, *up*, *child,* and *food*. Negation markers (*not*, *no*) are contained exclusively in the top 25 SWOW-UK EC items, which could link

back to the UK subset containing a higher share of definitions via antonym/negation as mentioned in Chapter 4.2.2.4 (Example 3). When examining all other networks qualitatively, five groups of collocation networks can be distinguished between on the basis of their EC behaviour. The first group is EC values from the $\Delta P_{backward}$ and $\Delta P_{forward}$ networks. These containing large number of named entities, foreign terms and low frequency items such as *trinkaus* or *chaarat*. The second group is made up of high EC values from combinations of AMs which contain $\Delta P_{forward}$. The items observed in these lists consist almost exclusively of individual letters and abbreviations (*i*, *e*, *sch*, *acc*). Two other groups contain only one AM each, Poisson and OddsRatio. Both of these groups are largely unique. Poisson contains mostly lexical items and is populated with verbs and adjectives such as *early*, *carry*, and *major*, whereas OddsRatio presents a strong focus on items resulting from lists, especially recipes with *tbsp*, *tsp*, *garlic*, and *chilli* being high EC items. The last group contains all networks based on χ^2, log Dice, LL, and all combinations thereof. For all of these lists a heavy grammatical focus and a focus on high frequency items such as *i*, *be*, *you*, *of* is observed.

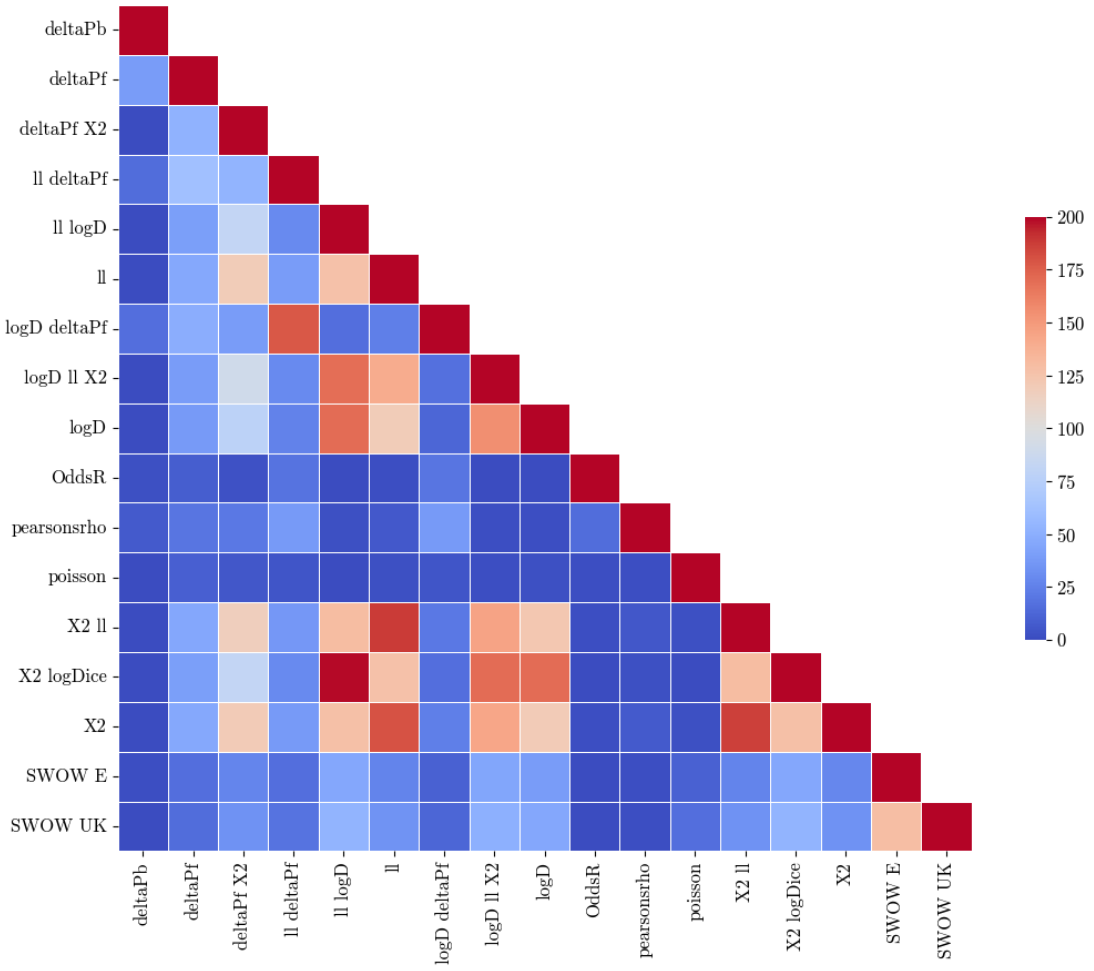


Figure 48: Heatmap showing the number of top 200 EC words shared between each of the seventeen analysed networks.

This more extensive heat map depicted in Figure 48 shows the total number of the top 200 EC words which are shared between each of the analysed networks. Red values denote a higher number of shared high EC items between the networks, blue values denote less overlap. This corroborates the general grouping described above Especially when focusing on the networks based on χ^2, log Dice, LL, and combinations thereof. The highest overlap is observed between χ^2 log Dice and LL log Dice at 199/200. SWOW-UK and SWOW-EN share 129 linguistic items, and the closest fit for SWOW-UK occurs in the χ^2 log Dice and LL log Dice networks with 53 identical terms each.

The second property to be explored is betweenness centrality. An initial look at the 25 highest scoring BC items in SWOW-UK and SWOW-EN shows that the networks share a slightly lower number of the top BC items when compared to top EC items: Thirteen of them are identical, examples for these are *red*, *white*, *good*, and *bad*. The non-shared items found in SWOW-UK tend to be prepositions or conjunctions (*in*, *of*, *and*) whereas the non-shared items found in SWOW-EN tend to be nouns (*car*, *man*, *health*).

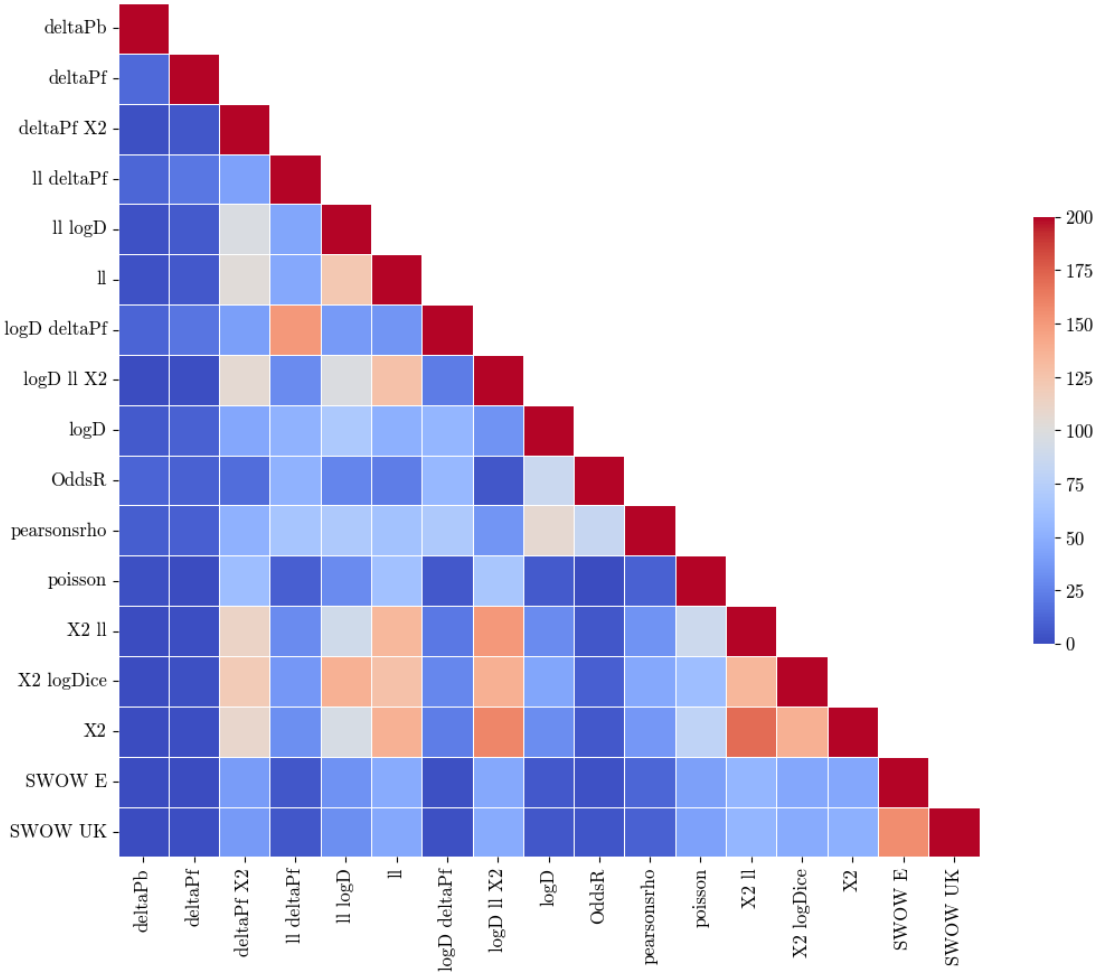


Figure 49: Heatmap showing the number of top 200 BC words shared between each of the seventeen analysed networks.

When examining these qualitatively, five (albeit different) groups of collocation networks can be distinguished between on the basis of their BC behaviour. The first group consists of BC values from the χ^2 log Dice, χ^2 LL, LL log Dice, log Dice LL χ^2, and LL networks (the same as EC group five except for log Dice itself). These networks contain a mix of financial and grammatical terms such as *share*, *financial*, *the*, *of*, and *i* in their top 25 BC words. The second group is made up of high BC values from OddsRatio, r_φ, LL $\Delta P_{forward}$, log Dice $\Delta P_{forward}$, and log Dice networks. These lists are abbreviation-heavy and also contain financial terminology but reference individual companies more frequently than group one. The third group, identical to EC group one, contains $\Delta P_{forward}$ and $\Delta P_{backward}$ networks which are largely unique. Their top BC values contain proper names, rare spellings, and foreign terms. The last two groups, again, form outliers that contain only one AM each: Poisson and $\Delta P_{forward}$ χ^2. Poisson is mostly unique, contains no reference to financial terms and is focused around adverbs and conjunctions such as *however*, *fortunately*, and *well*. $\Delta P_{forward}$ χ^2 also references financial terms but also contains high-frequency verbs like *do*, *say*, and *know*. As becomes apparent when examining Figure 49, the overall similarity between BC centrality items is lower than the overall similarity between high EC items. The heat map further substantiates the claims that the groups are overall rather similar, with the marked exception of log Dice not behaving in line with χ^2 log Dice, χ^2 LL, LL log Dice, log Dice LL χ^2, and LL as it did in the EC comparison. The number of shared high BC words between SWOW-UK and SWOW lies at 156, the next best collocation-based candidate for approximating SWOW-UK is χ^2 LL at 54 shared high BC items. χ^2 log Dice and LL log Dice which performed well when comparing EC values contain 47 and 32 shared words respectively.

The last centrality measure discussed here is degree centrality. An initial look at the 25 highest scoring DC items in SWOW-UK and SWOW-EN shows that the networks share only eight of their top DC items, which is significantly fewer than both BC and EC. Examples of shared items include *food*, *black*, *good*, and *hot*. The non-shared items in SWOW-UK tend to be prepositions or verbs (*off*, *of*, *go*), whereas the non-shared items in SWOW-EN are more often adjectives and nouns (*friend*, *love*, *america*). When examining all other networks qualitatively, three groups of collocation networks can be distinguished based on their DC behaviour. The first group consists only of the Poisson network, which is an outlier in that it contains many high-frequency grammatical items alongside a high share of items unique to this network such as *time*, *make*, and *there*. The second group includes high DC items from the OddsRatio, r_φ, and log Dice $\Delta P_{forward}$ networks which are characterised by single letters, abbreviations, and list items like *v*, *et*, and *de*. The last group encompasses all other networks, which mostly feature high-frequency grammatical items and financial terms and, unlike Poisson, do not contain a large share of nouns and verbs.

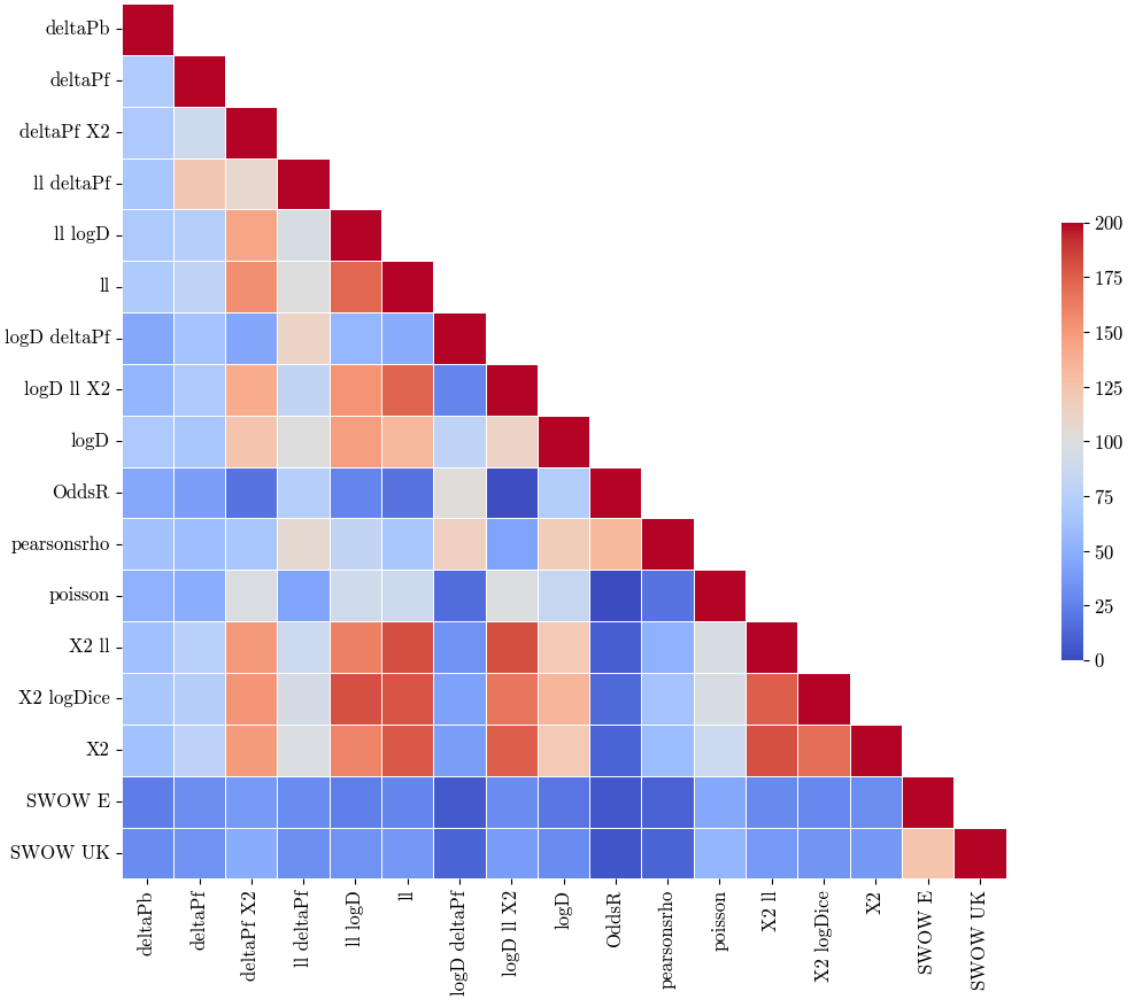


Figure 50: Heatmap showing the number of top 200 DC words shared between each of the seventeen analysed networks.

Looking at the heatmap depicting the similarities between the top 200 DC items, it becomes obvious that this centrality measure overall shows the most overlap between all networks. It also underlines the finding that OddsRatio, r_{φ}, and log Dice $\Delta P_{forward}$ are outliers which share very few items with any other networks. When looking at similarities compared to SWOW-UK, SWOW-EN is, again, the best match with 125/200 shared top DC items. The closest collocation-based match is, as was the case for EC, χ^2 log Dice at 54/200, followed by $\Delta P_{forward}$ χ^2 at 48. LL log Dice shares only 33 of the top 200 DC items, and χ^2 LL shares 38. This makes high centrality items from the χ^2 log Dice network overall most similar to SWOW-UK based high centrality items.

Having explored different centrality measures, it is also of interest to examine the clustering coefficient, especially given its prominent role in existing cognitive and linguistic research (Baronchelli et al., 2013; Chen et al., 2018, p. 6; Cong & Liu, 2014; Liu & Li, 2010; Steyvers & Tenenbaum, 2005; Utsumi, 2015).

The exploration of the clustering coefficient paints a completely different picture when compared to the centrality measures explored above. The first notable feature is that there is much less overlap between the top 25 (as well as the top 200) high ClCoef items each of the networks. The only significant overlap is measured between log Dice and r_{φ}, OddsRatio and r_{φ}, and LL $\Delta P_{forward}$ and log Dice $\Delta P_{forward}$. Generally speaking, high clustering coefficient items are far more lexical than high centrality items but still display a preference for abbreviations and single letters. While the results for clustering coefficient are presented alongside the centrality measures for reasons of comparability, A major limitation of this approach is that a large number of nodes in the networks possess a maximal clustering coefficient of 1 making it impossible to sort them meaningfully. For most networks, over 200 items exhibit this maximal score. This means that only their general nature can be examined, but the exact ordering of the items in these lists has to be dismissed. This might also skew the heat map displayed in Figure 51.

