universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Mapping Meaning Across Indo-European Languages: A Semantic Network Analysis Using Subs2Vec Embeddings

verfasst von | submitted by

Lale Tüver BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 647
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Digital Humanities
Betreut von | Supervisor:	Ass.-Prof. Mag. Mag. Dr. Andreas Baumann

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Andreas Baumann, for his invaluable guidance and support throughout the writing of this thesis. His rigorous scientific approach, thoughtful feedback, and attention to detail have helped me grow both as a researcher and as an academic writer. I am deeply grateful for his patience and guidance during this process.

Special thanks go to my friends Fırat, Rustem, Elena, and Manuel for their support throughout my studies, and for always bringing laughter and joy along the way. I would also like to thank my best friends, Nika and Karolina, for their endless support, kindness, and encouragement. Thank you for always listening to my ideas, for spending long hours with me in the library, and for believing in me throughout this process. Your friendship has been a constant source of joy, strength, and love, and I could not have asked for better friends.

My heartfelt thanks also go to my family, who have always supported me no matter what. Even across the distance, their love, encouragement, and interest in my work gave me strength and reminded me that I was never alone in this process. I feel incredibly lucky to have them by my side, even from afar.

Last, but never least, I would like to thank my partner, David. You have been my most honest critic, greatest supporter, and most patient companion throughout this process. Thank you for listening to my ideas about my research, giving me thoughtful advice, standing by me through the challenges of the final months, and always cheering me on. Your kindness, humor, and unwavering support helped me more than words can express. Thank you for believing in me and for reminding me, again and again, that I could finish this. This thesis would not have been possible without your support.

Abstract

Understanding patterns of semantic variation has emerged as a central research interest of computational linguistics. Within this research, the relationship between semantic organization and genealogical relatedness has received growing attention in cross-linguistic studies. Traditional comparative linguistics usually establishes language relationships through shared phonological, morphological, and lexical innovations, while many computational approaches rely on lexical overlap, cognates, or aligned multilingual material. However, less attention has been paid to whether languages can also be compared through the internal structure of semantic similarity, such as the way words form semantic neighborhoods, clusters, and larger network patterns.

This thesis uses semantic networks derived from word embeddings for cross-linguistic comparison within the Indo-European language family. In particular, it focuses on whether languages from the same subgroup exhibit similar semantic network properties and whether these properties can recover subgroupings resembling the established family tree. The analysis uses Subs2Vec word embeddings for 32 Indo-European languages from six branch categories: Germanic, Italic, Balto-Slavic, Indo-Iranian, Greek, and Albanian. For each language, cosine similarity between words was used to construct semantic networks with a percentile threshold method and a mutual k-nearest-neighbor method. Graph-theoretic properties describing degree structure, connectivity, clustering, community structure, and assortativity were compared across languages and branches, and used for hierarchical clustering.

The results show that semantic network properties reveal meaningful cross-linguistic patterns, but only partially correspond to genealogical classification. Some branches show a relatively compact structural profile, whereas others display stronger internal variation and language-specific outliers. Although hierarchical clustering recovers several local similarities between closely related languages, these do not always develop into stable higher-level branch clusters. These findings suggest that embedding-derived semantic networks capture partial genealogical signal, especially at the level of close language pairs and smaller subgroups. The thesis contributes a graph-based computational approach to cross-linguistic semantic comparison and language relatedness.

Kurzfassung

Die Analyse semantischer Variation hat sich zu einem zentralen Forschungsinteresse der Computerlinguistik entwickelt. In sprachvergleichenden Studien richtet sich dabei zunehmend die Aufmerksamkeit auf die Beziehung zwischen semantischer Organisation und genealogischer Verwandtschaft von Sprachen. Während traditionelle vergleichende Linguistik Sprachverwandtschaften meist anhand gemeinsamer phonologischer, morphologischer und lexikalischer Entwicklungen bestimmt, stützen sich viele rechnergestützte Ansätze auf lexikalische Überschneidungen, Kognaten oder mehrsprachig ausgerichtete Datensätze. Ob Sprachen auch über die interne Struktur semantischer Ähnlichkeit verglichen werden können, also anhand der Bildung semantischer Nachbarschaften, Cluster und größerer Netzwerkmuster, wurde bisher noch nicht ausreichend untersucht.