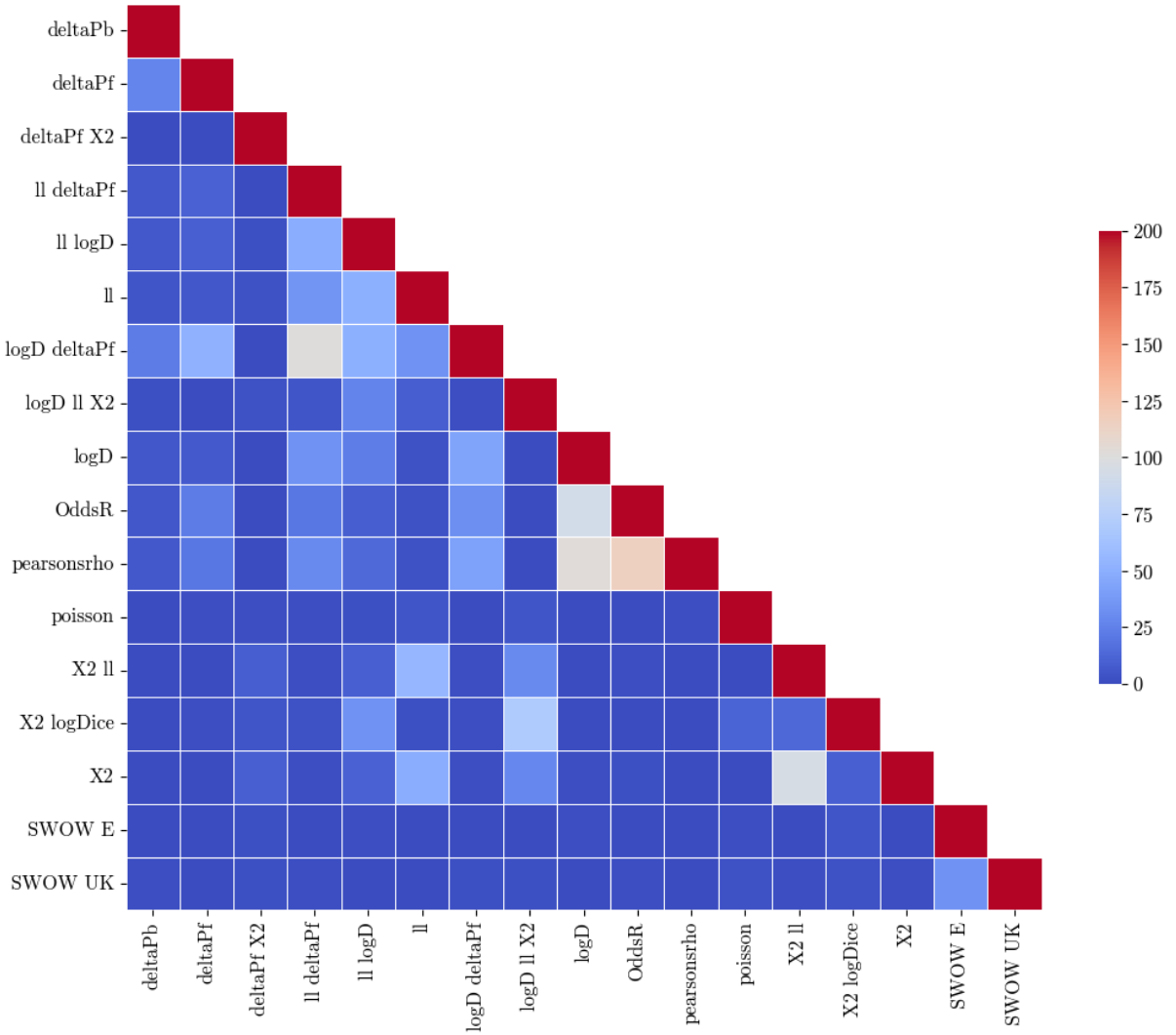


Figure 51: Heatmap showing the number of top 200 ClCoef words shared between each of the seventeen analysed networks.

Bearing this in mind, examples for high clustering coefficient features found in SWOW-UK are *deaf*, *rod*, *nerve*, and *exist*. While most collocation-based high ClCoef items are represent (parts of) named entities such as *hemel*, *hempstead* (OddsRatio, r_φ) or *epirubicin* (log Dice, log Dice $\Delta P_{forward}$, LL $\Delta P_{forward}$, r_φ, OddsRatio), the Poisson subsection is fully unique and exhibits markedly more high frequency items and almost exclusively nouns, verbs, and adjectives such as *surface*, *security*, and *weekend*. Due to the abovementioned limitations no groupings are established for high ClCoef items.

Before concluding the chapter on micro-level results, it is evaluated if words with an exceptionally high frequency of occurrence in the corpus are overrepresented in the top 200 BC/EC/DC/ Clustering Coefficient items as would be expected on the basis of Veremyev et al.'s (2019, p. 3) research. In order to examine this, five categories of word frequencies have been established and words have been classed into these groups on the basis of their frequencies in the BNC 2014. All words occurring above the 99.9th percentile have been classed as extremely high frequency, words less frequent than this but above the 99th percentile have been classed as very high frequency, words between the 90th and 99th percentile have been classed as high frequency, words between the 50th and 90th percentile have been classed as mid frequency, and all others (i.e. all words below the 50th percentile) have been classed as low frequency. Figure 52 displays the frequency distributions for each of the networks and each of the metrics.

This evaluation shows that SWOW-based high centrality items show a similar distribution for all centrality measures with about 10% extremely high frequency words, 38% very high frequency words, 50% high frequency words, and 2% mod frequency words. None of the top 200 centrality words fall into the 50% least frequent words in the BNC 2014 for either SWOW-based network. Looking at the collocation based high-centrality items, OddsRatio, r_φ, and for EC and BC also $\Delta P_{forward}$ and $\Delta P_{backward}$ again prove to be outliers. Across the board they show a tendency to represent lower frequency terms than either other collocation networks or word-association based networks in their top centrality list. The remaining networks with the exception of LL $\Delta P_{forward}$ and log Dice $\Delta P_{forward}$ who present an intermediate distribution show a proclivity to include higher frequency terms. Looking beyond the centrality measures alone, this figure further underlines that clustering coefficient and centrality measures exhibit a fundamentally different behaviour. The rate of extremely high frequency items in the top ClCoef list is much lower, with only χ^2 log Dice containing any.

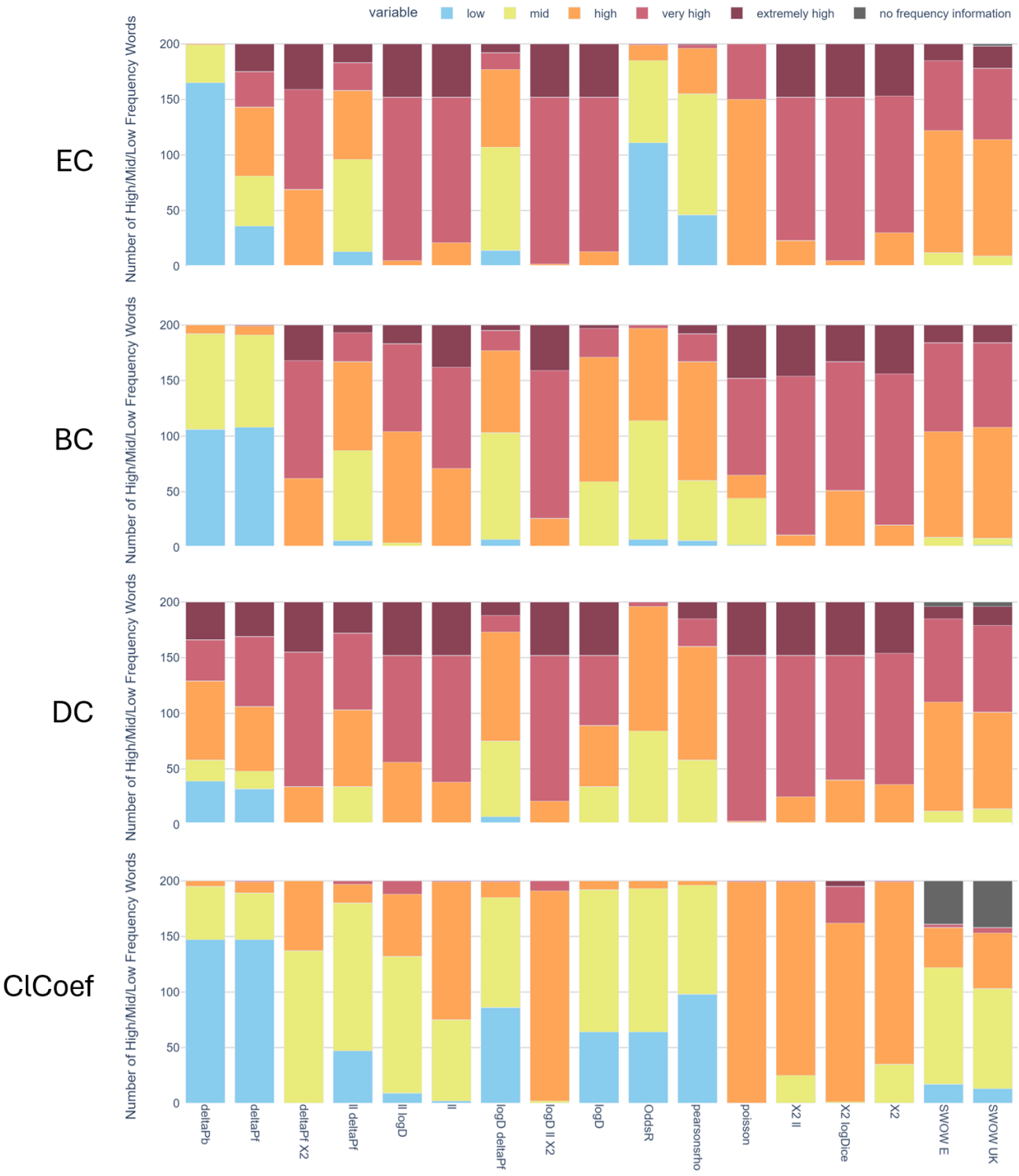


Figure 52: Frequency distribution of words with the 200 highest BC/DC/EC/ClCoef scores across all seventeen networks.

5 Discussion

This Chapter aims to provide comprehensive answers to the three research questions outlined in Chapter 2.9, synthesising the findings with the existing literature. By doing so, it seeks to bridge the gap between the theoretical framework and empirical results. The discussion examines each research question, drawing on the data presented in Chapters 3 and 4 and positioning it within existing research to highlight the contributions and implications of this research. Following the discussion of the research questions, Chapter 5.6 contains a section on limitations. This is crucial as it acknowledges the constraints inherent in the study, providing a balanced view of the findings, and, most importantly, clearly states what steps can be taken in future work in order to control for further factors such as inter-linguistic variation and change over time. Beyond this, this chapter includes a section on the practical applications of the LLN pipeline and general large linguistic network methodology developed as part of this thesis. This is particularly important because the development of a new methodology is only effective if practical applications are presented alongside it. By demonstrating how the network methodology can be applied in real-world scenarios, this section underscores the practical relevance and potential impact of the research.

5.1 Interpretation of Findings relating to RQ1

This very brief Chapter serves as a summary of a much more extensive discussion of the findings from RQ1 in Chapter 3.4 and Chapters 3.2.1-3.2.4. The brevity of this chapter is a consequence of the methodological nature of the thesis and an avoidance of repetition: Without discussing these results in their respective chapters, the motivation for methodological choices in RQs 2 and 3 could not have been provided effectively.

The key points regarding findings from RQ1 are the following: As explored more extensively in Chapter 3.4, the answer to the first research question of how current approaches to AM extraction can be harnessed in a psycholinguistically plausible manner identified the following factors for association measure extraction:

- Choice of Association Metric
- Choice of possible filters
- Choice of collocation window
- Directionality
- Choice of Graph Theoretical Analytics
- Limitations to a single layer of linguistic representation

Distinct recommendations for each of these considerations are made in the respective Chapters 3.2.1-3.2.4. Chapter 3.4 further contains a flowchart (Figure 18) positioning individual AM

approaches within this schema. Harnessing current approaches to AM extraction in a psycholinguistically plausible manner further builds on the general suitability of network approaches for representing language. As shown in Chapter 2.2.1, not only do network approaches allow for linguistic representation without a mandatory focus on an intuition-based starting point[22] for the analysis (Castro & Siew, 2020; Jihua Dong & Buckingham, 2018, p. 120; Sinclair & Coulthard, 1975), but also show robustness when applied to data of poor quality (Veremyev et al. (2019)), and, most importantly, they can account for all connections exhibited by individual across the entire dataset.

5.2 Interpretation of Findings Relating to RQ2

RQ2 asks which AMs lead to collocation networks that best approximate the content and structure of large word association networks. This question has been examined on the aforementioned three levels:

- Micro-level Analysis (Chapter 4.3.4) which entails exploring individual words that fulfil a special function in the network.
- Meso-level Analysis (Chapter 4.3.3) which entails exploring clusters emerging from the network qualitatively and classifying the types of collocations found within the clusters as well as the represented word classes.
- Macro-level Analysis (Chapter 4.3.2) which entails exploring the network holistically in terms of its overall structural properties.

The combination of quantitative and qualitative analysis in all of these levels shows the following results: When comparing collocation networks against word association networks, some are much worse candidates than others. On the macro level, $\Delta P_{backward}$ and $\Delta P_{forward}$ performed worst due to an almost inverted set of network-wide graph theoretical parameters. $\Delta P_{forward}$ and $\Delta P_{backward}$ contain far more nodes than SWOW-UK, but their number of edges in relation is much smaller. These networks further possess a very low average clustering coefficient and mean eigenvector centrality of near zero, while these values for SWOW-UK are among the highest observed across all networks at 0.26 and 0.76 respectively. A look at the macro-level graph theoretical parameters for Poisson and log Dice $\Delta P_{forward}$ shows similarly bad results. Since the examination of these metrics does not result in a clear answer as to which of the remaining AMs is optimal for approximating SWOW-UK, further investigations have been carried out. NetSimile (Berlingerio et al., 2012) results show that the best performance of a collocation network, i.e., the smallest distance to the SWOW-UK

[22] This is the case since alternative approaches e.g. regarding collocation extraction commonly necessitate determining a term or terms of interest used by the researcher to initialise the analysis.

network, is measured for χ^2 log Dice, χ^2, and log Dice LL χ^2. Complementary results obtained from an Adjacency Spectral Distance analysis (Wilson and Zhu (2008)) paint an overall similar picture and show log Dice LL χ^2, χ^2 log Dice, and χ^2 LL as the best candidates for approximating SWOW-UK. Balancing these largely quantitative analyses with a more qualitative approach, the networks have also been compared in terms of the number of top 100/500/1000 strongest collocations they share with the respective strongest word associations. The best performance here is recorded for log Dice LL χ^2, χ^2, and LL. Summing up the findings from the macro level analysis, the best performing AMs in terms of approximation of word associations are **log Dice LL χ^2.** In other words, the best performing metric extracts pairs of words which display an overall high mutual exclusivity (log Dice), where the relative proportion of collocates following the node and not following the node are unexpectedly different (χ^2), and where the observed and expected frequencies of the collocation and all other combinations of its collocates are unexpectedly different (LL). This suggests that Statistical Learning may particularly depend contingency learning (as expressed through χ^2 as attested in previous literature by Peterson and Beach (1967, p. 42) and LL), and mutual exclusivity as attested in previous literature by Tribushinina and Gillis (2017) and Gries (2012, p. 49) and expressed through log Dice.

Looking at the meso-level and thus the cluster analysis, three networks were selected for analysis for spatial reasons: SWOW-UK as the baseline, log Dice LL χ^2 as the best macro-level performer, and log Dice as a status quo and representation of common practices in corpus linguistics. This section also presents the most involved qualitative analysis, which considers the cluster topic as well as word class distribution within clusters, and the different types of collocation making up their edges. In terms of topical similarity, log Dice LL χ^2 performs better than log Dice. A particular issue found with log Dice clusters that log Dice LL χ^2 does not suffer from to the same extent is the overrepresentation of items from lists and terms indicative of a highly specialised register terms. All three networks showed similar shares of grammatical items represented in the clusters. Neither of the two analysed sets of clusters approximate SWOW-UK well in terms of the word classes represented within closed class items. This indicates that filtering for forms such as numbers could improve the fit between collocation networks and word association networks. Similarly, in terms of type of collocation, no great similarity between the collocation-based clusters and the word association clusters could be identified. The major difference observed here is that encyclopaedic knowledge as well as hyponyms and synonyms are more strongly represented in the word association network. This is in line with previous research into semantic maps and their structures (Chapter 2.5.2), especially Gravino et al. (2012) who attest a high frequency for synonymy, hypernymy and hyponymy in word association datasets. Beyond this, the research carried out in

this thesis provides pointers for future semantic relation database efforts since the most frequently encountered relationship between words in the word association clusters is 'encyclopaedic'. This does not neatly map onto the usual WordNet categories which are limited to 'keywords', synonymy, hyper/hyponymy, antonymy, similarity more widely, membership/part/material, cause, participle, attributive, verbal or derivational, or domain relationships (Trustees of Princeton University, 2024). An example for this is the encyclopaedic relationship between *crystal* and *future* as found in the SWOW-UK key cluster. The WordNet domain for *crystal*[23] contains a list of domains with varying levels of abstraction (e.g. *jewellery* (very concrete), and *physics* (very abstract)) and relevancy (*e.g. skiing, humanities*). It does not contain a mention of the concept of *future* found in the SWOW-UK clusters. While the 'is domain' attribute, especially the 'is domain (use)'-category comes closest to capturing encyclopaedic knowledge in WordNet, a strong argument for introducing a higher resolution for the nature of 'usage domain' and other representations of encyclopaedic knowledge would be strongly encouraged given the high frequency of encyclopaedic key word associations.

Lastly, the micro-level results are presented. The first area of interest are long-range nodes and words which carry out special functions within the network. For this purpose, the overlap between high centrality terms across all networks has been calculated for eigenvector centrality, betweenness centrality, and degree centrality. The results show that χ^2 log Dice and LL log Dice show the most overlap with SWOW-UK in terms of eigenvector centrality, χ^2 LL shows the highest overlap in terms of betweenness centrality, and χ^2 log Dice, again, shows the highest overlap in terms of degree centrality. This makes high centrality items from the χ^2 **log Dice** network overall most similar to SWOW-UK based high centrality items. When analysing the frequency categories of the words with the 200 highest centrality scores across all seventeen networks, the best distribution of matches is observed for networks based on χ^2 or LL and combinations thereof. When observing the results from the clustering coefficient, no collocation-based network provides a good approximation. As described in further detail in Chapter 4.3.4, this may be due to more than 200 items having a perfect score from any networks which makes this measure less reliable.

Based on the results observed in this thesis, the question of which AMs lead to collocation networks that best approximate the content and structure of large word association networks can be answered as follows: Generally speaking, a good approximation can be provided by combinations of association measures. In this thesis, particularly the combination of log Dice, LL, and χ^2 showed promising results. Measures such as OddsRatio, r_φ, and log Dice $\Delta P_{forward}$, alongside $\Delta P_{backward}$ and $\Delta P_{forward}$ on their own, showed a bad performance across the board. It remains to be

[23] Full domain list: *geology, jewellery, chemistry, geometry, oceanography, gastronomy, meteorology, physics, applied_science, color, skiing, photography, optics, electronic, electricity, electrotechnology, radio, metrology, humanities, occultism*

said that none of the association measure-based networks succeeded in generating qualitatively similar clusters. The combination of different AMs for applied research has been called for (Garcia et al., 2019, p. 57; Gries, 2022b, p. 14, 2022b, p. 30; Pecina, 2010, p. 153) and, less frequently, also been implemented (e.g. Seretan (2011) and Gabrielatos and Baker (2008)) in previous literature. The findings presented in this thesis underscore the importance of using a combination of association measures to achieve the best approximation of word association networks.

5.3 Interpretation of Findings relating to RQ3

Finally, RQ3, the question if there is a general structural difference between usage-based network patterns and patterns found in word association networks is answered alongside its key subcomponents as specified in Chapter 2.9

. The answers to this question are grounded in the data explored in RQ2 which establishes which collocation-based network approximates a word-association-based network the best, and focuses on general trends observed among *all* collocation-based networks.

The first sub-question to be answered is whether the networks are intuitively interpretable. This is the case, though only a full analysis of their properties, especially on the meso-level as explored in Chapter 4.3.3, allows for a holistic interpretation. Chapter 4.3.2 further answers the next sub-question regarding the macro-level properties of the resulting networks. What has been observed among all collocation-based networks as opposed to the word-association ones is the following: Firstly, word-association networks exhibit a low rate of self-loops when compared to collocation networks. Secondly, word-association networks exhibit a high average clustering coefficient paired with a high mean eigenvector centrality. This signifies that word association networks contain many connections that follow the preferential attachment model (Castro & Siew, 2020, p. 16; Mak & Twitchell, 2020, p. 1059; Sheridan & Onodera, 2018, p. 1), and, at the same time, strongly clustered words in a tight-knit context. This quantitative finding is further corroborated by the visually radial nature of word association clusters presented in Figure 47 as opposed to more chained collocation-based clusters. Another general pattern observed for word association networks is a node to edge ratio of roughly 1:3.7. All collocation networks with an otherwise similar network property profile exhibit a much higher node to edge ratio, e.g. log Dice LL χ^2 with a ratio of 1:14.5. To sum up, a general observation that sets word associations apart from collocations is that word associations are contextually rich, and heavily structured in a way that connects high-degree nodes to other high-degree nodes despite a relatively low node to edge ratio. Collocation-based networks, on the other hand, generally vary strongly in their structural makeup. One type (log Dice, LL log Dice, r_φ, OddsRatio, log Dice $\Delta P_{forward}$, $\Delta P_{forward}$, $\Delta P_{backward}$, LL, χ^2 log Dice) results in fractured networks

consisting of a large number of un-connected sub-networks despite a reasonably high number of nodes and edges which also exhibit a low overall density and low mean eigenvector centrality. Another type, primarily consisting of combined AMs with a χ^2 component (χ^2 LL, log Dice LL χ^2, χ^2, Poisson, and $\Delta P_{forward}$ χ^2) generally results in networks more similar to word association ones with a higher mean eigenvector centrality, but, as explored above, a disproportional node to edge ratio.

The next sub-question concerns itself with the extent of the percentage overlap between the strongest unweighted nodes and edges in the different collocation networks and the word association network. While statements such as "It is shown that basic language processes such as the production of free word associations and the generation of synonyms can be simulated using statistical models that analyse the distribution of words in large text corpora." (Rapp, 2002) have been made over 20 years ago, suggesting that linguistic feature extraction for simulating word association is a solved problem, results from the present thesis show that this is very much not the case. Findings from the network-based comparisons of word associations and collocations in this thesis show a percentage overlap of no more than 4.4% even for the best scoring collocation metric out of 15 different collocation metrics applied to a large-scale high-quality dataset, the BNC 2014, demonstrate clearly that there is still research to be done into approximating word association via "analyzing the distribution of words in large text corpora". The results from this thesis further show that collocations should not be employed as a single proxy for word association. This is especially relevant e.g. in the context of corpus-assisted discourse analysis where collocation methods are routinely employed used with the ultimate aim of generalising to stance and attitude (S. Chen, 2013; Galasinski & Marley, 1998). Results from this thesis therefore make a strong case for thoroughly evaluating how directly repeated textual co-occurrence influence listeners'/readers' mental processing via methods other than collocational analyses alone. These differences can be partially explained by the modality specific nature of SL (Sandoval et al., 2017, pp. 10–11). This makes a strong case for creating and analysing multimodal corpora to tackle the lack of auditory/visual collocational input.

The remaining four sub-questions fall into the category of more qualitative comparisons on the cluster and word level. Firstly, the degree of similarity between topics represented in word association-based clusters and collocation clusters is to be discussed. The clusters analysed here can be seen as *semantic spaces* (Deerwester et al., 1990, p. 391) a notion which also serves as the basis for latent semantic analysis (LSA). When looking at the three different groups of clusters analysed in detail for this thesis, log Dice log Dice LL χ^2, and SWOW-UK, the following observations can be made: None of the observed clusters across different data sources exhibit similarities when

assessing whether the same nodes are both present in the clusters and connected to a high number of identical edges. There is, however, a certain amount[24] of overlap between topics represented in clusters from all three networks.

Looking at systematic differences, the corpus-based clusters strongly feature topics surrounding heavily specialised terms from distinct genres/registers such as the automotive domain and legal matters (log Dice), sport, or financial affairs (log Dice and log Dice LL χ^2). This is in line with previous literature where genre effects have been found to significantly interfere with a network analysis originally carried out as part of a study investigating language typology (Liu & Li, 2010, p. 3461). While this is a limitation of both Liu & Li and the present thesis, the same genre-sensitivity can be harnessed to refine genre detection and improve explainable document summarisation.

In contrast to this, word association-based clusters feature actions such as showing or taking much more prominently than collocation-based clusters. Secondly, the question of which words display a particularly high linguistic availability as indicated by their high degree in the respective networks is raised. Looking at lists of high degree centrality items (Appendix C) which represent words with the highest number of outgoing and incoming connections (be that associations or collocations) leads to several observations.

Firstly, words with a high associative availability fall into the following groups (all taken due to their presence in the top 25 highest degree centrality items:

1. Prepositions (*of*, *off*, *out*, *down*, *up*, *in*, *away*, *to*, *on,* and *back*)
2. Terms expressing evaluation/judgement (*bad*, *good*, *well, old, hot*)
3. Action verbs (*go*, *do*, *show*, *take*)
4. Colour terms (*black*, *red*, *white*)
5. Food (*food*)
6. Body (*hair*)
7. Negation (*not*)

In comparison to this, as explored more extensively in Chapter 4.3.4, words with a high associative availability fall into three groups, one of which comprises primarily abbreviations and individual letters due to an overrepresentation of list items. The following groups of words are prevalent in at least five of the remaining lists (duplicates marked in bold):

1. **Prepositions** (***of***, ***to***, ***in***, ***on***, *for)*
2. Determiners (*the*, *a*)
3. Personal Pronouns (*I*, *she*, *he, it, you ,his, her, that*)
4. **Negation** (***not***)
5. **Action verbs** (***do***)

[24] The extent of overlap and similarities in particular are further qualified and discussed in Chapter 4.3.3, especially Figure 43.

6. Stative Verbs (*be, have*)
7. Conjunctions (*and*)
8. Finance (*financial*, *share*)
9. Possessive marker (*'s*)
10. Ambiguous (*like*[25])

This shows that a fundamental difference between highly connected terms in word associations versus collocations is a stronger focus on pronouns, determiners, and descriptive terms as well as specialised terms (*financial*, *share*) than word associations which are, again, more action focused, embodied (*hair*, *food*), and evaluative.

Extending this question to all centrality measures observed in this thesis, eigenvector centrality, degree centrality, and betweenness centrality, the following candidate words for kernel lexicon membership emerge from the **word-association**-based high centrality lists[26]:

1. **Prepositions** (*up*)
2. Terms expressing evaluation/judgement (*bad*, *good*, *old*)
3. Colour terms (*red, black*)
4. Food (*food*)
5. Ambiguous (*work*)

Using the same approach, the items extracted from the **collocation**-based high centrality list[27] fall into the following groups:

1. **Prepositions** (*to, of, in)*
2. Determiners (*the*, *a*)
3. Personal Pronouns (*I*, *you*, *she*, *it*, *he*)
4. Stative Verbs (*be*)

The difference here is largely similar to the one observed for degree centrality only, with the exception of more prepositions dominating the collocation-based list. It is of interest to note that terms such as food and work, in their role as frequent abstract nouns, can be seen as representative of common, nearly universal, experiences and concepts that, in terms of their Signified in a semiotic context, strongly vary. While this is not the case for stable high centrality components of collocations, they do contain a large number of personal pronouns which are, again, represent a large variety of different Signified individuals and entities. The findings presented in this section are broadly in line with Bordag's (2003, p. 330) observation that common verbs, articles and

[25] Like has been classified as ambiguous since the collocational profile does not allow for establishing the function of this versatile term (e.g. *I like her* vs. *I am like her* etc.)
[26] The requirement for being included in this list is appearing in at least five of the six examined top 25 centrality item lists (BC/DC/EC for SWOW-EN and SWOW-UK respectively).
[27] The requirement for being included in this list is appearing in at least twenty of the 45 examined top 25 centrality item lists (BC/DC/EC for all fifteen collocation-based networks respectively).

function words are the most common word classes of long-range nodes, but expands that list to also contain personal pronouns, evaluative terms, and colours.

Finally, the last remaining sub-question asks if there are qualitative differences between word-association and collocation network-central hubs in terms of the high/low frequency of represented words expected on the basis of previous research such as Veremyev et al. (2019, p. 3). This question is doubly important since it also has the potential to inform recommendations for good practice in corpus linguistics. The results aiming to answer this question are obtained in Chapter 4.3.4 and presented in Figure 52 which depicts the individual frequency distributions of words with the 200 highest BC/DC/EC scores across all seventeen networks. This data shows that word association networks follow a largely stable distribution containing about 10% extremely high frequency (top 0.01% most frequent in the BNC 2014) terms, about 38% other very high frequency words (top 1%), 50% high frequency words (top 10%), and 2% mid frequency words (top 50%). Collocation-based network-central hubs on the other hand either fall into the category of across the board disproportionally lower-frequency terms (OddsRatio, r_{φ}, $\Delta P_{backward}$, $\Delta P_{forward}$, LL $\Delta P_{forward}$, log Dice $\Delta P_{forward}$), and broadly proportional terms ($\Delta P_{forward}$ χ^2, LL log Dice, LL, log Dice LL χ^2, log Dice, Poisson, χ^2 LL, χ^2 log Dice, and χ^2). Out of the group with comparable frequency patterns, the best approximation for EC is reached by $\Delta P_{forward}$ χ^2, the best approximation for BC is reached by LL log Dice, and overall no good approximation for DC is reached, but $\Delta P_{forward}$ χ^2, again, is the most comparable. These findings can also contribute to a methodological debate surrounding recent papers such as Gries (2022a), Gries (2022b), and Gries & Durrant (2020) questioning the strong ties of many commonly used AMs to raw frequency over pure association. While it is certainly desirable to introduce concise terminology for referring to measures which rely on both association and frequency the results observed here indicate that employing a 'pure association' approach as is the case for OddsRatio overcorrects for a perceived overreliance on high frequency items since the crucial long-range hubs connecting word associations themselves, which are commonly seen as the gold standard for textual associations to be extracted by AMs as a proxy for semantic salience (Dekalo & Hampe, 2017, p. 165), do present a reasonably high share of extremely frequent words. OddsRatio, r_{φ}, and $\Delta P_{backward}$ systematically overrepresent mid and low frequency words in terms of the frequency profile of all high (top 200) centrality items. The common reliance on frequency in AM extraction may therefore be seen as actively desirable as long as it is monitored and not displaying a bad fit with typical word-association frequency profiles.

To summarise, there is a general structural difference between usage-based network patterns and patterns found in word association networks which manifests itself in:

1) A focus on actions in key word association clusters and a focus on specialised terminology in collocation clusters
2) A higher clustering coefficient and mean eigenvector centrality in word association networks, indicating more tightly-knit clusters
3) A lower node to edge ratio in word association networks, suggesting an overall more associatively rich embedding
4) Network central hubs in word association networks being more action focused, embodied and evaluative than network central hubs in collocation networks which exhibit a stronger focus on pronouns, determiners, stative verbs and descriptive terms

5.4 Implications and Broader Impact

This chapter outlines the broader implications and impact of the research findings, focusing on their significance for future research, and contributions to existing literature. Overall, the emphasis on the role of statistical knowledge in processing abstractions (Kapatsinski, 2014, p. 29) is evident in the results, which show that different association measures can capture varying aspects of lexical relationships. The findings also align with the idea that linguistic knowledge is emergent and ever-changing (Bybee, 2013, p. 50), as the analysis of balanced corpora like the BNC 2014 provides insights into the types of constructions that speakers are likely to have encountered, highlighting the dynamic nature of language use (Herbst, 2018, p. 6). Looking at individual aspects of results emerging from this thesis, the following observations can be made:

Firstly, the findings of this research can be used to further probe the concept of spreading activation. The identification of individual high clustering coefficient items in collocation networks provides insights into how spreading activation might function. The theory posited by Chan and Vitevitch (2009) suggests that items with a low clustering coefficient receive higher relative spreading activation due to reduced competition whereas high clustering coefficient items would receive a lower relative spreading activation. The findings in this thesis, e.g. the identification of the top scoring clustering coefficient terms *deaf*, *rod*, *nerve*, and *exist* in SWOW-UK alongside the full clustering coefficient values available for all other networks in the Results section of Appendix A can be used as the basis for future experimental studies. These could, for instance, register processing times for high and low clustering coefficient items derived from different network types to probe the notion that these network properties influence the ease and speed of lexical access (Siew et al., 2019) and allow for a simulation of spreading activation to further inform our understanding of human memory retrieval behaviour.

Beyond this, the importance of function words is highlighted by their significant presence in word association networks and crucial role for forming clusters in this thesis. This supports the

constructionist view that lexicon, morphology, and grammar exist on a continuum (Langacker, 2008; Berber Sardinha, 2020) and ought not to be treated separately. These findings challenge the routine removal of function words (as part of stop word lists) in linguistic analyses, as they play a crucial role in shaping and structuring clusters, enabling contextualised meaning representation. This further underlines the importance of integrated approaches that highlight the interconnectedness of syntax and pragmatics (Newmeyer, 2010, p. 302) as well as emphasising the connection between corpus linguistics and Functional Linguistics by demonstrating how the functionalist view of language as a dynamic and context-dependent system (van Valin, 2003, p. 320) is crucial for providing meaningful analyses of clusters. Amongst commonly removed stop words prepositions are especially important and should be retained since they make up the largest share of closed class items in the word association networks and fulfil crucial functions as part of compound verbs which could otherwise not be semantically analysed at all (see Figure 45 for an illustration of the core role they can play in key clusters).

On a larger scale, the results of this research further show a prevalence of value judgements in clusters containing otherwise very concrete terms as illustrated by the *phone/touch* cluster (see Figure 53). This strengthens the argument that usage-based linguistic datasets are strongly impacted by emotional and affective relations which are universally present when communicating or during the word association task (Kempe et al., 2013; Out et al., 2020; Sereno et al., 2015) and encourages research expanding research expanding Kousta et al. 's (2011) findings that systematic mental and linguistic processing differences emerge based on whether emotionally charged or neutral concepts are processed. The present data indicates that this could affect even seemingly value neutral lexical items such as *phone* as exemplified in Chapter 4.3.1.

Another outcome of the large-scale network comparison not of word associations and collocations, but more narrowly just within different types of collocation networks carried out in this thesis regards linguistic item that self-collocate. The findings show that changing association measures used on the same corpus leads to radically different outcomes in terms of the number of self-loops, that is words that collocate with themselves, constituting collocations. Log Dice produces a very high number of about 4000 self-loops (11.4% of all collocations identified via log Dice in the entire corpus) whereas the Poisson network exhibits none. For reference, the word association network of SWOW-UK does contain self-loops, in this context they represent self-associations, but these only amount to 0.9% of the total word associations in this network. This has implications for AM selection in lexicography and language learning since self-collocating items cannot be harnessed to extract or refine meaning in the former context or to create pedagogical recommendations for vocabulary expansion in the latter context.

Lastly, the findings of the present thesis are broadly in line with existing research positing that there is a partial thematic organisation of highly connected items in large-scale word association networks (Deyne et al., 2016, p. 58). In Deyne et al. (2016), the authors use a Dutch dataset and identify the themes of *water*, *food*, *money*, *car* and *pain* as core. The highest centrality items identified in SWOW-UK and SWOW-EN form similar thematic central hubs, also containing the *food* theme. This supports the notion that concepts strongly connected to basic needs form central hubs in the mental lexicon across languages. Beyond this, another prevalent theme in the examination of high eigenvector centrality, betweenness centrality, and degree centrality items in the present study are value judgements (*good* and *bad*) and colours (*black*, *white* and *red*). It is crucial to note here that previous studies primarily focus on nouns in identifying the central themes, in the present study this distinction has not been made which may lead to adjectives being dominant examples in this context.

In conclusion, the research outlined in this thesis offers significant implications for numerous facets of linguistic research and practice, especially as a basis for testing, for instance, how well clustering coefficient values derived from different network types predict processing times, an application of the method established in this thesis in other (linguistic) sub-disciplines (Chapters 5.6 and 5.7), and informing methodological choices in corpus linguistics specifically (Chapter 5.5).

5.5 Recommendations for Future Approaches to Psycholinguistically Plausible Collocation Extraction

Lastly, the findings described in Chapters 5.1 - 5.3 can be used to highlight the distinct structural characteristics of word association networks compared to collocation-based networks and can therefore be used as the basis for recommending new approaches to psycholinguistically plausible and relevant collocation extraction for corpus linguists. Six specific recommendations are addressed here:

Firstly, introducing a ceiling for AM scores of words collocating with themselves (should be considered given that word-association-based networks contain a comparatively low number of self-loops. Secondly, corpus pre-processing, or, more optimally, corpus compilation should contain a basic step identifying tables and other highly regularised structural patterns and mark these up so that collocations resulting from lists or tables are weighed systematically lower than collocations emerging from un-structured running text. This is especially relevant when using metrics such as $\Delta P_{forward}$ and $\Delta P_{backward}$ and derived AMs. Thirdly, in a similar vein, when extracting collocations from corpora spanning multiple registers a minimum threshold that has to be met in all sub-registers should be introduced to limit register bias which may arise from preferential-attachment-

style processes leading to discourses being more and more specialised as they progress. Fourthly, the AM values of action verbs (such as *go*, *do*, *show*; see Chapter 5.3 for further context) could be increased systematically in order to better approximate the action-focus in word-association based hubs and clusters. The fifth recommendation regards current practices in NLP where stop-word filtering is commonly employed as described in the preceding Chapter. This approach should be considered very carefully since ambiguity and polysemy are key in interpreting language and the gold-standard of word association networks display a rate of 'grammatical' elements in key positions that is perfectly in line with collocation-based results obtained without the removal of stop-words. The sixth and final recommendation concerns corpus linguistic methods and posits that advances in AM development should not be dismissive of frequency effects since these effects can be observed for true word associations themselves. This is exacerbated by the fact that frequency profiles are easily controllable as secondary factors e.g. using values modified from frequency information as a multiplier to be applied to AM results.

5.6 Limitations and Starting Points for Future Work

Understanding the limitations of a study is crucial, especially in a methodologically oriented thesis, as it provides a framework for interpreting the findings and identifying areas for future research. This Chapter aims to demonstrate how the findings from this thesis, particularly methodological groundwork presented when answering RQ1 and the interactive and adaptable code in the accompanying Jupyter lab could be expanded, tested, and replicated on multiple levels.