Diese Arbeit nutzt semantische Netzwerke, die aus Word Embeddings abgeleitet wurden, um indoeuropäische Sprachen zu vergleichen. Im Fokus steht die Frage, ob Sprachen derselben Untergruppe ähnliche semantische Netzwerkeigenschaften aufweisen und ob sich daraus Gruppierungen ableiten lassen, die dem etablierten indoeuropäischen Stammbaum ähneln. Die Analyse basiert auf Subs2Vec Word Embeddings für 32 indoeuropäische Sprachen aus sechs Zweigen, nämlich Germanisch, Italisch, Balto-Slawisch, Indo-Iranisch, Griechisch und Albanisch. Für jede Sprache wurde die Kosinus-Ähnlichkeit zwischen Wörtern berechnet. Daraus wurden semantische Netzwerke mithilfe von perzentilbasierten Methode und der mutual k-nearest-neighbor-Methode konstruiert. Anschließend wurden graphentheoretische Eigenschaften wie Degree-Struktur, Konnektivität, Clustering, Community-Struktur und Assortativität zwischen Sprachen und Zweigen verglichen und für hierarchisches Clustering verwendet.

Die Ergebnisse zeigen, dass semantische Netzwerkeigenschaften sprachübergreifende Muster sichtbar machen, diese jedoch nur teilweise mit der genealogischen Klassifikation übereinstimmen. Einige Zweige weisen ein vergleichsweise kompaktes strukturelles Profil auf, während andere stärkere interne Variation und sprachspezifische Ausreißer zeigen. Das hierarchische Clustering rekonstruiert mehrere lokale Ähnlichkeiten zwischen eng verwandten Sprachen, bildet jedoch nicht immer stabile übergeordnete Zweig-Cluster. Die Ergebnisse deuten darauf hin, dass embedding-basierte semantische Netzwerke ein partielles genealogisches Signal erfassen, besonders auf der Ebene enger Sprachpaare und kleinerer Untergruppen. Damit leistet die Arbeit einen Beitrag zur Entwicklung rechnergestützter Ansätze für den sprachübergreifenden Vergleich semantischer Strukturen und sprachlicher Verwandtschaft.

Contents

List of Tables

List of Figures

List of Algorithms

1. Introduction

Language is not only a system of forms, but also a system of meaning. Words do not exist in isolation, but are embedded in broader structure of semantic relations, associations, and contrasts that together shape how meaning is organized within a language. Traditional work in lexical semantics has long emphasized that lexical meaning emerges through relations between words, whether in the form of synonymy, antonymy, hierarchical relations, or semantic fields. Such relations have been central for understanding how vocabularies are internally structured and how concepts relate to one another within a linguistic system (Murphy, 2003, 2010; Cruse, 1986; Lehrer, 1974).

This connection between vocabulary and meaning is also relevant for cross-linguistic comparison. In comparative linguistics, languages are often compared through lexical items that express similar or equivalent concepts across languages. Such comparison relies on conceptual similarity, since words can be treated as comparable when they refer to similar meanings, even if their forms differ. However, lexical comparison is not sufficient enough for establishing historical relatedness. In historical comparative linguistics, vocabulary is interpreted together with regular phonological correspondences, shared morphological innovations, and other structural evidence (Bowern and Evans, 2015).

Languages can therefore be compared both as systems of meaning and as historically related structural systems. In the case of Indo-European language family, linguists have long relied on comparative methods, using shared phonological, morphological and lexical innovations to build subgroupings and historical relationships. This has made Indo-European linguistics one of the most thoroughly developed areas of historical comparative research. Because the family is well documented, with extensive textual traditions, dictionaries, annotated corpora, and scholarly descriptions, it is one of the best described language families in the world. Its considerable internal diversity makes it also a good setting for exploring how far different kinds of linguistic structure relate to genealogical relationships (Clackson, 2007; Bowern and Evans, 2015).

In recent decades, computational approaches have opened new ways of studying semantic structure in language. In particular, distributional semantics and word embeddings make it possible to model lexical meaning from patterns of language use by representing words as vectors in a semantic space derived from corpora. Because such representations are learned directly from linguistic data, they can capture semantic and syntactic regularities without relying on manually annotated features. They also make semantic similarity computationally measurable through methods such as cosine similarity, allowing meaning to be represented not as a fixed set of discrete categories but as a system of graded similarities that emerge from distributional behavior and patterns of usage (Turney and Pantel, 2010; Lenci and Sahlgren, 2023; Torregrossa et al., 2021).

A complementary perspective is also provided by graph theory or network-based approaches. In such models, words or concepts can be represented as nodes and their relations as links, making it possible to study semantic organization at the level of the language system as a whole. Semantic networks allow researchers to analyze broader structural properties such as clustering, connectivity, and community structure. In this way, they make it possible to ask not only which words are similar, but also how semantic neighborhoods are organized and how larger relational patterns emerge across the lexicon.(Steyvers and Tenenbaum, 2005; Veremyev et al., 2019).

These developments show that semantic structure can now be studied from different perspectives, including lexical semantic theory, historical language comparison, vector and graph based modeling. Against this background, the next question is how far these perspectives can be brought together for cross-linguistic research, especially when the goal is to compare not only individual lexical items, but the broader organization of semantic structure across languages.

1.1. Problem Statement

Despite major advances in both computational linguistics and historical language comparison, the relationship between semantic structure and genealogical relatedness remains relatively underexplored. Traditional Indo-European subgroupings relies mainly on phonological, morphological and lexical innovations established through the comparative method (Clackson, 2022; Bowern and Evans, 2015). In this tradition, genealogical relationships are inferred from shared historical developments rather than from broader structural properties of language systems. Computational approaches to Indo-European comparison have expanded the range of evidence used for language comparison, including cognate sets, lexical lists, and translation-based corpora. These studies have shown that language similarity can be modeled quantitatively, but they generally focused on lexical overlap or aligned multilingual material (Rabinovich et al., 2017; Serva and Petroni, 2008; Schrader and Gultepe, 2023).

A parallel development has taken place in computational semantics. Word embeddings have made it possible to model lexical meaning from usage patterns in corpora, while semantic network research has shown that lexical and conceptual structure can be represented and analyzed as graphs. Yet, these two strands of research have rarely been combined in a way that allows languages to be compared through the the graph-theoretic properties of embedding-derived semantic networks. In many embedding based studies, languages are compared only through semantic spaces, whereas graph based studies have more often modeled language structure through co-occurrence networks, lexical databases, or explicitly defined semantic relations (Thompson et al., 2018; Lewis et al., 2023; Gao et al., 2014; Budel et al., 2023).

A further limitation concerns the data used in multilingual comparison. Recent studies have shown that multilingual corpora and embedding-based methods can be used to compare languages or even reconstruct phylogenetic relations, especially when parallel corpora, such as Bible translations, are used (Schrader and Gultepe, 2023). However, such

approaches usually depend on aligned content and on the assumption that translation provides sufficiently comparable semantic material across languages. Even when network methods are used successfully, as in the case of Schrader and Gultepe (2023) network-partitioning approach over document vectors, the underlying comparison still depends on parallel translations and matched multilingual structure. As a result, these studies often capture translation effects, lexical overlap, or cross-linguistic alignment more directly than the broader internal semantic organization of each language (Thompson et al., 2018; Schrader and Gultepe, 2023; Serva and Petroni, 2008).

The present study addresses this gap by asking whether the overall structure of semantic networks derived from word embeddings can be used for cross-linguistic comparison within the Indo-European family. More specifically, it investigates whether graph-theoretic properties of semantic networks can reveal meaningful similarities and differences between languages, and whether such structural properties can produce subgroupings that resemble established Indo-European family branches. This study is motivated by the fact that word embeddings provide a quantitative representation of semantic similarity, while graph-based analysis makes it possible to move beyond pairwise similarity and examine broader patterns of semantic organization.

The motivation of this approach is twofold. First, it offers a way of comparing languages that does not depend on manually crafted linguistic features, explicit cognate coding or direct translation alignment. Second, it shifts the focus from isolated lexical correspondences to the system-level organization of semantic structure. In this way, the study explores whether semantic-network topology can provide an additional perspective on language relatedness within the Indo-European family, while also clarifying the extent to which such structural patterns align with established subgroupings.

1.2. Research Question

Against this background, the main research question of the present study is as follows:

> *To what extent do Indo-European languages show similar or different structural properties in semantic networks derived from word embeddings, and can these properties be used for cross-linguistic comparison?*

This question can be divided into two more specific subquestions:

1. *Do languages from the same Indo-European subgroup exhibit similar semantic network properties?*
2. *Can semantic network properties be used to recover subgroupings that resemble the established Indo-European family tree?*

To address these questions, the thesis combines distributional semantics with graph-based analysis. Word embeddings are used to construct semantic similarity networks for a selection of Indo-European languages, and graph-theoretic measures are then extracted in

order to capture broader structural properties of semantic organization. These structural profiles are subsequently compared using hierarchical clustering to examine whether the semantic network structure yields groupings that resemble established Indo-European subgroupings.