One practical limitation is the missing implementation of filters based on dispersion, which measures the rate at which specific nodes and collocates co-occur in different contexts. Future work should incorporate dispersion metrics, such as DP, to better understand and filter results, as dispersion is relevant to the psycholinguistic reality of collocations and has been linked to reaction times, fluency, and processing productivity (Gries, 2013, p. 155; Gries & Ellis, 2015, p. 233). One approach to including dispersion information could be using the dispersion metric DP, one of a range of different dispersion measures available to Corpus Linguists. DP is calculated by obtaining the size of every subsection of the corpus relative to the entire dataset, as well as the relative frequency of the collocation of interest within it. The difference in these values is then calculated and summed up for all subsets of the corpus, the result is halved. This metric thus ranges from 0 to 1, 0 being a perfectly even distribution (Gries & Ellis, 2015, p. 233). This links back to the recommendation of limiting register effects made in Chapter 0. Another area for future research is an exploration of small-worldedness of collocation networks using the small world measurement ω in order to categorise the networks at hand. This was not feasible in this thesis due to high

computational costs but could provide insights into the structural makeup of collocations and link them to known networks structures found in different domains.

Additionally, considering participant or speaker age and health conditions as a separate variable could be valuable. The relevancy of this can be exemplified when looking at semantic dementia. The progression of this disease affects the ability to name low-frequency objects and specific attributes before general ones (Divjak & Caldwell-Harris, 2019, p. 65) and would thus heavily skew the baseline word association dataset. Despite the fact that they could not be explored at length for reasons of brevity in this thesis, future work should further encourage comprehensive studies taking into account equally interesting and worthwhile alternative network representations of linguistic relationships such as orthographic networks (Korkiakangas & Lassila, 2018; Siew, 2018; Trautwein & Schroeder, 2018, p. 12), phonological networks (Neergaard et al., 2019; Siew & Vitevitch, 2019; Vitevitch, 2008), or syntactic networks (Cong & Liu, 2014; Jiang et al., 2019). While making no claim to be exhaustive, these limitations highlight areas for future research to enhance the understanding and application of collocation and word association networks. The following subchapters explore a select few further fundamental limitations.

5.6.1 Language as a Complex Adaptive System

Having explored different storage models in the ML as well as concepts such as Statistical Learning as the theoretical foundation of this thesis in Chapter 2.5, it is also important to consider other factors influencing language production that are not rooted in word co-occurrences. The presented model of the mechanism driving the circular relationship between language perception and language production (Figure 1) that serves as the basis of the present study is linked to a larger underlying structure: a Complex Adaptive System (CAS; Clark (1996, p. 25) that encompasses social interaction, cognitive mechanisms and patterns of experience (here co-occurrences of linguistic elements). All human behaviour and therefore also information exchange via language operates within a CAS. The underlying theory emphasises that language does not exist in a vacuum and only interacts with itself – it is also a deeply social phenomenon governed by cognitive limits. This is the case in the sense that language is a social act of an interaction between people, and the conventionalization of specific norms is therefore essential for effective interaction (here: communication) between speakers. According to the CAS framework, there are four distinct levels of language: production, identification, and understanding of utterances, as well as finally the execution of behaviour (ibid.). The network-based corpus analyses in this project are largely focused on the first level of this system, the production of the utterance, whereas the cue-association networks provide insights into the identification and recognition phases respectively. The snapshot views into these domains which are available via this methodology are useful, but it

is important to stress that whole segments such as the execution of behaviour cannot be researched using these methods alone. It is furthermore important to make the point that language in use is heavily influenced by social norms and conventions, cultural preferences, and taboos, as well as individual relationships between the author/speaker and the reader/listener – all of these variables inevitably add complexity to the systems at hand and cannot be investigated using networks containing linguistic information alone.

The approach taken in this thesis is nevertheless motivated by the fact that linguistic conventions used to communicate in any given situation by default rely on prior use of these conventions. What network approaches can contribute to further understanding of linguistic processes at large is visualising and displaying this use of conventions. Beyond this, network approaches can also serve to contrast this with representations of mental and primarily non-communicative relationships between said conventions. While not related to the approach taken in this thesis, social networks are another area of network science that could significantly contribute to furthering the understanding of language change; for example through analysing social clusters alongside specific collocations or key terms (Beckner et al., 2009, p. 17) in future work.

5.6.2 Multilingual and Longitudinal Data

A further caveat of this thesis the fact that only one language has been considered as the basis for all analyses. It is important to note that English has not been chosen as the medium for this study by default, this choice was merely motivated by the fact that the largest and highest quality comparable datasets that could constitute the basis for this study happened to be collections of (British) English texts. While the BNC 2014 is not the most extensive English corpus by a large margin (see iWeb with a size of 14 billion words (M. Davies, 2018), or the 2020 update[28] of COCA (M. Davies, 2008-) which has been extended to a size of 1 billion words by incorporating web-based data), it is the largest collection of continuously balanced and adequately pre-processed British English whose sampling timeframe perfectly overlaps that of the SWOW project. A further motivation for choosing the largest available suitable dataset is that the computational hurdles grow exponentially with the number of words in the corpus/database. A methodology developed to deal with a less extensive dataset as the primary test case would render it impossible to use the methodology instantly on a larger dataset; this would then require re-writing large parts of the code and incorporating new computational optimisation strategies on top of language-specific idiosyncrasies.

[28] https://www.english-corpora.org/coca/help/new2020.asp

While no claims can be made about the exact structure and graph theoretical properties of networks stemming from languages other than English, the underlying theoretical foundation of Statistical Learning and language as a CAS is applicable universally. This is the case since influences of frequency of occurrence and Statistical Learning present universal properties of human mental processing. Observing and registering reoccurring patterns in everyday life is the foundation for a large number of non-linguistic processes such as problem-solving, motor tasks etc., and can be seen as a universal property of human learning. Consequentially and despite the limitations of this project to a (British) English corpus and an English word association database as the network basis, inferences on a more general, structural level might still be possible. In order to go beyond surface level observations of learnability and to enable insights into specific clusters or individual nodes, contrastive research involving languages other than English as the basis for large-scale linguistic networks is absolutely essential. Particularly research into non-Indo-European languages would be invaluable for assessing how universally generalisable claims pertaining to the structure of the ML on the basis of solely English data really are. Laudable efforts towards examining non-Indo-European language networks have been made by Kovács et al. (2021) who examine word association networks and their emergent community structures. Contrastive studies comparing the obtained results of this and other studies of non-Indo-European languages to the existent body of research covering both word-association and collocation networks of Indo-European languages are therefore highly recommended.

Beyond the urgent need for multilingual research in this area, another promising starting point for furthering the understanding of the structure of the ML lies in the empirical examination of long-term knowledge of regularities for word recognition studies. As identified by Frost et al. (2019, p. 1136), there is a lack of psycholinguistic research examining learning patterns in a timeframe larger than a few minutes; this knowledge is, however, crucial for investigating the construction and shape of the mental lexicon. The results from this thesis might serve as the basis for constructing a representation of a participant's mental lexicon using volunteered existing language data. An observation of their learning mechanisms over a much longer time span through continuous data collection and dynamic extension of the network should be used to examine said heavily under-researched effects in future research.

5.7 Practical Application of the LLN method

While taking the current limitations into account, it is important to remark on the potential network methodologies hold for advancing specific areas of linguistic research in alignment with the aim of this thesis to advance corpus linguistic methodology. Exploring future applications alongside the development of a new processing pipeline is essential in order to make domain experts aware of

the potential reward gained from obtaining a skillset required to carry out large linguistic network analyses. In this Chapter, a broad overview over some other areas of linguistic research, namely language pedagogy, translation, conceptual metaphors, and semantic prosody is therefore given. Apart from listing applications for collocation research in these areas, existing applications of graph theoretical methods to these research areas are mentioned, highlighting the immediate utility of the novel approaches proposed in this thesis which could serve to expand this body of work.

5.7.1 Language Pedagogy and Phrase Extraction

The first area of research where collocations play a considerable role is language pedagogy, chiefly in terms of L2 learning, but secondarily also with regards to Child Language Acquisition. Existing literature consistently reports differences between the use of collocational patterns by L1 and L2 speakers (Gablasova et al., 2017, p. 172), which suggests that the use of appropriate collocations is mastered relatively late in the learning process. The competent use of collocational structures contributes to higher fluency, an improved ad hoc language production (Brezina, 2018, p. 71; Webb & Kagimoto, 2009, p. 55), and prevents comprehension difficulties on the L1 speaker's end when confronted with atypical collocations (Sonbul & Siyanova-Chanturia, 2021, p. 8). Systematically identifying common and structurally relevant collocations on the basis of a psycholinguistically plausible methodology and, for example, high network centrality in the desired variety of the target language. The results of this thesis can therefore be used to guide learners to the most central and connective terms and concepts first, which holds great potential as a strategy for improving learning materials and enhancing learner fluency.

Related to this is the generation of lexicographic resources in a broader context, especially with regards to phrase extraction. There have been efforts to construct phrase banks such as the *Academic Formulas List* constructed by (Simpson-Vlach & Ellis, 2010, p. 510) on the basis of collocational statistics. Major dictionaries such as the Oxford Dictionary of English (NODE, Stevenson (2010)) are informed by corpora that provide frequency information regarding individual words to give indications as to which words and phrases are most essential (Hanks, 2012, p. 220). In the case of the NODE, the underlying dataset used for these calculations is the Oxford English Corpus, a 2.1-billion-word sample of English writing predominantly composed of online sources spanning different varieties of global Englishes. Specific collocation dictionaries that are often directly based on corpus methodologies are also available, particularly for the use of language learners (e.g. the Oxford Collocation Dictionary (McIntosh et al., 2009)). Beyond raw frequency information, major dictionaries with an exceptionally large readership such as the abovementioned NODE, the Macmillan English Dictionary for Advanced Learners (Rundell, 2007), as well as the Cambridge Advanced Learner's Dictionary (McIntosh, 2013) rely on collocational frequencies in

particular in order to provide phraseological information for illustrative purposes. A clear difference between corpus-based and non-corpus-based dictionaries can be observed in terms of the breadth of examples given – intuition alone rarely led to the type of near comprehensive phraseological descriptions of dictionary entries that can be observed in dictionaries that take collocation frequencies into account (Hanks, 2012, pp. 228–229). Interestingly, a trend towards more similar dictionary entries can be observed in corpus-based dictionaries when compared to older dictionaries that had been established on the basis of the editors' intuitions (Hanks, 2012, p. 224). These introspective definitions were more subjective, while a clearly defined common ground is generally desirable over biased definitions, this development also lowers domain specificity. Novel methods for retrieving network-based collocational information based on texts from specific registers or user groups might be innovative solutions to this issue. Using such an approach, specific dictionaries for English in the courtroom, English in casual conversation etc. could be constructed near automatically on top of existing templates.

5.7.2 Enhancing Translation Accuracy and Accounting for Semantic Prosody

In a similar vein, collocations can also be used to help identify near-synonyms and to gain insights into the semantic prosody of specific lexical items (Xiao & McEnery, 2006, p. 125); this in turn helps to improve not only the establishment of precise pedagogical rules but also translation accuracy. Collocation data is, for example, directly used in valency dictionaries (Herbst, 2018, p. 12) which are relied on as resources for high-precision translations. Comparisons of collocational density between original and translated texts have furthermore shown that similar effects to the ones observed in language pedagogy can be found in this domain: Translated texts have been shown to contain a limited range of collocations when compared to the untranslated source texts (Dayrell, 2007, p. 398). On the basis of this, collocation frequency-based metrics can be used for assessing translational quality and language proficiency.

All of these investigations rely on the establishment of a gold standard where specific collocations are listed that can then be used to count the presence or absence of these constructions in L2 language or translations – the more accurate and suitable the gold standard, the better the evaluation. Corpus methods and thus also information regarding collocational patterns is also directly used to inform language teaching and translational practices, for example as the basis for the abovementioned valency dictionaries as well as in the form of distributional approaches to assess formulaity (Gablasova et al., 2017, p. 158). New methodologies such as the use of large-scale collocation networks and graph theoretical analyses hold a great potential to facilitate substantial improvements in translation and L2 proficiency research in two ways: by introducing new metrics to identify different types of collocations which are then fed into databases directly

and in refining the basis for gold standards in proficiency and translation accuracy assessments. First steps in this direction have for example been made by Zhao et al. (2018, p. 904) in their study on quantitative learning strategies. They found that since learning costs for particular words can be assessed quantitatively, network models allow for analysing strategic benefits and drawbacks of particular learning strategies such as high frequency words first, alphabetical, or random learning. The outcome of this study is a data-driven recommendation for strengthening the focus on words with a high degree in the learner networks in order to optimised L2 English learning strategies. Using the LLN pipeline, further investigations into highly central terms via network centrality measurements and topic clustering using MCODE clusters could be employed to strengthen this area of research.

5.7.3 Improvement of Linguistic Processing Models

One such area is applied Statistical Learning research. Results from the LLN pipeline could, for instance, be used to optimise biologically plausible systems aimed at modelling the inner workings of linguistic processes directly. Three prominent examples for such systems are simple recurrent networks (SRNs; originally proposed by Elman (1990)), Parser models (Perruchet & Peereman, 2004, p. 109), and ACT-R (Adaptive Control of Thought-Rational or Atomic Components of Thought (Ritter et al., 2019, p. 1)). All of these models are designed to directly model the inner workings of linguistic processes themselves. SRNs are computational models built on explicit underlying statistical rules. These are fed to the system paired with snippets of linguistic information; the model then predicts the most likely candidate for the next language element on the basis of this. This linear approach is then refined by backpropagation which leads to a stepwise improvement in performance (Perruchet & Peereman, 2004, pp. 107–108). The Parser model, on the other hand, does not directly rely on statistical calculations. Here, perceived sequences are stored in memory as individual chunks that are either strengthened through re-occurrence of the same pattern or weakened/divided into new chunks in the absence of repetitions and the presence of too many similar chunks. This model is designed to mimic human associative learning processes. ACT-R, lastly, is a theory about the human cognitive architecture that aims to model semantic processing steps via incorporating current knowledge of human cognition (Ritter et al., 2019, p. 1). The focus of ACT-R primarily lies on human working memory which also represents memory transformation and procedural knowledge. It has been applied successfully to a number of linguistic tasks, one of which is sentence production (Reitter et al., 2011, p. 589).

Feeding the models graph-theoretical parameters such as centrality or cluster membership could enhance these approaches and test the usability of purely graph-theoretical measures in an applied psycholinguistic context. Beyond this, a study involving the creation of idiosyncratic collocation

networks for different participants and expanding said networks by adding new words could be contrasted with SRNs, ACT-R and the Parser model in order to highlight strengths and weaknesses of the respective approaches.

5.7.4 Conceptual Metaphors

Beyond these well-established research areas, new advances in corpus methodologies such as the incorporation of graph theoretical analyses proposed in this thesis also bring with them the potential for new applications, e.g. regarding conceptual metaphors. A collocation network-based verification and identification of conceptual metaphors lies at the intersection between corpus linguistics, lexicography and cognitive linguistics and therefore holds great potential for triangulation. The idea behind conceptual metaphors (such as ARGUMENT IS WAR) is that one concept is used to talk about the other which, in turn, has a profound effect on how listeners/readers perceive *and act* in the situation at hand (Lakoff & Johnson, 1980). In the presented example, expressing a verbal disagreement as WAR will, according to conceptual metaphor theory, influence the interaction between participants of the argument: Sides are established, individual battles are fought, and the focus will lie on establishing a clear winner and a loser. Conceptual metaphors have been the subject of linguistic study for many decades and significant contributions to quantifying this research have been made by CL researchers in recent years; the development of a refined, user-friendly methodology in this thesis could further contribute to this. Small-scale collocation networks have, for instance, already been used to explore the empirical reality of conceptual metaphors in everyday language via the shared collocates of the elements within the conceptual metaphor (Brezina, 2018, p. 80). Larger-scale, fully connected collocation networks might then provide more refined and less intuition-driven ways of exploring conceptual metaphors through measuring node distances. Intuitively, nodes that end up in close proximity to one another in a force-directed network carry the potential to be used as conceptual metaphors since the foundation of their proximity lies on them not only being connected to the same linguistic items but being connected to these items with a similar strength of association. Research into this area using detailed graph theoretical analyses of collocation networks does, to the knowledge of the author, not exist at the time of publication and is strongly encouraged since it transgresses the functionality of existing methods: It can not only be used to verify or falsify existing conceptual metaphors, but it can also automatically detect new conceptual metaphors that are not yet discussed in the literature.

5.7.5 Meaning Disambiguation: Semantic Prosody & Stance

In addition to that, collocations are also used to investigate lexicographic meanings in greater detail, especially in terms of meaning disambiguation. One such application is an analysis of the semantic prosody of a given word on the basis of its collocates. The collocates indicate what contexts the node is commonly used in and thereby serve as an indication of whether or not the node displays a tendency to have positive or negative opaque connotations associated with it (Ellis et al., 2009, p. 90), see Xiao and McEnery (2006) and Borba and Jaeger (2011) for applied examples of this approach. Collocations further help shed light on discourse strategies such as stance which is defined as a way of communicating an author's position regarding the reader and the subject matter and a way of expressing notions such as the author's integrity or involvement (Jihua Dong & Buckingham, 2018, p. 121). Collocation networks, specifically the GraphColl tool available in #LancsBox (Brezina et al., 2020), have been used to investigate collocations of stance phrases; the findings indicate that there are differences in the distribution of four categories of stance phrases (cognitive, attitude, hedges and reference) depending on the research community in which they are used. These differences consequently indicate that there different communicative norms are shaped by and shape different research communities (Jihua Dong & Buckingham, 2018, p. 130).

The toolkit that is currently available to examine semantic prosody and stance could also be expanded using graph theoretical analyses of collocation networks. Graph theoretical parameters of individual nodes such as centrality and connectivity measures could enrich existing research strategies. Information regarding the centrality of a node in terms of its semantic prosody alongside information on its shortest paths to words with a particular negative or positive connotation could, for example, be carried out. Individual communicative strategies and stance could furthermore be explored in new ways through network analyses that focus on the underlying grammatical patterns. These might provide further indications as to which communicative strategies are employed, how complex or easily understandable a sample text is and what central topics a specific author regards more important than others. This can be carried out in a similar fashion as keyword analyses: By comparing the normalised centrality scores of nodes present in a given author's work with the normalised centrality in larger networks comprising a large number of texts of the same genre.

6 Conclusion

The motivation behind this thesis encompassed three aims. The first aim is methodological innovation, specifically the development of LLN, a novel approach for generating large linguistic networks by integrating corpus linguistics, psycholinguistics, and graph theory, which lies at the heart of the project. This interdisciplinary framework was designed to bridge the gap between psycholinguistics and corpus linguistics, thereby enhancing the analysis of linguistic patterns and structures. The second aim is a critical evaluation of current practices in collocation extraction and the generalisability of findings from collocation studies. This involves a thorough assessment of the reliability and validity of different association measures to ensure they accurately represent human linguistic knowledge in a psycholinguistically plausible manner. This means that the proposed network generation pipeline must be adaptable to new psycholinguistic findings and alternative theories. Despite the uncertainties associated with relying on empirical knowledge, the pursuit of this research remains worthwhile. The methodological decisions underpinning the proposed network generation pipeline can be dynamically adjusted in response to new evidence, ensuring ongoing relevance and accuracy. This adaptability is crucial for maintaining the robustness and applicability of the research findings in the face of evolving scientific knowledge. The concept of psycholinguistic plausibility is central to this aim, as it requires that the generated networks align with current theories and experimental findings from cognitive linguistics and psycholinguistics. This critical evaluation was used as the foundation for developing the new methodology employed here.