1.3. Roadmap

The remainder of the thesis is structured as follows. Chapter 2 provides the theoretical background relevant to the study. It introduces key concepts from lexical semantics, outlines the classification of Indo-European language family, and discusses distributional semantics and word embeddings as the basis for modeling semantic similarity. The chapter also introduces semantic networks and the network properties used to describe their structure. Chapter 3 describes the data and methodology used in the analysis. It outlines the selection, the preprocessing of the embedding data, the calculation of cosine similarities, the construction of semantic similarity networks using two different methods, and the extraction of network properties for comparison. Chapter 4 presents the results of the network analysis and hierarchical clustering. Chapter 5 discusses these findings in relation to the research questions, addresses the limitations of the study, and outlines directions for future research. Finally, Chapter 6 summarizes the main conclusions of the thesis.

2. Theoretical Background

> "When I use a word," Humpty Dumpty said in a rather scornful tone, "it means just what I choose it to mean—neither more nor less."
> "The question is," said Alice, "whether you can make words mean so many different things."
>
> Lewis Carroll, *Through the Looking-Glass*

2.1. Lexical Semantics and the Structure of Meaning

The lexicon has come to play an increasingly central role in linguistic theories of meaning and language structure. In a broad sense, lexical semantics is concerned with how words convey meaning and how those meanings are structured within the vocabulary of a language. Rather than existing as isolated units, lexical items form interconnected systems in which meaning emerges through relationships with other words or the situations in which they are used. Early work in lexical semantics emphasized that the lexicon should be understood as an organized semantic system, in which the meaning of individual words is partly determined by their relations to other lexical items (Cruse, 1986; Lyons, 1977). These relationships form patterns of similarity, contrast, and inclusion that shape how speakers categorize and interpret concepts through language.

This view also suggests that word meaning cannot be reduced to a dictionary definition alone. As Nunberg (1978, p. iii) argues, knowledge of word meaning is tied to the "collective beliefs of the speech community", and therefore interacts with conventions of use, shared assumptions, and pragmatic interpretation. In a similar way, Murphy (2003) notes that speakers recognize particular senses of words not only because of formal semantic structure, but also because they understand the conventional and contextual ways in which words are used. Therefore, lexical meaning should be understood as both relational and context sensitive. Within this broader perspective, semantic relations provide one of the main ways of describing how the lexicon is structured.

2.1.1. Semantic Relations

One of the central insights of lexical semantics is that words are systematically connected through semantic relations. These relations help organize the lexicon into a coherent system, rather than a simple list of unrelated items. As Lyons (1977, p. 3) notes, the meaning of words forms "a network of similarities and differences", which means that

individual lexical items can rarely be understood in complete isolation from others. Words are linked through patterns of resemblance, opposition, inclusion, and association.

This relational view of meaning shows that vocabulary reflects conceptual structures. Some words are related because they overlap in meaning, or because they stand in contrast to one another, or they belong to broader systems of classification or part-whole organization. These lexical semantic relations provide an important framework for understanding how meaning is structured within the lexicon (Murphy, 2010, 2003).

A useful way to organize these semantic relations is to distinguish between non-hierarchical and hierarchical relations. Non-hierarchical relations include synonymy and antonymy, which are based primarily on similarity and contrast. Hierarchical relations include hyponymy, hypernymy, and meronymy, which involve asymmetrical relations of inclusion or part-whole structure. All these relations together show that lexical meaning is not isolated or random, but systematically organized through different types of semantic connections.

Non-hierarchical relations

One of the most widely discussed non-hierarchical relations is synonymy, which refers to the relationships between lexical items with similar or closely related meanings. Words such as *begin* and *start* or *big* and *large* are often treated as synonyms because they can occur in similar contexts and express comparable meanings. However, complete synonymy is rare. Words that appear synonymous usually differ in some respect, such as collocation, register, or pragmatic use (Murphy, 2010, 2003). For example, while *begin* and *start* are often interchangeable, one can *start a car* but not *begin a car*. Such cases show that semantic similarity is often a matter of degree rather than absolute equivalence. Synonymy, therefore, highlights fine-grained distinctions that structure the lexicon (Murphy, 2010, 2003; Cruse, 1986; Lyons, 1977).

Another important non-hierarchical relation is antonymy, which describes oppositional relationships between words. Antonymy captures how languages express contrast between concepts, as in pairs such as *hot* and *cold*, *young* and *old*, or *alive* and *dead*. These oppositions are not all of the same type. Gradable antonyms, such as *hot* and *cold*, occupy opposite ends of a scale and allow intermediate values such as *warm* or *cool*. Complementary antonyms, such as *alive* and *dead*, represent mutually exclusive states (Murphy, 2010; Cruse, 1986). Antonymy shows that meaning can be structured through contrast and that lexical categories can also be defined in relation to oppositional terms.

Hierarchical relations

Hierarchical relations differ from synonymy and antonymy because they involve asymmetrical relations between lexical items. Among hierarchical relationships, hyponymy and hypernymy are especially important. These relations describe inclusion between meanings, where one term is more general and the other more specific. A hypernym is a broader category term, while a hyponym is a more specific instance within a category. For example, *vehicle* is a hypernym, while *car* is hyponym of *vehicle*. This inclusion

relation is asymmetrical, meaning that every *car* is a *vehicle*, but not every *vehicle* is a *car*. Hyponymy and hypernymy therefore form taxonomic structures, allowing words to be organized into semantic hierarchies (Murphy, 2003; Jurafsky and Martin, 2026; Cruse, 1986). Such relations are particularly useful because they show how lexical meaning reflects broader systems of conceptual classification.

A related but not entirely hierarchical relation is meronymy, which refers to a part-whole structure. In meronymic relations, one word denotes a part of something denoted by another word. For example, *wheel* is part of a *car*, *branch* is part of a *tree*, *chapter* is part of a *book* (Jurafsky and Martin, 2026). Unlike hyponymy, meronymy is not based on category inclusion, such as a *wheel* is not kind of a *car*, but rather one of its components. Meronymy therefore shows that lexical structure can reflect not only taxonomic classification but also the internal composition of objects and entities (Cruse, 1986; Jurafsky and Martin, 2026). Together, hyponymy, hypernymy, and meronymy illustrate that lexical meaning is structured through several types of asymmetrical relations.

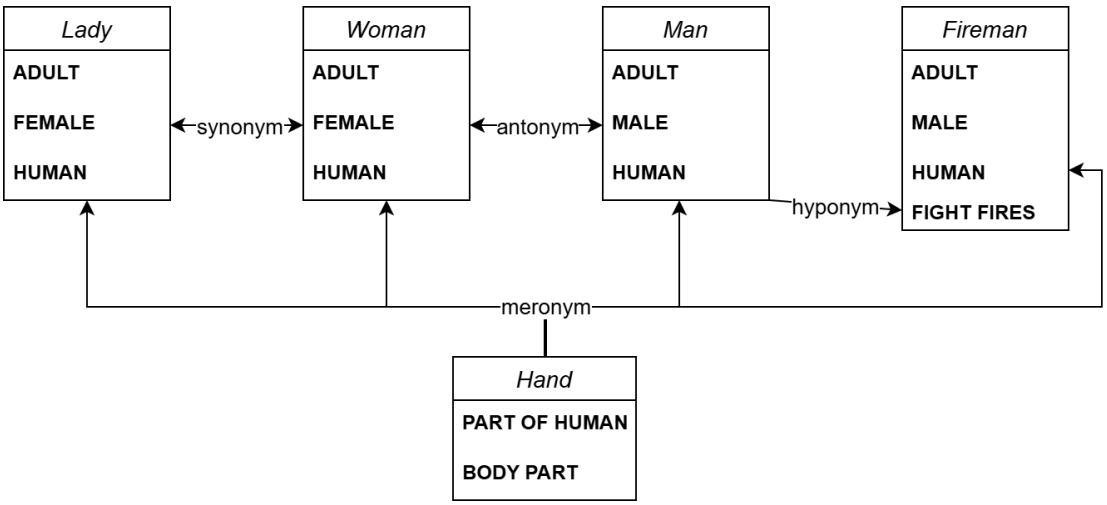

Figure 2.1.: Schematic representation of selected semantic relations between word meanings, including synonymy, antonymy, hyponymy, and meronymy. Adapted from Murphy (2010, p. 125) and expanded to include meronymy.

These distinctions can also be illustrated through the Figure 2.1. In this representation, lexical meanings are represented as bundles of semantic features. *Lady* and *woman* share the features **HUMAN**, **ADULT**, and **FEMALE**, which explains their near-synonymous relationship. *Woman* and *man*, by contrast differ in the feature **FEMALE** versus **MALE**, which captures antonymy. Similarly, *fireman* shares several features with *man*, but includes the additional component **FIGHTS FIRES**, representing a hyponymic relation. The figure also includes *hand* as an example of meronymy, since it denotes a body part that stand in a part-whole relation to a human being. Although such analyses simplify actual language use, they are useful for illustrating how semantic similarity, contrast, and asymmetrical relations can be represented within a framework.