The third aim is to examine the generalisability of collocations to mental associations through a contrastive analysis of a large word association network (SWOW-EN) with holistic collocation networks (BNC 2014). This evaluation is crucial for understanding how repeated textual co-occurrence influences mental processing in readers and listeners. By systematically comparing these networks, the thesis aims to provide insights into the cognitive processes underlying language use and comprehension, and to recommend a change in methodology based on this.

In following these aims, this thesis has provided comprehensive answers to the three research questions outlined in Chapter 2.9. The first research question explored whether current approaches to AM extraction can be harnessed in a psycholinguistically plausible manner. Chapter 3.4 posits that psycholinguistic plausibility is a multifaceted concept requiring careful consideration of various parameters and that the choice of association metric is pivotal, with some metrics better capturing human cognition nuances than others. Given that principles such as the choice of suitable filters, collocation windows, and Graph Theoretical analytics are adhered to and that it the limitation to a single layer of linguistic representation is expressed, this thesis concludes that psycholinguistically plausible collocation extraction is possible. An especially crucial aspect is the directionality of

collocations—whether bidirectional or unidirectional — since this is pivotal information for interpreting collocational relationships. By addressing the factors outlined in Chapter 3.4, this thesis provides a robust framework for future research, ensuring that AM extraction methods are scientifically sound, interpretable, and applicable to theoretical questions.

The second research question investigated which AMs lead to collocation networks that best approximate the content and structure of large word association networks. The analysis, conducted at the levels of individual words, clusters, and entire networks, indicates that combinations of association measures, particularly log Dice, LL, and $\chi 2$, provide the best approximation. However, systematic discrepancies remain across all three levels, leading to the discussion of RQ3.

The final research question explored structural differences between usage-based network patterns and word association networks both on a qualitative and on a quantitative level. The first sub-question addressed the intuitive interpretability of these networks, revealing that while they are interpretable, a holistic network understanding requires a full analysis of their properties, especially at the meso-level where qualitative analyses of clusters are paramount. Key differences that have been observed between the network types include a lower rate of self-loops and a higher average clustering coefficient and mean eigenvector centrality in word association networks when compared to collocation networks. These findings suggest that word association networks are more tightly-knit and contextually rich than the examined collocation networks.

The second sub-question examined the percentage overlap between the strongest unweighted nodes and edges in different collocation networks and the word association network. The best scoring collocation metric showed only a 4.4% overlap with the word association network, indicating that collocations should not be used as a direct and sole proxy for word association. This is particularly relevant in corpus-assisted discourse analysis, where collocation methods are often employed to generalize towards stance and attitude. The findings advocate for a more nuanced approach, evaluating how repeated textual co-occurrence influences mental processing beyond collocational analyses alone.

The remaining sub-questions focused on qualitative comparisons at the cluster and word level. The analysis revealed that while there is some overlap in topics represented in clusters from different networks, notable differences exist. Word association clusters prominently feature actions and evaluative terms, whereas collocation clusters are more specialized, often reflecting distinct registers. Additionally, words with high associative availability in word association networks tend to be more action-focused and embodied, while collocation networks emphasize pronouns, determiners, and specialized terms. These differences underscore the distinct structural

characteristics of word association networks, which are more contextually rich and action-oriented compared to the more specialized and fragmented nature of collocation networks.

In conclusion, this thesis has made specific contributions to the fields of corpus linguistics and cognitive linguistics as well as provide starting points for future psycholinguistic research via the development of a new methodological pipeline, evaluating current practices in collocation extraction, and exploring similarities and differences between word associations and collocations. An essential quality of the networks generated here is that they are based on the entire corpus or word association database which allowed for an exploration of more than just the sum of the words they contain, capturing the rich contextual interlinking between words and clusters, and providing insights into the structure of the dataset and the underlying linguistic patterns as a whole. These insights underscore the importance of using a combination of association measures, especially in CL where single-AM approaches are common, and highlight the potential for future research to build on these results. The development of the LLN pipeline and large linguistic network methodology offers a valuable tool for further exploration and application in various linguistic contexts. By providing a comprehensive framework for analysing linguistic networks, this research paves the way for future studies to build on these findings and develop new methodologies that further the current understanding of language. The insights gained from this research have the potential to impact a wide range of fields, from natural language processing to cognitive linguistics, and underscore the importance of interdisciplinary approaches in advancing the understanding of complex linguistic phenomena.

Appendix A

Appendix A (Schmück, 2024) is an online appendix and can be found at https://doi.org/10.17605/OSF.IO/8ANGY, future releases can be accessed at https://github.com/hannaschmueck/LLN. The visualisation below provides a preview of the commented codebase. The individual network files are provided in the Results subfolder. The final codebase contains five separate files and equates a total of 119 pages of pdf printout.

The resulting split df can be used to generate contingency tables as the basis for AM calculations. Since there are differences in how existing software generates contingency tables, the approach taken here will be exemplified in the following section:

observed	word2	~~word2~~	x
word1	O11	O12	R1
~~word1~~	O21	O22	R2
x	C1	C2	N

expected	word2	~~word2~~	rows
word1	(R1C1)/N	(R1C2)/N	R1
~~word1~~	(R2C1)/N	(R2C2)/N	R2
cols	C1	C2	N

This generates the contingency dfs for each subsection of the corpus.

[91]:
```
def createdirectionaldf(split_df):
    split_df = split_df.copy()
    startswithdf = split_df.groupby('node').sum(numeric_only=True)
    endswithdf = split_df.groupby('collocate').sum(numeric_only=True)
    endswithdf = endswithdf.reset_index()
    endswithdf.rename({'index' : 'collocate'})
    startswithdf = startswithdf.reset_index()
    startswithdf.rename({'index' : 'node'})
    return [startswithdf,endswithdf]
```

This will be used to calculate MI and related scores.

[92]:
```
def MI(contingency_df):
    MI = contingency_df.copy()
    MI["MI"] = np.log2(MI["O11"]/MI["E11"])
    return MI["MI"]

def MI2(contingency_df):
    MI = contingency_df.copy()
    MI["MI2"] = np.log2(MI["O11"]**2/MI["E11"])
    return MI["MI2"]


def MI3(contingency_df):
    MI = contingency_df.copy()
    MI["MI3"] = np.log2(MI["O11"]**3/MI["E11"])
    return MI["MI3"]

def MI4(contingency_df):
    MI = contingency_df.copy()
```

Appendix B

Appendix B (Schmück, 2024) is the online appendix for Chapter 4.3.3 and can be found at https://doi.org/10.17605/OSF.IO/8ANGY in the Appendix B subfolder.

It contains 17 files in total, cys (network), png files, svg files, as well as excel files showing word class distributions and categorisations for each of the three clustered networks and a pdf file showing Figure 41 in full resolution. It also contains one excel file showing types of collocation found in all three clustered networks.

- 432_CCDFs.pdf
- 433_BNC_logDice_X2_clusters_wordclass_distribution.xlsx
- 433_SWOW_UK_clusters_wordclass_distribution.xlsx
- 433_BNC_logDice clusters_wordclass_distribution.xlsx
- 433_Types_Of_Collocation_Clusters.xIsx
- 433_BNC_logDice ll X2_Annotated_Clusters.xlsx
- 433_BNC_logDice_Annotated_Clusters.xlsx
- 433_SWOW-UK_Annotated_Clusters.xlsx
- 433_BNC_logD_Clusters.cys
- 433_BNC_logD_ll_X2_Clusters.cys
- 433_SWOW-UK_Clusters.cys
- 433_BNC_logD_Annotated_Clusters.svg
- 433_BNC_logD_ll_X2_Annotated_Clusters.svg
- 433_SWOW-UK_Annotated Clusters.svg
- 433_BNC_logD_Annotated_Clusters.png
- 433_BNC_logD_ll_X2_Annotated_Clusters.png
- 433_SWOW-UK_Annotated Clusters.png

Appendix C

Appendix C (Schmück, 2024) is the online appendix for Chapter 4.3.4 and can be found at https://doi.org/10.17605/OSF.IO/8ANGY in the Appendix C subfolder.

It contains four excel files with top centrality and clustering coefficient items for each of the networks.

- 434_Betweenness Centrality.xlsx
- 434_ClusteringCoefficient.xlsx
- 434_Degree Centrality.xlsx
- 434_Eigenvector Centrality.xlsx

List of Abbreviations

ACOM	Automatic Contexonym Organising Model
ACT-R	Adaptive Control of Thought-Rational or Atomic Components of Thought
AM	Association Measure
ASD	Adjacency Spectral Distance
BC	Betweenness Centrality
CAS	Complex Adaptive System
CCDF	Complementary Cumulative Distribution Function
ClC	Closeness Centrality
ClCoef	Clustering Coefficient
DC	Degree Centrality
$\Delta P_{backward}$	Delta P Forward (AM)
$\Delta P_{forward}$	Delta P Forward (AM)
EC	Eigenvector Centrality
fMRI	functional Magnetic Resonance Imaging
EEG	Electroencephalography
ERP	Event Related Potential
L2	Second Language
LL	Log Likelihood (AM)
LLN	Large Linguistic Network Methodology
LNRE	Large Number of Rare Events
LPM	Limited Paradigmatic Modifiability
LSA	Latent Semantic Analysis
LSM	Limited Syntagmatic Modifiability
MI	Mutual Information (AM)
ML	Mental Lexicon
MWE	Multi Word Expression
NODE	Oxford Dictionary of English
POS	Part Of Speech
r_φ	Pearson's Rho (AM)
SL	Statistical Learning
SRN	Simple Recurrent Networks
USF	University of South Florida norms
χ^2	Chi-Squared (AM)

7 References

Adult literacy Program for the International Assessment of Adult Competencies. (2012). *Survey of Adult Skills: Figure 0.3 Literacy skills gap between older and younger generations.* https://doi.org/10.1787/888932903671

Aitchison, J. (2008). *Words in the mind: An introduction to the mental lexicon* (3. ed., reprinted.). Blackwell.

Akmajian, A., Demers, R. A., Farmer, A. K., & Harnish, R. M. (Eds.). (2010). *Linguistics: An introduction to language and communication* (6th ed.). MIT Press.

Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, *29*(3), 626–688. https://doi.org/10.1007/s10618-014-0365-y

Anthonisse, J. M. (1971). The rush in a directed graph. *Journal of Computational Physics*, 1–10.

Anthony, L. (2022). *AntConc* (Version 4.0.5) [Computer software]. https://www.laurenceanthony.net/software

Arbesman, S., Strogatz, S., & Vitevitch, M. (2010). Comparative Analysis of Networks of Phonologically Similar Words in English and Spanish. *Entropy*, *12*(3), 327–337. https://doi.org/10.3390/e12030327

Arppe, A., Gilquin, G., Glynn, D., Hilpert, M., & Zeschel, A. (2010). Cognitive Corpus Linguistics: five points of debate on current theory and methodology. *Corpora*, *5*(1), 1–27. https://doi.org/10.3366/cor.2010.0001

Bader, G. D., & Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, *4*(2). https://doi.org/10.1186/1471-2105-4-2

Baker, P. (2016). The shapes of collocation. *International Journal of Corpus Linguistics*, *21*(2), 139–164. https://doi.org/10.1075/ijcl.21.2.01bak

Baker, P., & Collins, L. (2023). Creating and analysing a multimodal corpus of news texts with Google Cloud Vision's automatic image tagger. *Applied Corpus Linguistics*, *3*(1), 100043. https://doi.org/10.1016/j.acorp.2023.100043

Bales, M. E., & Johnson, S. B. (2006). Graph theoretic modeling of large-scale semantic networks. *Journal of Biomedical Informatics*, *39*(4), 451–464. https://doi.org/10.1016/j.jbi.2005.10.007

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In L. Kennedy (Ed.), *Visual Word Recognition: Models and Methods, Orthography and Phonology* (pp. 90–115). Psychology Press.

Barlow, M., & Kemmer, S. (2000). *Usage-based models of language.* CSLI Publications.

Barnbrook, G., Mason, O., & Krishnamurthy, R. (2013). The concept of collocation. In G. Barnbrook, O. Mason, & R. Krishnamurthy (Eds.), *Collocation: Applications and Implications* (1st ed., pp. 3–31). Palgrave Macmillan. https://doi.org/10.1057/9781137297242_1

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in Cognitive Sciences*, *17*(7), 348–360. https://doi.org/10.1016/j.tics.2013.04.010

Bartsch, S. (2004). *Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Gunther Narr.

Bassett, D. S., Wymbs, N. F., Porter, M. A., Mucha, P. J., Carlson, J. M., & Grafton, S. T. (2011). Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(18), 7641–7646. https://doi.org/10.1073/pnas.1018985108

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In ICWSM (Ed.), *ICWSM: Proceedings of the third International AAAI Conference on Weblogs and Social Media : 17-20 May 2009, San Jose, California USA*. AAAI Press.

Bauer, L. (2019). Compounds and multi-word expressions in English. In B. Schlücker (Ed.), *Konvergenz und Divergenz: Vol. 9. Complex Lexical Units: Compounds and Multi-Word Expressions* (pp. 45–68). De Gruyter. https://doi.org/10.1515/9783110632446-002

Beckage, N. M., & Colunga, E. (2016). Language Networks as Models of Cognition: Understanding Cognition through Language. In A. Mehler, A. Lücking, S. Banisch, P. Blanchard, & B. Job (Eds.), *Understanding Complex Systems. Towards a Theoretical Framework for Analyzing Complex Linguistic Networks* (pp. 3–28). Springer. https://doi.org/10.1007/978-3-662-47238-5_1

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language Is a Complex Adaptive System: Position Paper. *Language Learning*, *59*, 1–26. https://doi.org/10.1111/j.1467-9922.2009.00533.x

Benedek, M., Kenett, Y. N., Umdasch, K., Anaki, D., Faust, M., & Neubauer, A. C. (2017). How semantic memory structure and intelligence contribute to creative thought: a network science approach. *Thinking & Reasoning*, *23*(2), 158–183. https://doi.org/10.1080/13546783.2016.1278034

Berber Sardinha, T. (2020). Lexicogrammar. In C. Chapelle (Ed.), *The concise encyclopedia of applied linguistics* (First edition, pp. 1–5). Wiley Blackwell. https://doi.org/10.1002/9781405198431.wbeal0698.pub2

Berlingerio, M., Koutra, D., Eliassi-Rad, T., & Faloutsos, C. (2012). *NetSimile: A Scalable Approach to Size-Independent Network Similarity*. https://doi.org/10.48550/arXiv.1209.2684

Bestgen, Y. (2018). Getting rid of the Chi-square and Log-likelihood tests for analysing vocabulary differences between corpora. *Quaderns De Filologia - Estudis Lingüístics*, *22*(22), 33. https://doi.org/10.7203/qf.22.11299

Bickenbach, J. E., Davies, J. M [Jacqueline MacGregor], Davies, J. M [Jackie M.], & MacGregor Davies, J. (1997). *Good reasons for better arguments: An introduction to the basic skills and values of critical thinking*. Broadview Press.

Biemann, C. (2012). *Structure discovery in natural language. Theory and applications of natural language processing*. Springer.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Bonacich, P. (1972). Technique for Analyzing Overlapping Memberships. *Sociological Methodology*, *4*, 176. https://doi.org/10.2307/270732

Borba, R., & Jaeger, A. (2011). "They Never Realized That, You Know": Linguistic Collocations and Interactional Functions of You Know in Contemporary Academic Spoken English. *The ESPecialist*, *32*(2), 195–215.

Bordag, S. (2003). Sentence Co-occurrences as Small-World Graphs: A Solution to Automatic Lexical Disambiguation. In A. Gelbukh (Ed.), *Lecture Notes in Computer Science: Vol. 2588, Computational Linguistics and Intelligent Text Processing: 4th International Conference, CICLing 2003, Mexico City, Mexico, February 16-22, 2003* (pp. 329–332). Springer. https://doi.org/10.1007/3-540-36456-0_34

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55–71. https://doi.org/10.1016/j.socnet.2004.11.008

Borge-Holthoefer, J., & Arenas, A. (2010). Semantic Networks: Structure and Dynamics. *Entropy*, *12*(5), 1264–1302. https://doi.org/10.3390/e12051264

Brezina, V. (2016). Collocation Networks: Exploring Associations in Discourse. In P. Baker & J. Egbert (Eds.), *Routledge Advances in Corpus Linguistics: Vol. 17. Triangulating methodological approaches in corpus-linguistic research* (pp. 90–107). Routledge.

Brezina, V. (2018). Collocation Graphs and Networks: Selected Applications. In P. Cantos Gómez & M. Almela-Sánchez (Eds.), *Quantitative Methods in the Humanities and Social Sciences: Vol. 10. Lexical collocation analysis: Advances and applications* (Vol. 16, pp. 59–83). Springer. https://doi.org/10.1007/978-3-319-92582-0_4

Brezina, V., Hawtin, A., & McEnery, T. (2021). The Written British National Corpus 2014 – design and comparability. *Text & Talk*, *41*(5-6), 595–615. https://doi.org/10.1515/text-2020-0052

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139–173. https://doi.org/10.1075/ijcl.20.2.01bre

Brezina, V., & Meyerhoff, M. (2014). Significant or random? *International Journal of Corpus Linguistics*, *19*(1), 1–28. https://doi.org/10.1075/ijcl.19.1.01bre

Brezina, V., & Platt, W. (2024). *#LancsBox X* [Computer software]. Lancaster University.

Brezina, V., Weill-Tessier, P., & McEnery, T. (2020). *#LancsBox v. 5.1* [Computer software]. http://corpora.lancs.ac.uk/lancsbox.

Broca, P. (1865). Sur le siege de la Farclte de langage articule. *Bulletins De La Societe D' Anthropologie De Paris*, *6*, 337–393. https://ci.nii.ac.jp/naid/10005608590/

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *5*(4), 325–337. https://doi.org/10.1016/S0022-5371(66)80040-3

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J [Jeffrey], Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*.

Brysbaert, M., Lagrou, E., & Stevens, M. (2017). Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. *Bilingualism: Language and Cognition*, *20*(3), 530–548. https://doi.org/10.1017/S1366728916000353

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*, *27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Burghardt, M. (2018). Visualization as a Key Factor for the Usability of Linguistic Annotation Tools. In N. Bubenhofer & M. Kupietz (Eds.), *Visualisierung sprachlicher Daten: Visual Linguistics – Praxis – Tools* (pp. 315–329). Heidelberg University Publishing.

Butterworth, B. (1983). Lexical representation. In B. Butterworth (Ed.), *Language Production: Vol. 2. Development, writing and other language processes* (pp. 257–294).

Bybee, J. L. (2013). Usage-based Theory and Exemplar Representations of Constructions. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (pp. 49–69). Oxford University Press.

Carrol, G., & Conklin, K. (2020). Is All Formulaic Language Created Equal? Unpacking the Processing Advantage for Different Types of Formulaic Sequences. *Language and Speech*, *63*(1), 95–122. https://doi.org/10.1177/0023830918823230

CASS. (11/2018). *The British National Corpus 2014: User manual and reference guide. Version 1.1*. Centre of Corpus Approaches to Social Science, Lancaster University. http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf

Castro, N., & Siew, C. S. Q. (2020). Contributions of modern network science to the cognitive sciences: Revisiting research spirals of representation and process. *Proceedings. Mathematical, Physical, and Engineering Sciences*, *476*(2238), 1–25. https://doi.org/10.1098/rspa.2019.0825

Cecchini, F. M., Riedl, M., Fersini, E., & Biemann, C. (2018). A comparison of graph-based word sense induction clustering algorithms in a pseudoword evaluation framework. *Language Resources and Evaluation*, *52*(3), 733–770. https://doi.org/10.1007/s10579-018-9415-1

Chan, K. Y., & Vitevitch, M. S. (2009). The influence of the phonological neighborhood clustering coefficient on spoken word recognition. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(6), 1934–1949. https://doi.org/10.1037/a0016902

Chen, A. C.-H. (2022). Words, constructions and corpora: Network representations of constructional semantics for Mandarin space particles. *Corpus Linguistics and Linguistic Theory*, *18*(2), 209–235. https://doi.org/10.1515/cllt-2020-0012

Chen, H., Chen, X., & Liu, H. (2018). How does language change as a lexical network? An investigation based on written Chinese word co-occurrence networks. *PloS One*, *13*(2), e0192545. https://doi.org/10.1371/journal.pone.0192545

Chen, S. (2013). Corpus Linguistics in Critical Discourse Analysis: A Case Study on News Reports of the 2011 Libyan Civil War. *Stream: Interdisciplinary Journal of Communication*, *5*(1), 21–28. https://doi.org/10.21810/strm.v5i1.77

Chetail, F. (2017). What do we do with what we learn? Statistical learning of orthographic regularities impacts written word processing. *Cognition*, *163*, 103–120. https://doi.org/10.1016/j.cognition.2017.02.015

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *User-Oriented Content-Based Text and Image Handling* (pp. 609–623).

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, *51*(4), 661–703. https://doi.org/10.1137/070710111

Collins, A. (1975). A Spreading Activation Theory of Semantic Memory. *Psychological Review*, *82*(6), 407–428.

Cong, J., & Liu, H. (2014). Approaching human language with complex networks. *Physics of Life Reviews*, *11*(4), 598–618. https://doi.org/10.1016/j.plrev.2014.04.004

Covington, N. V., Brown-Schmidt, S., & Duff, M. C. (2018). The Necessity of the Hippocampus for Statistical Learning. *Journal of Cognitive Neuroscience*, *30*(5), 680–697. https://doi.org/10.1162/jocn_a_01228

Croft, W., & Cruse, D. A. (2004). *Cognitive linguistics*. *Cambridge textbooks in linguistics*. Cambridge University Press. http://www.loc.gov/catdir/description/cam032/2003053175.html

Csardi, G., & Nepusz, T. (2005). The Igraph Software Package for Complex Network Research. *InterJournal Complex Systems*(1695).

Dasgupta, T., Sinha, M., & Basu, A. (2016). Computational Models of the Representation of Bangla Compound Words in the Mental Lexicon. *Journal of Psycholinguistic Research*, *45*(4), 833–855. https://doi.org/10.1007/s10936-015-9367-1

Davies, M. (2008-). The Corpus of Contemporary American English (COCA).2008-. https://www.english-corpora.org/coca/

Davies, M. (2018). iWeb: The 14 Billion Word Web Corpus. https://www.english-corpora.org/iweb/

Davies, R. A. I., Arnell, R., Birchenough, J. M. H., Grimmond, D., & Houlson, S. (2017). Reading through the life span: Individual differences in psycholinguistic effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1298–1338. https://doi.org/10.1037/xlm0000366

Dayrell, C. (2007). A quantitative approach to compare collocational patterns in translated and non-translated texts. *International Journal of Corpus Linguistics*, *12*(3), 375–414.

Deerwester, S., Dumais, S. T [Susan T.], Furnas, G. W., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Dehmer, M., & Emmert-Streib, F. (Eds.). (2009). *Analysis of complex networks: From biology to linguistics*. Wiley-VCH.

Dehmer, M., Emmert-Streib, F., & Mehler, A. (Eds.). (2011). *Towards an information theory of complex networks: Statistical methods and applications*. Springer. https://doi.org/10.1007/978-0-8176-4904-3_11

Dehmer, M., Emmert-Streib, F., & Shi, Y. (2017). Quantitative Graph Theory: A new branch of graph theory and network science. *Information Sciences*, *418-419*, 575–580. https://doi.org/10.1016/j.ins.2017.08.009

Dekalo, V., & Hampe, B. (2017). Networks of meanings: Complementing collostructional analysis by cluster and network analyses. *Yearbook of the German Cognitive Linguistics Association*, *5*(1). https://doi.org/10.1515/gcla-2017-0011

Department for Business, Innovation and Skills. (2013). *The International Survey of Adult Skills 2012: Adult literacy, numeracy and problem solving skills in England* (Research Paper No. 139). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/246534/bis-13-1221-international-survey-of-adult-skills-2012.pdf

Deshors, S. C., & Gries, S. T. (2022). Using Corpora in Research on Second Language Psycholinguistics. In A. Godfroid & H. Hopp (Eds.), *The Routledge handbook of second language acquisition and psycholinguistics*. Routledge.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018, October 11). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://arxiv.org/pdf/1810.04805.pdf

Deyne, S. de, Navarro, D. J., Collell, G., & Perfors, A [Andrew] (2021). Visual and Affective Multimodal Models of Word Meaning in Language and Mind. *Cognitive Science*, *45*(1), 1-44. https://doi.org/10.1111/cogs.12922

Deyne, S. de, Navarro, D. J., Perfors, A [Amy], Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006. https://doi.org/10.3758/s13428-018-1115-7

Deyne, S. de, & Storms, G. (2008). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, *40*(1), 198–205. https://doi.org/10.3758/brm.40.1.198

Deyne, S. de, & Storms, G. (2015). Word Associations. In J. R. Taylor (Ed.), *The Oxford handbook of the word* (pp. 465–480). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199641604.013.018

Deyne, S. de, Verheyen, S., & Storms, G. (2016). Structure and Organization of the Mental Lexicon:: A Network Approach Derived from Syntactic Dependency Relations and Word Associations. In A. Mehler, A. Lücking, S. Banisch, P. Blanchard, & B. Job (Eds.), *Understanding Complex Systems. Towards a Theoretical Framework for Analyzing Complex Linguistic Networks* (pp. 47–79). Springer.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, *26*(3), 297–302. https://doi.org/10.2307/1932409

Dimitropoulos, D., Golovin, A., John, M., & Krissinel, E. (2009). Applications of Graph Theory in Chemo- and Bioinformatics. In M. Dehmer & F. Emmert-Streib (Eds.), *Analysis of complex networks: From biology to linguistics*. Wiley-VCH.

Divjak, D., & Caldwell-Harris, C. L. (2019). Chapter 3: Frequency and entrenchment. In E. Dąbrowska & D. Divjak (Eds.), *Cognitive linguistics* (pp. 61–86). De Gruyter Mouton.

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, *11*(03), 357. https://doi.org/10.1017/S0140525X00058027

Dong, J [Jihua], & Buckingham, L. (2018). The collocation networks of stance phrases. *Journal of English for Academic Purposes*, *36*, 119–131. https://doi.org/10.1016/j.jeap.2018.10.004

Dong, J [Jun], & Horvath, S. (2007). Understanding network concepts in modules. *BMC Systems Biology*, *1*, 1–24. https://doi.org/10.1186/1752-0509-1-24

Dorogovtsev, S. N., & Mendes, J. F. (2001). Language as an evolving word web. *Proceedings. Biological Sciences*, *268*(1485), 2603–2606. https://doi.org/10.1098/rspb.2001.1824

Dunning, T. (1993a). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, *19*(1), 61–74.

Dunning, T. (1993b). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, *19*(1), 61–74. https://www.aclweb.org/anthology/J93-1003.pdf

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, *6*(2). https://doi.org/10.1515/cllt.2010.006

Durrant, P., & Siyanova-Chanturia, A. (2015). Learner corpora and psycholinguistics. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 57–78). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.004

Ellis, N. C. (2006). Language Acquisition as Rational Contingency Learning. *Applied Linguistics*, *27*(1), 1–24. https://doi.org/10.1093/applin/ami038

Ellis, N. C. (2019). Essentials of a Theory of Language Cognition. *The Modern Language Journal*, *103*, 39–60. https://doi.org/10.1111/modl.12532

Ellis, N. C., & Frey, E. (2009). The psycholinguistic reality of collocation and semantic prosody (2): Affective Priming. In R. Corrigan, E. A. Moravcsik, H. Ouali, & K. M. Wheatley (Eds.), *Typological studies in language: Vol. 83. Formulaic language: Volume 2* (pp. 473–497). John Benjamins.

Ellis, N. C., Frey, E., & Jalkanen, I. (2009). The psycholinguistic reality of collocation and semantic prosody (1). In U. Römer & R. Schulze (Eds.), *Studies in Corpus Linguistics: Vol. 35. Exploring the lexis-grammar interface* (pp. 89–116). John Benjamins.

Ellis, N. C., & O'Donnell, M. B. (2014). Construction learning as category learning: A cognitive analysis. In H.-J. Schmid, S. Faulhaber, & T. Herbst (Eds.), *Trends in linguistics. Studies and monographs: Vol. 282. Constructions collocations patterns* (71-93). De Gruyter Mouton.

Ellis, N. C., & O'Donnell, M. B. (2012). Statistical construction learning: Does a Zipfian problem space ensure robust language learning? In P. Rebuschat & J. N. Williams (Eds.), *Statistical Learning and Language Acquisition* (pp. 265–304). De Gruyter.

Ellis, N. C., Simpson-Vlach, R., Römer, U., O'Donnell, M. B., & Wulff, S. (2015). Learner corpora and formulaic language in second language acquisition research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 357–378). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.016

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1016/0364-0213(90)90002-E

Emberson, L. L., Misyak, J. B., Schwade, J. A., Christiansen, M. H., & Goldstein, M. H. (2019). Comparing statistical learning across perceptual modalities in infancy: An investigation of underlying learning mechanism(s). *Developmental Science*, *22*(6), e12847. https://doi.org/10.1111/desc.12847

Erdős, P., & Rényi, A. (1960). On the Evolution of Random Graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*(5), 17–61.

Evans, V. (2019). *Cognitive linguistics: An introduction* (Second edition). Edinburgh University Press.

Evert, S. (2004). *Asymptotic hypothesis tests*. http://www.collocations.de/AM/index.html

Evert, S. (2005). *The Statistics of Word Cooccurrences: The Statistics of Word Cooccurrences* [PhD]. Universität Stuttgart, Stuttgart. https://elib.uni-stuttgart.de/bitstream/11682/2573/1/Evert2005phd.pdf

Evert, S. (2008). Corpora and collocations. In A. Lüdeling (Ed.), *Handbücher zur Sprach- und Kommunikationswissenschaft. Corpus linguistics: An international handbook* (pp. 1212–1248). Mouton de Gruyter. https://doi.org/10.1515/9783110213881.2.1212

Evert, S., & Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. In B. L. Webber (Ed.), *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01* (pp. 188–195). Association for Computational Linguistics. https://doi.org/10.3115/1073012.1073037

Evert, S., Uhrig, P., Bartsch, S., & Proisl, T. (2017). E-VIEW-Alation – a Large-Scale Evaluation Study of Association Measures for Collocation Identification. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & B. Vít (Eds.), *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2017 Conference* (pp. 531–549). Lexical Computing.

Fellbaum, C. (2006). WordNet(s). In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (2nd ed., pp. 665–670). Elsevier Science.

Ferrer-i-Cancho, R., & Solé, R. V. (2001). The small world of human language. *Proceedings. Biological Sciences*, *268*(1482), 2261–2265. https://doi.org/10.1098/rspb.2001.1800

Firth, J. R. (1957). *Studies in linguistic analysis*. *Philological Society: Special Volume*. Blackwell.

Fitzpatrick, T., & Izura, C. (2011). Word Association in L1 and L2: An Exploratory Study of Response Types, Response Times, and Interlingual Mediation. *Studies in Second Language Acquisition*, *33*(3), 373–398. https://doi.org/10.1017/S0272263111000027

Fitzpatrick, T., Playfoot, D., Wray, A., & Wright, M. J. (2015). Establishing the Reliability of Word Association Data for Investigating Individual and Group Differences. *Applied Linguistics*, *36*(1), 23–50. https://doi.org/10.1093/applin/amt020

Fitzpatrick, T., & Thwaites, P. (2020). Word association research and the L2 lexicon. *Language Teaching*, *53*(3), 237–274. https://doi.org/10.1017/S0261444820000105

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*(6), 627–635. https://doi.org/10.1016/S0022-5371(73)80042-8

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, *40*(1), 35. https://doi.org/10.2307/3033543

Friederici, A. D., & Gierhan, S. M. E. (2013). The language network. *Current Opinion in Neurobiology*, *23*(2), 250–254. https://doi.org/10.1016/j.conb.2012.10.002

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128–1153. https://doi.org/10.1037/bul0000210

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *Language Learning*, *67*(S1), 155–179. https://doi.org/10.1111/lang.12225

Gabrielatos, C., & Baker, P. (2008). Fleeing, Sneaking, Flooding. *Journal of English Linguistics*, *36*(1), 5–38. https://doi.org/10.1177/0075424207311247

Galasinski, D., & Marley, C. (1998). Agency in foreign news: A linguistic complement of a content analytical study. *Journal of Pragmatics*, *30*, 565–587.

Galton, F. (1879). Psychometric Experiments. *Brain*, *2*(2), 149–162. https://doi.org/10.1093/brain/2.2.149

Garcia, M., García Salido, M., & Alonso-Ramos, M. (2019). A comparison of statistical association measures for identifying dependency-based collocations in various languages. In A. Savary, C. P. Escartín, F. Bond, J. Mitrović, & V. B. Mititelu (Eds.), *Proceedings of the Joint Workshop*

on Multiword Expressions and WordNet (MWE-WN 2019) (pp. 49–59). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5107

Georgakopoulos, T., & Polis, S. (2018). The semantic map model: State of the art and future avenues for linguistic research. *Language and Linguistics Compass*, *12*(2), 1-33. https://doi.org/10.1111/lnc3.12270

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, *7*(5), 219–224. https://doi.org/10.1016/S1364-6613(03)00080-9

Goldstein, R., & Vitevitch, M. S. (2014). The influence of clustering coefficient on word-learning: How groups of similar sounding words facilitate acquisition. *Frontiers in Psychology*, *5*, Article 1307, 1–6. https://doi.org/10.3389/fpsyg.2014.01307

Gould, R. (2012). *Graph theory*. *Dover books on mathematics*. Dover Publications.

Gravino, P., Servedio, V. D. P., Barrat, A., & Loreto, V. (2012). Complex structures and semantics in free word association. *Advances in Complex Systems*, *15*(03n04), Article 1250054. https://doi.org/10.1142/S0219525912500543

Gries, S. T. (2012). Corpus linguistics, theoretical linguistics, and cognitive/ psycholinguistics:: Towards more and more fruitful exchanges. In J. Mukherjee & M. Huber (Eds.), *Language and Computers: Vol. 75. Corpus Linguistics and Variation in English: Theory and Description* (pp. 41–63). Brill.

Gries, S. T. (2013). 50-something years of work on collocations: What is or should be next … *International Journal of Corpus Linguistics*, *18*(1), 137–166. https://doi.org/10.1075/ijcl.18.1.09gri

Gries, S. T. (2022a). What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies*, *5*(2), 171–205. https://doi.org/10.1075/jsls.21029.gri

Gries, S. T. (2022b). What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies*, *5*(1), 1–33. https://doi.org/10.1075/jsls.21028.gri

Gries, S. T., & Durrant, P. (2020). Analyzing Co-occurrence Data. In M. Paquot & S. T. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 141–159). Springer International Publishing AG. https://doi.org/10.1007/978-3-030-46216-1_7

Gries, S. T., & Ellis, N. C. (2015). Statistical Measures for Usage-Based Linguistics. *Language Learning*, *65*(S1), 228–255. https://doi.org/10.1111/lang.12119

Gries, S. T., & Stefanowitsch, A. (2010). Cluster Analysis and the Identification of Collexeme Classes. In S. Rice & J. Newman (Eds.), *Empirical and experimental methods in cognitive* (pp. 73–90). Center for the Study of Language and Information Publications.

Grieve, J. (2021). Observation, experimentation, and replication in linguistics. *Linguistics*, *59*(5), 1343–1356. https://doi.org/10.1515/ling-2021-0094

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. *Psychological Science*, *18*(12), 1069–1076. https://doi.org/10.1111/j.1467-9280.2007.02027.x

Gyllstad, H. (2014). Grammatical Collocation. In *The Encyclopedia of Applied Linguistics* (pp. 1–6). American Cancer Society. https://doi.org/10.1002/9781405198431.wbeal1444

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In G. Varoquaux, T. Vaught, & J. Millman (Chairs), *Proceedings of the 7th Python in Science Conference: SciPy2008,* Pasadena, CA USA).

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (3rd ed.). Arnold. http://www.uel.br/projetos/ppcat/pages/arquivos/RESOURCES/2004_HALLIDAY_MATTHIESSEN_An_Introduction_to_Functional_Grammar.pdf

Halliday, M. A. K., & Webster, J. (Eds.). (2005). *The collected works of M.A.K. Halliday: Vol. 6. Computational and quantitative studies*. Continuum.

Hamilton, C., Adolphs, S., & Nerlich, B. (2007). The meanings of 'risk': a view from corpus linguistics. *Discourse & Society*, *18*(2), 163–181. https://doi.org/10.1177/0957926507073374

Hanks, P. (2012). The Impact of Corpora on Dictionaries. In P. Baker (Ed.), *Contemporary studies in linguistics. Contemporary Corpus Linguistics* (1st ed., pp. 214–236). Bloomsbury.

Hasan, R. (2009). Rationality in Everyday Talk: From Process to System. In J. J. Webster (Ed.), *Collected Works of Ruqaiya Hasan: Vol. 2. Semantic variation: Meaning in society and in sociolinguistics* (1st ed., pp. 309–352). Equinox.