2.1.2. Semantic Fields and Categorization

While lexical semantic relations describe structured connections between individual words, lexical organization also extends beyond isolated word pairs. Semantic fields refer to broader conceptual domains in which groups of related words are organized around shared areas of experience. Within such fields, the meaning of a lexical item is partly shaped by its position in relation to other items in the same domain (Lehrer, 1974; Lyons, 1977; Murphy, 2003).

This broader level of lexical organization can be seen in domains such as kinship, color, food, or time. Kinship terms such as *mother*, *father*, *sister* and *brother*, for example, do not function as isolated labels. Rather, they form part of a larger conceptual system in which each term is interpreted in relation to the others. The same applies to other areas of vocabulary. For example, color terms gain meaning through their relation to neighboring and contrasting color categories, while terms for food or time are embedded in culturally and linguistically structured domains. Semantic fields, therefore, extend the analysis of lexical semantics from relations between individual words to larger conceptual systems within the lexicon (Murphy, 2010, 2003; Jurafsky and Martin, 2026; Berlin and Kay, 1991; Wallace, 1970).

The existence of semantic fields also suggest that lexical categories are not always sharply bounded or internally uniform. This point is developed further in prototype theory, which offers an account of the internal structure of categories. Rather than assuming that categories consist of equally representative members defined by fixed boundaries, prototype theory assumes that many categories are organized around more central and more peripheral members (Rosch, 1978; Geeraerts, 2006). As Geeraerts (2006) argues, prototypical categories are not defined by a single set of necessary and sufficient features, but instead tend to show family resemblance structure, degrees of representativeness, and often fuzzy boundaries. A well-known example is the category *bird*, in which robins and sparrows are often judged to be better examples of the category *bird* than penguins or ostriches, even though all belong to the same category (Rosch, 1978). In lexical semantics, this suggests that many lexical categories are structured through degrees of typicality, centrality, and family resemblance rather than through fixed boundaries (Geeraerts, 2006; Murphy, 2003). Semantic fields and prototype theory together show that lexical meaning is structured not only by the relationships among individual words but also by broader conceptual domains and even internal category structures.

2.1.3. Resources for Lexical Semantics

Theoretical work in lexical semantics also inspired the development of formal lexical resources that explicitly encode semantic relations. One of the most influential examples is WordNet, a large lexical database of English developed at Princeton University. In WordNet, nouns, verbs, adjectives, and adverbs are organized into sets of cognitive synonyms, known as synsets, each of which represents a distinct lexicalized concept. These synsets are linked through a range of semantic and lexical relations, including synonymy, antonymy, hypernymy, hyponymy, and meronymy. In this way, WordNet offers

a structured representation of the relational organization of the lexicon and can also be understood as an early example of a semantic network, since lexical meaning is modeled as a system of interconnected concepts rather than as isolated dictionary entries (Miller, 1995; Fellbaum, 1998; Brenndoerfer, 2025; Jurafsky and Martin, 2026). Since its creation, the WordNet model has been extended beyond English.The Open Multilingual Wordnet currently brings together 60 wordnets for 49 languages, with additional wordnets being added as they become available (Bond et al., 2016).

Other resources approach cross-linguistic lexical comparison from the perspective of concepts, semantic domains, or patterns of lexical association. Concepticon provides a standardized reference system for linking concept lists used in linguistic research, allowing concepts such as hand, water, mother, and tree to be compared across different datasets and studies (List et al., 2016). CLICS, the Database of Cross-Linguistic Colexifications, provides a complementary perspective by documenting cases in which the same lexical form expresses multiple meanings across languages, thereby revealing recurrent semantic associations between concepts (CLICS, 2026). NorthEuraLex covers 1,016 concepts across 107 languages, with a particular focus on Indo-European and Uralic languages, and makes it possible to compare lexical items across broader domains such as kinship, body parts, space, and time (Dellert et al., 2020).

These resources show that lexical meaning can be represented not only descriptively, but also formally and computationally. WordNet models meaning through explicitly encoded semantic relations, while other mentioned resources support cross-linguistic comparison through concepts and semantic domains. This is especially relevant for the present thesis, which also approaches lexical meaning as structured and relational. However, instead of relying on manually constructed lexical databases or predefined concept lists, the thesis investigates whether comparable semantic structures can be derived from distributional data and represented as semantic similarity networks. Since lexical organization may vary across languages while still reflecting historical relatedness, this approach raises the question of whether related languages show similar semantic network properties. The following section therefore turns to the Indo-European language family, which provides the comparative framework for this study.