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, *45*(1). https://doi.org/10.1515/flin.2011.002

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic Representation of Action Words in Human Motor and Premotor Cortex. *Neuron*, *41*(2), 301–307. https://doi.org/10.1016/S0896-6273(03)00838-9

Hebb, D. O. (2002). *The Organization of Behavior* (1st ed.). Psychology Press. https://doi.org/10.4324/9781410612403

Herbst, T. (2018). Is Language a Collostructicon? A Proposal for Looking at Collocations, Valency, Argument Structure and Other Constructions. In P. Cantos Gómez & M. Almela-Sánchez (Eds.), *Quantitative Methods in the Humanities and Social Sciences: Vol. 10. Lexical collocation analysis: Advances and applications* (pp. 1–22). Springer. https://doi.org/10.1007/978-3-319-92582-0_1

Hitch, G. J., Hurlstone, M. J., & Hartley, T. (2022). Computational Models of Working Memory for Language. In J. W. Schwieter & Z. Wen (Eds.), *The Cambridge Handbook of Working Memory and Language* (pp. 143–174). Cambridge University Press.

Hitchcock, D. B. (2009). Yates and Contingency Tables: 75 Years Later. *Electronic Journal for History of Probability and Statistics*, *5*(2). https://www.jehps.net/Decembre2009/Hitchcock.pdf

Hodges, J. S. (1996). Statistical Practice as Argumentation: A Sketch of a Theory of Applied Statistics. In J. C. Lee, W. O. Johnson, & A. Zellner (Eds.), *Modeling and prediction: Honoring Seymour Geisser* (pp. 19–45). Springer.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language* (1st ed.). Taylor & Francis. https://ebookcentral.proquest.com/lib/gbv/detail.action?docID=178100

Hoffmann, T., & Trousdale, G. (2013). Construction Grammar: Introduction. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195396683.013.0001

Holten, D., & van Wijk, J. J. (2009). Force ‐ Directed Edge Bundling for Graph Visualization. *Computer Graphics Forum*, *28*(3), 983–990. https://doi.org/10.1111/j.1467-8659.2009.01450.x

Hu, Y. (2012). Algorithms for Visualizing Large Networks. In U. Naumann & O. Schenk (Eds.), *CRC Computational Science Series. Combinatorial scientific computing* (pp. 525–549). CRC Press.

Isbilen, E. S., & Christiansen, M. H. (2022). Statistical Learning of Language: A Meta-Analysis Into 25 Years of Research. *Cognitive Science*, *46*(9). https://doi.org/10.1111/cogs.13198

Ji, H [Hyngsuk], Lemaire, B., Choo, H., & Ploux, S. (2008). Testing the cognitive relevance of a geometric model on a word association task: A comparison of humans, ACOM, and LSA. *Behavior Research Methods*, *40*(4), 926–934. https://doi.org/10.3758/BRM.40.4.926

Jiang, J [Jingyang], Yu, W., & Liu, H. (2019). Does Scale-Free Syntactic Network Emerge in Second Language Learning? *Frontiers in Psychology*, *10*, 925. https://doi.org/10.3389/fpsyg.2019.00925

Johnson, S. J., Murty, M. R., & Navakanth, I. (2024). A detailed review on word embedding techniques with emphasis on word2vec. *Multimedia Tools and Applications*, *83*(13), 37979–38007. https://doi.org/10.1007/s11042-023-17007-z

Jurman, G., Visintainer, R., & Furlanello, C. (2011). An introduction to spectral distances in networks. In B. Apolloni (Ed.), *Frontiers in artificial intelligence and applications. Knowledge-based intelligent engineering systems: v. 226, Neural nets WIRN10: Proceedings of the 20th Italian Workshop on Neural Nets* (pp. 227–234). IOS Press.

Kallen, J. L. (2000). Two languages, two borders, one Island: Some linguistic and political borders in Ireland. *International Journal of the Sociology of Language*, *145*(1), 29–64.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, *31*(1), 7–15. https://doi.org/10.1016/0020-0190(89)90102-6

Kang, B.-M. (2018). Collocation and word association. *International Journal of Corpus Linguistics*, *23*(1), 85–113. https://doi.org/10.1075/ijcl.15116.kan

Kapatsinski, V. (2014). What is grammar like? A usage-based constructionist perspective. *Linguistic Issues in Language Technology*, *11*(1), 1–41. https://www.aclweb.org/anthology/2014.lilt-11.2.pdf

Karuza, E. A., Thompson-Schill, S. L., & Bassett, D. S. (2016). Local Patterns to Global Architectures: Influences of Network Topology on Human Learning. *Trends in Cognitive Sciences*, *20*(8), 629–640. https://doi.org/10.1016/j.tics.2016.06.003

Keating, G. D. (2013). Eye-tracking with text. In J. Jegerski & B. VanPatten (Eds.), *Second Language Acquisition Research Series. Research Methods in Second Language Psycholinguistics* (pp. 69–92). Taylor and Francis.

Kempe, V., Rookes, M., & Swarbrigg, L. (2013). Speaker emotion can affect ambiguity production. *Language and Cognitive Processes*, *28*(10), 1579–1590. https://doi.org/10.1080/01690965.2012.755555

Kenett, Y. N., Beaty, R. E., Silvia, P. J., Anaki, D., & Faust, M. (2016). Structure and flexibility: Investigating the relation between the structure of the mental lexicon, fluid intelligence, and creative achievement. *Psychology of Aesthetics, Creativity, and the Arts*, *10*(4), 377–388. https://doi.org/10.1037/aca0000056

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2015). *Statistics used in the Sketch Engine*. Lexical Computing Ltd.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. Bailey, & N. Hamilton-Smith (Eds.), *The Computer and Literary Studies: An associative thesaurus of English and its computer analysis.* Edinburgh University Press.

Klapaftis, I. P., & Manandhar, S. (2008). Word Sense Induction Using Graphs of Collocations. In M. Ghallab, C. D. Spyropoulos, N. Fakotakis, & N. Avouris (Eds.), *Frontiers in artificial intelligence and applications: Vol. 178, ECAI 2008: 18th European Conference on Artificial Intelligence, July 21-25, 2008, Patras, Greece.* IOS Press.

Kolesnikova, O. (2016). Survey of Word Co-occurrence Measures for Collocation Detection. *Computación Y Sistemas*, *20*(3), 327–344. https://doi.org/10.13053/cys-20-3-2456

Korkiakangas, T., & Lassila, M. (2018). Visualizing linguistic variation in a network of Latin documents and scribes. *Journal of Data Mining & Digital Humanities*, *Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages*, Article 4472. https://doi.org/10.46298/jdmdh.4472

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology. General*, *140*(1), 14–34. https://doi.org/10.1037/a0021446

Kovács, L., Bóta, A., Hajdu, L., & Krész, M. (2021). Networks in the mind – what communities reveal about the structure of the lexicon. *Open Linguistics*, *7*(1), 181–199. https://doi.org/10.1515/opli-2021-0012

Kovaleva, O., Romanov, A., Rogers, A., & Rumshisky, A. (2019). Revealing the Dark Secrets of BERT. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4364–4373). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1445

Krug, M. G., Schlüter, J., & Rosenbach, A. (2013). Introduction: Investigating language variation and change. In M. G. Krug & J. Schlüter (Eds.), *Research methods in language variation and change* (pp. 1–14). Cambridge University Press. https://doi.org/10.1017/CBO9780511792519.002

Kumova Metin, S., & Karaoğlan, B. (2011). Measuring Collocation Tendency of Words. *Journal of Quantitative Linguistics*, *18*(2), 174–187. https://doi.org/10.1080/09296174.2011.556005

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Lachnit, H. (2003). The Principle of Contiguity. In R. H. Kluwe, G. Lüer, & F. Rösler (Eds.), *Principles of Learning and Memory* (pp. 3–13). Birkhäuser Basel; Imprint; Birkhäuser. https://doi.org/10.1007/978-3-0348-8030-5_1

Lakoff, G. (1991). Cognitive versus generative linguistics: How commitments influence results. *Language & Communication*, *11*(1-2), 53–62. https://doi.org/10.1016/0271-5309(91)90018-Q

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Univ. of Chicago Press. http://www.loc.gov/catdir/description/uchi051/80010783.html

Landauer, T., & Dumais, S. T [S. T.] (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, *104*, 211–240. https://pcl.sitehost.iu.edu/rgoldsto/courses/concepts/landauer.pdf

Langacker, R. W. (1986). An Introduction to Cognitive Grammar. *Cognitive Science*, *10*(1), 1–40. https://doi.org/10.1016/S0364-0213(86)80007-6

Langacker, R. W. (1987). *Foundations of cognitive grammar* (Vol. 1). Stanford University Press.

Langacker, R. W. (1999). *Grammar and Conceptualization*. *Cognitive Linguistics Research: Vol. 14*. De Gruyter Mouton. https://doi.org/10.1515/9783110800524

Langacker, R. W. (2008). *Cognitive grammar: A basic introduction*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195331967.001.0001

Li, W [Wanyin], Lu, Q., & Xu, R. (2005). Similarity Based Chinese Synonym Collocation Extraction. *Computational Linguistics and Chinese Language Processing*, *10*(1), 123–144. http://140.109.19.106/clclp/v10n1/v10n1a6.pdf

Lindley, D. V., & Scott, W. F. (2018). *New Cambridge Statistical Tables*. Cambridge University Press. https://doi.org/10.1017/CBO9780511811906

Liu, H., & Li, W [WenWen] (2010). Language clusters based on linguistic complex networks. *Chinese Science Bulletin*, *55*(30), 3458–3465. https://doi.org/10.1007/s11434-010-4114-3

Locke, J. (1700). Chapter XXXIII. Of the Association of Ideas. In J. Locke (Ed.), *An essay concerning humane understanding: In four books* (4th ed., pp. 223–226).

Louwerse, M. M. (2021). Mapping out the road from corpus linguistics to psycholinguistics. *Revista Signos*, *54*(107), 971–984. https://doi.org/10.4067/S0718-09342021000300971

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). Compiling and analysing the Spoken British National Corpus 2014. *International Journal of Corpus Linguistics*, *22*(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Lu, J., & Si, Y.-W. (2020). Clustering-based force-directed algorithms for 3D graph visualization. *The Journal of Supercomputing*, *76*(12), 9654–9715. https://doi.org/10.1007/s11227-020-03226-w

Lyons, J. (1977). *Semantics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139165693

Mak, M. H. C., & Twitchell, H. (2020). Evidence for preferential attachment: Words that are more well connected in semantic networks are better at acquiring new links in paired-associate learning. *Psychonomic Bulletin & Review*, *27*(5), 1059–1069. https://doi.org/10.3758/s13423-020-01773-0

Matusevych, Y., & Stevenson, S. (2019). Analyzing and modeling free word associations. In C. Kalish, M. A. Rau, Zhu, Xiaojin, & T. T. Rogers (Chairs), *CogSci 2018,* Madison, WI, USA.

McConnell, K., & Blumenthal-Dramé, A. (2019). Effects of task and corpus-derived association scores on the online processing of collocations. *Corpus Linguistics and Linguistic Theory*, *aop*, 1–44. https://doi.org/10.1515/cllt-2018-0030

McCosker, A., & Wilken, R. (2014). Rethinking 'big data' as visual knowledge: the sublime and the diagrammatic in data visualisation. *Visual Studies*, *29*(2), 155–164. https://doi.org/10.1080/1472586X.2014.887268

McEnery, T., & Brezina, V. (2019). Collocations and colligations: Visualizing lexicogrammar. In B. Busse & R. Moehlig-Falke (Eds.), *Topics in English linguistics: Vol. 104. Patterns in language and linguistics: New perspectives on a ubiquitous concept* (pp. 97–124). De Gruyter Mouton. https://doi.org/10.1515/9783110596656-005

McEnery, T., & Brezina, V. (2022). *The Fundamental Principles of Corpus Linguistics*. Cambridge University Press.

McEnery, T., Tono, Y., & Xiao, R. (2006). *Corpus-based language studies: An advanced resource book. Routledge applied linguistics*. Routledge.

McIntosh, C. (Ed.). (2013). *Cambridge advanced learner's dictionary* (4. ed.). CUP.

McIntosh, C., Francis, B., & Poole, R. (2009). *Oxford collocations dictionary: For students of English* (2nd ed.). Oxford University Press.

McKeown, K. R., Smadja, F., & Hatzivassiloglou, V. (1996). Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, *22*(1), 1–38. https://doi.org/10.7916/D8C82M3R

McRae, K., Sa, V. R. de, & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99–130. https://doi.org/10.1037/0096-3445.126.2.99

Meara, P. (2009). *Connected Words* (Vol. 24). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.24

Mehler, A. (2008). Large text networks as an object of corpus linguistic studies. In A. Lüdeling (Ed.), *Handbücher zur Sprach- und Kommunikationswissenschaft. Corpus linguistics: An international handbook* (pp. 328–382). Mouton de Gruyter.

Menn, L., & Dronkers, N. (2017). *Psycholinguistics: Introduction and applications* (Second edition). Plural Publishing Inc.

Messaoudi, S. (2019). The Efficiency of Association Measures in Automatic Extraction of Collocations: Exclusivity and Frequency. *World Academy of Science, Engineering and Technology International Journal of Cognitive and Language Sciences*, *13*(4), 222–225. https://doi.org/10.5281/ZENODO.2643974

Metcalf, L., & Casey, W. (2016). *Cybersecurity and applied mathematics*. Syngress.

Michelbacher, L., Evert, S., & Schütze, H. (2011). Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory*, *7*(2). https://doi.org/10.1515/cllt.2011.012

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, January 16). *Efficient Estimation of Word Representations in Vector Space*. https://arxiv.org/pdf/1301.3781.pdf

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013, October 16). *Distributed Representations of Words and Phrases and their Compositionality*. arXiv:1310.4546

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191–1195. https://doi.org/10.1126/science.1152876

Mollin, S. (2009). Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory*, *5*(2), 175–200. https://doi.org/10.1515/CLLT.2009.008

Moon, R. (2015). Multi-word items. In J. R. Taylor (Ed.), *The Oxford handbook of the word* (pp. 120–140). Oxford University Press.

Morais, A. S., Olsson, H., & Schooler, L. J. (2013). Mapping the structure of semantic memory. *Cognitive Science*, *37*(1), 125–145. https://doi.org/10.1111/cogs.12013

Moretti, F. (2011). Network theory, plot analysis. *New Left Review*, *68*, 80–102.

Mukherjee, J. (2004). Corpus data in a usage-based cognitive grammar. In K. Aijmer & B. Altenberg (Eds.), *Language and Computers: Vol. 49. Advances in corpus linguistics: Papers from the 23rd international conference on English language research on computerized corpora (ICAME 23), Göteborg 22-26 May 2002 / edited by Karin Aijmer and Bengt Altenberg* (pp. 83–100). Brill. https://doi.org/10.1163/9789004333710_006

Navarro, E., Macnamara, B. N., Glucksberg, S., & Conway, A. R. A. (2020). What Influences Successful Communication? An Examination of Cognitive Load and Individual Differences. *Discourse Processes*, *57*(10), 880–899. https://doi.org/10.1080/0163853X.2020.1829936

Nazar, R. (2011). A statistical approach to term extraction. *International Journal of English Studies*, *11*(2), 159–182. https://doi.org/10.6018/ijes/2011/2/149691

Neergaard, K. D., Luo, J., & Huang, C.-R. (2019). Phonological network fluency identifies phonological restructuring through mental search. *Scientific Reports*, *9*(1), 15984. https://doi.org/10.1038/s41598-019-52433-w

Nelson, D. L., Canas, J., & Bajo, M. T. (1987). The effects of natural category size on memory for episodic encodings. *Memory & Cognition*, *15*(2), 133–140. https://doi.org/10.3758/bf03197024

Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition*, *28*(6), 887–899.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

Newmeyer, F. J. (2010). Formalism and functionalism in linguistics. *Wiley Interdisciplinary Reviews. Cognitive Science*, *1*(3), 301–307. https://doi.org/10.1002/wcs.6

Ni, C., Sugimoto, C. R., & Jiang, J [Jiepu] (2011). Degree, Closeness, and Betweenness: Application of group centrality measurements to explore macro-disciplinary evolution diachronically. In E. Noyons, P. Ngulube, & J. Leta (Eds.), *Proceedings of ISSI 2011 - the 13th International Conference of the International Society for Scientometrics and Informetrics* (pp. 605–617). Leiden University and University of Zululand.

Noble, H., & Heale, R. (2019). Triangulation in research, with examples. *Evidence-Based Nursing*, *22*(3), 67–68. https://doi.org/10.1136/ebnurs-2019-103145

Oakes, M. P. (2020). Statistical significance for measures of collocation strength. In G. Corpas Pastor & J.-P. Colson (Eds.), *IVITRA research in linguistics and literature: Vol. 24. Computational phraseology* (Vol. 24, pp. 189–206). John Benjamins Publishing Company. https://doi.org/10.1075/ivitra.24.10oak

O'Brien, B. A. (2014). The Development of Sensitivity to Sublexical Orthographic Constraints: An Investigation of Positional Frequency and Consistency Using a Wordlikeness Choice Task. *Reading Psychology*, *35*(4), 285–311. https://doi.org/10.1080/02702711.2012.724042

Ofcom. (2021). *Media Nations 2021: UK*. https://www.ofcom.org.uk/__data/assets/pdf_file/0023/222890/media-nations-report-2021.pdf

Ofcom. (2023a). *Media Nations 2023: UK*. https://www.ofcom.org.uk/__data/assets/pdf_file/0029/265376/media-nations-report-2023.pdf

Ofcom (Ed.). (2023b). *Technology tracker*. https://www.ofcom.org.uk/__data/assets/pdf_file/0016/262510/technology-tracker-2023-data-tables.pdf

Office for National Statistics. (2018). *Population estimates for the UK, England and Wales, Scotland and Northern Ireland: mid-2018*. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2018

Out, C., Goudbeek, M., & Krahmer, E. (2020). Do Speaker's emotions influence their language production? Studying the influence of disgust and amusement on alignment in interactive reference. *Language Sciences*, *78*, 101255. https://doi.org/10.1016/j.langsci.2019.101255

Oxley, J. (2014). Matroidal Methods in Graph Theory: Section 6.6. In J. L. Gross, J. Yellen, & P. Zhang (Eds.), *Discrete mathematics and its applications. Handbook of graph theory* (pp. 691–717). CRC Press.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources & Evaluation*, *44*, 137–158. https://doi.org/10.1007/s10579-009-9101-4

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In Q. Alessandro Moschitti, G. Bo Pang, & U. o. A. Walter Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Perruchet, P., & Peereman, R. (2004). The exploitation of distributional information in syllable processing. *Journal of Neurolinguistics*, *17*, 97–119. https://doi.org/10.1016/S0911-6044(03)00059-9

Perruchet, P., & Poulin-Charronnat, B. (2012). Word segmentation: Trading the (new, but poor) concept of statistical computation for the (old, but richer) associative approach. In P. Rebuschat & J. N. Williams (Eds.), *Statistical Learning and Language Acquisition* (pp. 119–143). De Gruyter.

Peruani, F. (2009). Advances in the Theory of Complex Networks. In N. Ganguly, A. Deutsch, & A. Mukherjee (Eds.), *MSSET - Modeling and Simulation in Science, Engineering & Technology. Dynamics On and Of Complex Networks: Applications to Biology, Computer Science, Economics, and the Social Sciences* (pp. 275–293). Birkhäuser Boston.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*(1), 29–46.

Pexman, P. M., Holyk, G. G., & Monfils, M.-H. (2003). Number-of-features effects and semantic processing. *Memory & Cognition*, *31*(6), 842–855. https://doi.org/10.3758/BF03196439

Phillips, M. (1985). *aspects of text structure: an investigation of the lexical organisation of text*. Elsevier.

Piaget, J. (1936). *La naissance de l'intelligence chez l'enfant*. Delachaux & Niestle.

Ploux, S., Boussidan, A., & Ji, H [Hyungsuk] (2010). The Semantic Atlas: an Interactive Model of Lexical Representation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp. 356–360). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/592_Paper.pdf

Ploux, S., Dabic, S., Paulignan, Y., Cheylus, A., & Nazir, T. A. (2012). Toward a neurolexicology: A method for exploring the organization of the mental lexicon by analyzing electrophysiological signals. *The Mental Lexicon*, *7*(2), 210–236. https://doi.org/10.1075/ml.7.2.04plo

Pradhan, P., C.U., A., & Jalan, S. (2020). Principal eigenvector localization and centrality in networks: Revisited. *Physica a: Statistical Mechanics and Its Applications*, *554*, 124169. https://doi.org/10.1016/j.physa.2020.124169

Rajeg, G. P. W. (2020). *collogetr* [Computer software]. Monash University. https://gederajeg.github.io/collogetr/index.html

Rapp, R. (2002). The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches. In *COLING '02, Proceedings of the 19th International Conference on Computational Linguistics - Volume 1* (pp. 1–7). Association for Computational Linguistics. https://doi.org/10.3115/1072228.1072235

Raviv, L., & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science*, *21*(4), 1-14. https://doi.org/10.1111/desc.12593

Rebuschat, P., & Williams, J. N. (2012). Introduction: Statistical learning and language acquisition. In P. Rebuschat & J. N. Williams (Eds.), *Statistical Learning and Language Acquisition* (pp. 1–12). De Gruyter.

Reitter, D., Keller, F., & Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cognitive Science*, *35*(4), 587–637. https://doi.org/10.1111/j.1551-6709.2010.01165.x

Rijk, M. de, & Mareček, D. (2020). Using Word Embeddings and Collocations for Modelling Word Associations. *Prague Bulletin of Mathematical Linguistics*, *114*(1), 35–57. https://doi.org/10.14712/00326585.002

Ritter, F. E., Tehranchi, F., & Oury, J. D. (2019). Act-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews. Cognitive Science*, *10*(3), e1488. https://doi.org/10.1002/wcs.1488

Rowlands, M. (2010). *The new science of the mind: From extended mind to embodied phenomenology*. MIT Press.

Rundell, M. (Ed.). (2007). *Macmillan English dictionary for advanced learners* (2nd ed.). Macmillan/A. & C. Black.

Rychlý, P. (2008). A Lexicographer-Friendly Association Score. In P. Sojka & A. Horák (Chairs), *RASLAN 2008: Recent Advances in Slavonic Natural Language Processing. Second Workshop on Recent Advances in Slavonic Natural Language Processing*. https://www.fi.muni.cz/usr/sojka/download/raslan2008/13.pdf

Sabidussi, G. (1966). The centrality of a graph. *Psychometrika*, *31*(4), 581–603. https://doi.org/10.1007/BF02289527

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1998). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27–52.

Sandoval, M., Patterson, D., Dai, H., Vance, C. J., & Plante, E. (2017). Neural Correlates of Morphology Acquisition through a Statistical Learning Paradigm. *Frontiers in Psychology*, *8*, Article 1234, 1–13. https://doi.org/10.3389/fpsyg.2017.01234

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *372*(1711), 1–15. https://doi.org/10.1098/rstb.2016.0049

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, *26*(1), 3–8. https://doi.org/10.1002/hipo.22523

Schmid, H.-J. (2000). *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. *Topics in English linguistics: Vol. 34*. De Gruyter. https://doi.org/10.1515/9783110808704

Schmück, H. (2024). *LLN: A pipeline for generating and comparing Large Linguistic Networks*. https://doi.org/10.17605/OSF.IO/8ANGY

Schmück, H., & Malone, D. (2023). *A pack of lone wolves*. https://doi.org/10.17605/OSF.IO/MW4JT

Scott, M. (2024). *WordSmith Tools version 9 (64 bit version)* [Computer software].

Sereno, S. C., Scott, G. G., Yao, B., Thaden, E. J., & O'Donnell, P. J. (2015). Emotion word processing: Does mood make a difference? *Frontiers in Psychology*, *6*, 1191. https://doi.org/10.3389/fpsyg.2015.01191

Seretan, V. (2011). *Syntax-Based Collocation Extraction*. *Text, Speech and Language Technology: Vol. 44*. Springer Science & Business Media.

Shan, T., Tang, W., Dang, X., Li, M., Yang, F., Xu, S., & Wu, J [Ji]. (2017, December 15). *Study on a Poisson's Equation Solver Based On Deep Learning Technique*. https://arxiv.org/pdf/1712.05559

Shanks, D. R. (1995). *The Psychology of Associative Learning*. Cambridge University Press. https://doi.org/10.1017/CBO9780511623288

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated

models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Sheridan, P., & Onodera, T. (2018). A Preferential Attachment Paradox: How Preferential Attachment Combines with Growth to Produce Networks with Log-normal In-degree Distributions. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-21133-2

Shukla, M., Gervain, J., Mehler, J., & Nespor, M. (2012). Linguistic constraints on statistical learning in early language acquisition. In P. Rebuschat & J. N. Williams (Eds.), *Statistical Learning and Language Acquisition* (pp. 171–202). De Gruyter. https://doi.org/10.1515/9781934078242.171

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, *81*, 105–120. https://doi.org/10.1016/j.jml.2015.02.001

Siew, C. S. Q. (2018). The orthographic similarity structure of English words: Insights from network science. *Applied Network Science*, *3*(1), 13. https://doi.org/10.1007/s41109-018-0068-1

Siew, C. S. Q., & Vitevitch, M. S. (2016). Spoken word recognition and serial recall of words from components in the phonological network. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *42*(3), 394–410. https://doi.org/10.1037/xlm0000139

Siew, C. S. Q., & Vitevitch, M. S. (2019). The phonographic language network: Using network science to investigate the phonological and orthographic similarity structure of language. *Journal of Experimental Psychology. General*, *148*(3), 475–500. https://doi.org/10.1037/xge0000575

Siew, C. S. Q., Wulff, D. U., Beckage, N. M., Kenett, Y. N., & Meštrović, A. (2019). Cognitive Network Science: A Review of Research on Cognition through the Lens of Network Representations, Processes, and Dynamics. *Complexity*, *2019*, 1–24. https://doi.org/10.1155/2019/2108423

Simmons, W. K., Hamann, S. B., Harenski, C. L., Hu, X. P., & Barsalou, L. W. (2008). fMRI evidence for word association and situated simulation in conceptual processing. *Journal of Physiology-Paris*, *102*(1-3), 106–119. https://doi.org/10.1016/j.jphysparis.2008.03.014

Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, *31*(4), 487–512. https://doi.org/10.1093/applin/amp058

Sinclair, J., & Carter, R. (2004). *Trust the text: Language, corpus and discourse*. Routledge.

Sinclair, J., & Coulthard, M. (1975). *Towards an analysis of discourse: The English used by teachers and pupils*. Oxford University Press.

Sinclair, J., Jones, S., Daley, R., & Krishnamurthy, R. (2004). *English collocation studies: The OSTI report. Corpus and discourse*. Continuum.

Siyanova-Chanturia, A., Conklin, K., Caffarra, S., Kaan, E., & van Heuven, W. J. B. (2017). Representation and processing of multi-word expressions in the brain. *Brain and Language*, *175*, 111–122. https://doi.org/10.1016/j.bandl.2017.10.004

Siyanova-Chanturia, A., & Martinez, R. (2015). The Idiom Principle Revisited. *Applied Linguistics*, *36*(5), 549–569. https://doi.org/10.1093/applin/amt054

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Sonbul, S. (2015). Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing. *Bilingualism: Language and Cognition*, *18*(3), 419–437. https://doi.org/10.1017/S1366728914000674

Sonbul, S., & Siyanova-Chanturia, A. (2021). Research on the on-line processing of collocation: Replication of Wolter and Gyllstad (2011) and Millar (2011). *Language Teaching*, *54*(2), 236–244. https://doi.org/10.1017/S0261444819000132

Stefanowitsch, A., & Gries, S. T. (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory*, *1*(1), 1–43.

Stella, M., Beckage, N. M., Brede, M., & Domenico, M. de (2018). Multiplex model of mental lexicon reveals explosive learning in humans. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-20730-5

Stevenson, A. (2010). *The Oxford dictionary of English* (3. ed.). OUP. http://www.oxfordreference.com/view/10.1093/acref/9780199571123.001.0001/acref-9780199571123

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78. https://doi.org/10.1207/s15516709cog2901_3

Stiller, J., Nettle, D., & Dunbar, R. I. M. (2003). The small world of shakespeare's plays. *Human Nature (Hawthorne, N.Y.)*, *14*(4), 397–408. https://doi.org/10.1007/s12110-003-1013-1

Stulpinaitė, M., Horbačauskiene, J., & Kasperavičienė, R. (2016). Issues in Translation of Linguistic Collocations: Lingvistinių kolokacijų vertimo ypatumai. *Studies About Languages*(29), 31–41. https://doi.org/10.5755/j01.sal.0.29.15056

Szewczyk, J. M., & Schriefers, H. (2018). The N400 as an index of lexical preactivation and its implications for prediction in language comprehension. *Language, Cognition and Neuroscience*, *33*(6), 665–686. https://doi.org/10.1080/23273798.2017.1401101

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 638–647. https://doi.org/10.1016/S0022-5371(75)80051-X

Teinonen, T., Fellman, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience*, *10*, Article 21, 1–8. https://doi.org/10.1186/1471-2202-10-21

Telesford, Q. K., Joyce, K. E., Hayasaka, S., Burdette, J. H., & Laurienti, P. J. (2011). The ubiquity of small-world networks. *Brain Connectivity*, *1*(5), 367–375. https://doi.org/10.1089/brain.2011.0038

Thiessen, E. D., & Saffran, J. R. (2007). Learning to Learn: Infants' Acquisition of Stress-Based Strategies for Word Segmentation. *Language Learning and Development*, *3*(1), 73–100. https://doi.org/10.1080/15475440709337001

Thurner, S. (2009). Statistical Mechanics of Complex Networks. In M. Dehmer & F. Emmert-Streib (Eds.), *Analysis of complex networks: From biology to linguistics*. Wiley-VCH. https://learning.oreilly.com/library/view/analysis-of-complex/9783527323456/07-chapter02.html#c02

Tomasello, R., Garagnani, M., Wennekers, T., & Pulvermüller, F. (2018). A Neurobiologically Constrained Cortex Model of Semantic Grounding With Spiking Neurons and Brain-Like Connectivity. *Frontiers in Computational Neuroscience*, *12*, Article 88, 1–17. https://doi.org/10.3389/fncom.2018.00088

Trautwein, J., & Schroeder, S. (2018). Orthographic Networks in the Developing Mental Lexicon. Insights From Graph Theory and Implications for the Study of Language Processing. *Frontiers in Psychology*, *9*, 1–12. https://doi.org/10.3389/fpsyg.2018.02252

Tribushinina, E., & Gillis, S. (2017). Advances and lacunas in usage-based studies of first language acquisition. In J. Evers-Vermeul & E. Tribushinina (Eds.), *Studies on language and acquisition: Volume 55. Usage-based approaches to language acquisition and language teaching* (pp. 13–114). De Gruyter Mouton.

Trustees of Princeton University (Ed.). (2024). *WordNet Documentation: wngloss(7WN)*. https://wordnet.princeton.edu/documentation/wngloss7wn

Tucker, B. V., & Ernestus, M. (2016). New Questions for the Next Decade. *The Mental Lexicon*, *11*(3), 375–400. https://doi.org/10.1075/ml.11.3.03tuc

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Uhrig, P., Evert, S., & Proisl, T. (2018). Collocation Candidate Extraction from Dependency-Annotated Corpora: Exploring Differences across Parsers and Dependency Annotation

Schemes. In P. Cantos Gómez & M. Almela-Sánchez (Eds.), *Quantitative Methods in the Humanities and Social Sciences: Vol. 10. Lexical collocation analysis: Advances and applications* (pp. 111–140). Springer. https://doi.org/10.1007/978-3-319-92582-0_6

Ulicheva, A., Harvey, H., Aronoff, M., & Rastle, K. (2020). Skilled readers' sensitivity to meaningful regularities in English writing. *Cognition*, *195*, Article 103810, 1–21. https://doi.org/10.1016/j.cognition.2018.09.013

Utsumi, A. (2015). A Complex Network Approach to Distributional Semantic Models. *PloS One*, *10*(8), e0136277. https://doi.org/10.1371/journal.pone.0136277

van Valin, R. D. (2003). Functional Linguistics. In M. Aronoff & J. Ress-Miller (Eds.), *The Handbook of Linguistics* (pp. 319–336). Blackwell Publishers Ltd. https://doi.org/10.1002/9780470756409.ch13

van Vu, D., & Peters, E. (2022). Incidental Learning of Collocations from Meaningful Input: A Longitudinal Study into Three Reading Modes and Factors that Affect Learning. *Studies in Second Language Acquisition*, *44*(3), 685–707. https://doi.org/10.1017/S0272263121000462

Vankrunkelsven, H., Verheyen, S., Storms, G., & Deyne, S. de (2018). Predicting Lexical Norms: A Comparison between a Word Association Model and Text-Based Word Co-occurrence Models. *Journal of Cognition*, *1*(1), Article 45, 1–14. https://doi.org/10.5334/joc.50

Veremyev, A., Semenov, A., Pasiliao, E. L., & Boginski, V. (2019). Graph-based exploration and clustering analysis of semantic spaces. *Applied Network Science*, *4*(1). https://doi.org/10.1007/s41109-019-0228-y

Vespignani, F., Canal, P., Molinaro, N., Fonda, S., & Cacciari, C. (2010). Predictive mechanisms in idiom comprehension. *Journal of Cognitive Neuroscience*, *22*(8), 1682–1700. https://doi.org/10.1162/jocn.2009.21293

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, *48*(4), 422–488. https://doi.org/10.1016/j.cogpsych.2003.09.001

Vilkaitė-Lozdienė, L. (2019). First steps towards the Lithuanian word association database. *Taikomoji Kalbotyra*, *12*, 226–258.

Vilkaitė-Lozdienė, L., & Conklin, K. (2021). Word order effect in collocation processing. *The Mental Lexicon*, *16*(2-3), 362–396. https://doi.org/10.1075/ml.20022.vil

Vitevitch, M. S. (2008). What Can Graph Theory Tell Us About Word Learning and Lexical Retrieval? *Journal of Speech, Language, and Hearing Research*, *51*(2), 408–422. https://doi.org/10.1044/1092-4388(2008/030)

Vitevitch, M. S., & Goldstein, R. (2014). Keywords in the mental lexicon. *Journal of Memory and Language*, *73*, 131–147. https://doi.org/10.1016/j.jml.2014.03.005

Vogel Sosa, A., & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: collocations involving the word. *Brain and Language*, *83*(2), 227–236. https://doi.org/10.1016/S0093-934X(02)00032-9

Wallis, S. (2013). z -squared: The Origin and Application of X^2. *Journal of Quantitative Linguistics*, *20*(4), 350–378. https://doi.org/10.1080/09296174.2013.830554

Ward, D., & Stapleton, M. (2012). Es are good: Cognition as enacted, embodied, embedded, affective and extended. In F. Paglieri (Ed.), *Advances in consciousness research (AICR): Vol. 86. Consciousness in interaction: The role of the natural and social context in shaping consciousness* (pp. 89–104). John Benjamins Pub. Co.

Watts, D. J., & Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Webb, S., & Kagimoto, E. (2009). The Effects of Vocabulary Learning on Collocation and Meaning. *TESOL Quarterly*, *43*(1), 55–77.

Webb, S., Newton, J., & Chang, A. (2013). Incidental Learning of Collocation. *Language Learning*, *63*(1), 91–120. https://doi.org/10.1111/j.1467-9922.2012.00729.x

Wermter, J. (2008). *Collocation and term extraction using linguistically enhanced statistical methods* [PhD].

Wermter, J., & Hahn, U. (2006). You can't beat frequency (unless you use linguistic knowledge). *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 785–792. https://doi.org/10.3115/1220175.1220274

Wernicke, C. (1908). The Symptom Complex of Aphasia. In A. C. Appleton (Ed.), *Diseases of the Nervous System* (pp. 265–324).

Wills, P., & Meyer, F. G. (2020). Metrics for graph comparison: A practitioner's guide. *PloS One*, *15*(2). https://doi.org/10.1371/journal.pone.0228728

Wilson, R. C., & Zhu, P. (2008). A study of graph spectra for comparing graphs and trees. *Pattern Recognition*, *41*(9), 2833–2841. https://doi.org/10.1016/j.patcog.2008.03.011

Wittgenstein, L., Anscombe, G. E. M., Hacker, P. M. S., & Schulte, J. (2009). *Philosophische Untersuchungen: Philosophical investigations* (Rev. 4th ed.). Wiley-Blackwell.

Wundt, W. (1897). *Grundriss der Psychologie* (2nd ed.). Wilhelm Engelmann.

Xiao, R., & McEnery, T. (2006). Collocation, Semantic Prosody, and Near Synonymy: A Cross-Linguistic Perspective. *Applied Linguistics*, *27*(1), 103–129. https://doi.org/10.1093/applin/ami045

Yamada, Y., & Neville, H. J. (2007). An ERP study of syntactic processing in English and nonsense sentences. *Brain Research*, *1130*(1), 167–180. https://doi.org/10.1016/j.brainres.2006.10.052

Yates, M. (2013). How the clustering of phonological neighbors affects visual word recognition. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(5), 1649–1656. https://doi.org/10.1037/a0032422

Yenduri, G., M, R., G, C. S., Y, S., Srivastava, G., Maddikunta, P. K. R., G, D. R., Jhaveri, R. H., B, P., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2023). *Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions*. https://doi.org/10.48550/arXiv.2305.10435

Yonelinas, A. P. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, *46*(3), 441–517. https://doi.org/10.1006/jmla.2002.2864

Zhao, Y.-T.-Y., Jia, Z.-Y., Tang, Y., Xiong, J. J., & Zhang, Y.-C. (2018). Quantitative learning strategies based on word networks. *Physica a: Statistical Mechanics and Its Applications*, *491*, 898–911. https://doi.org/10.1016/j.physa.2017.09.097

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.

Zortea, M., Menegola, B., Villavicencio, A., & Salles, J. F. de (2014). Graph analysis of semantic word association among children, adults, and the elderly. *Psicologia: Reflexão E Crítica*, *27*(1), 90–99. https://doi.org/10.1590/S0102-79722014000100011