2.2. Indo-European Language Family and Its Classification

Among the 143 language families listed by Ethnologue, six stand out as the world's major language families and together account for roughly five-sixths of the global population (Eberhard et al., 2026). Of these, the Indo-European language family is the largest by speaker population. Eberhard et al. (2026) reports that Indo-European languages account for approximately 3.3 billion first language speakers in 2025, while Clackson (2007) notes an earlier estimate of around 2.3. billion first language speakers in the early 2000s.

Beyond its demographic importance, Indo-European is also the most thoroughly studied language family in historical linguistics. One reason for this scholarly attention is the richness of the available evidence. Several Indo-European branches, including Greek, Latin and Sanskrit, are documented over long historical period and are supported by extensive

textual and grammatical traditions. This rich body of evidence made Indo-European studies central to the development of historical-comparative linguistics, including methods of linguistic reconstruction, sound change analysis, and subgrouping (Clackson, 2007). For this reason, introducing the Indo-European family is not only relevant to the languages analyzed in this thesis, but also important for understanding the historical comparative framework against which the semantic network results are interpreted.

2.2.1. Indo-European Relatedness and the Family Tree Model

The recognition of Indo-European relatedness is generally associated with Sir William Jones. In his well-known statement, Jones (1786, p. 26) pointed to the striking similarities between Sanskrit, Greek, and Latin, especially in their verbal roots and grammatical structures, and argued that such correspondences could not have been "produced by accident". He famously concluded that "no philologer could examine them all three, without believing them to have sprung from some common source, which, perhaps, no longer exists" (Jones, 1786, p. 26). In broad terms, this basic criterion has remained central to Indo-European classification: languages are identified as Indo-European because they display close similarity in core verbal roots and morphological paradigms than could reasonably be attributed to chance. Jones's "common source" later came to be understood as Proto-Indo-European , the reconstructed ancestral languages of the family.

At the same time, this early recognition remains largely intuitive. As Lehmann (1993, p. 3) notes, when Jones made his pronouncement in 1786, there was "almost no experience in two fundamental prerequisites of comparative linguistics: understanding of sound systems and historical relationships". There was, and still is, no absolute checklist that automatically determines Indo-European membership; rather, the evidence must be strong enough to convince both the individual scholar and the wider research community (Clackson, 2007). The problem for nineteenth-century linguistics was therefore not only to observe similarity, but to establish a method for evaluating it.

Interestingly, the recognition of individual Indo-European languages preceded a full understanding of the Indo-European language family as a whole. In the early nineteenth century, scholars focused more on identifying what are now known as subgroupings, such as Germanic and Slavic, rather than building a family tree. Clackson (2022) notes that early research treated these language groups as largely independent units rather than components of a single, structured genealogical system.

The family tree model is the most common way of representing genealogical relationships between languages. Usually associated with August Schleicher, the model treats languages as descendants of earlier parents stages (Schleicher, 1861). In this view, a language family develops through successive branching. An ancestral language gives rise to later branches, and these branches may in turn divide into further subgroups and individual languages.

Figure 2.2, gives a simplified illustration of this principle. Proto-Indo-European appears as the common ancestral source, while branches such as *Germanic* and *Italic* represent later subgroups within the family. Individual languages, such as English, German, Italian, and French, are placed within these branches. The important point is that these subgroupings indicate relative historical closeness. Languages within the same subgroup are assumed to

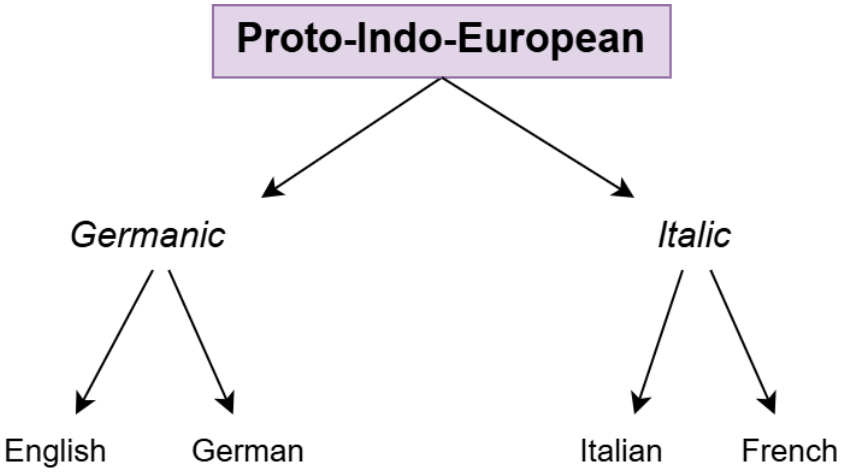

Figure 2.2.: Simplified schematic representation of the family-tree model using selected Indo-European branches and languages.

share a more recent common ancestor with each other than with languages outside of it. In some cases, this common ancestor is historically attested; in others, it is reconstructed through the comparative method (Clackson, 2007).

Schleicher (1861)'s own Indo-European tree, show in Figure 2.3, is an early and influential example of this genealogical way of representing language relationships. His diagram differs visually from modern tree diagrams, as the parent language is placed on the left, and the daughter branches extend toward the right. Notably, Schleicher chose to label the branches themselves rather than the nodes. While the specific details of his classification are no longer fully accepted [1], the model marked a critical turning point. It established genealogical representation as a central problem in Indo-European studies and demonstrated that the structure of a language tree depends on which shared developments are deemed historically significant.

2.2.2. Neogrammarian Regularity and Subgrouping Criteria

The first scholars to address Indo-European subgrouping in a more explicit methodological way were the Neogrammarians (*Junggrammatiker*), a group of scholars centered in Leipzig during the 1870s. They introduced more precise principles and procedures for establishing linguistic relationships, moving beyond the more intuitive approaches of their predecessors. One of their central tenet, emphasized by Leskien (1876) and later formalized by the so-called "Neogrammarian Manifesto"(Osthoff and Brugmann, 1878), was the regularity of sound change. This principle holds that sound change affects all relevant forms within the

[1] Schleicher's original model proposed an early split that grouped Germanic, Baltic, and Slavic together. This was largely attributed to the distribution of the *-m* and *-bh* markers in plural noun endings, which is a distinction that modern scholarship treates with much greater caution (Schleicher, 1861; Clackson, 2007).

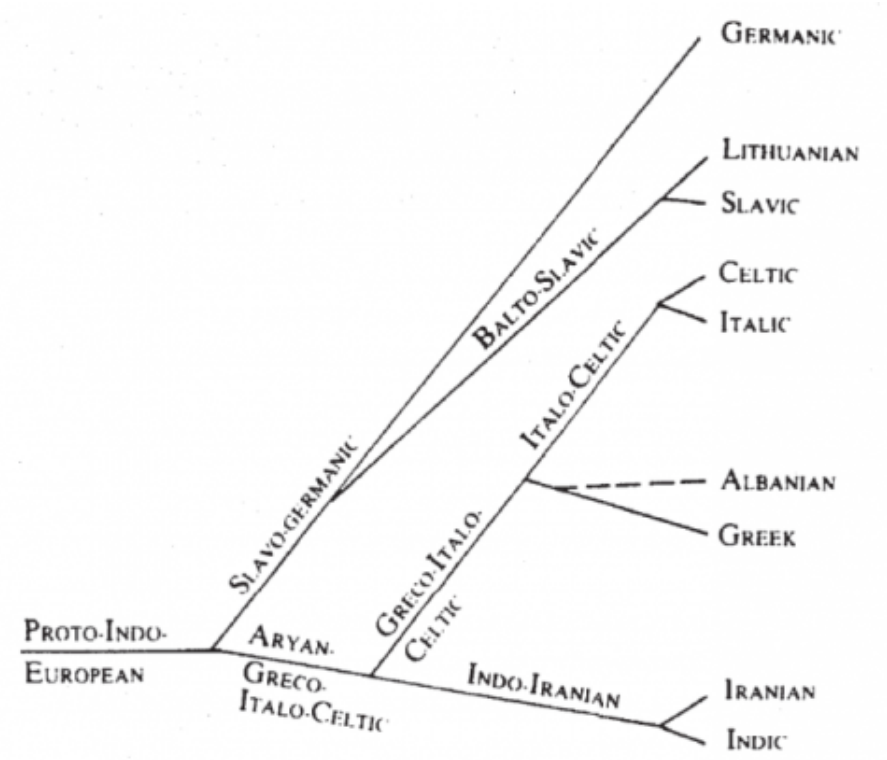

Figure 2.3.: Schleicher's Indo-European tree diagram (Schleicher, 1861, p. 9). English translation of the original German version, sourced from Wikimedia Commons.

same phonetic environment across a speech community, rather than occurring sporadically or irregularly. In this framework, sound change was treated as a regular and, in principle, exceptionless process (Baldi, 2002; Lehmann, 1993; Clackson, 2022).

The shift was foundational because it allowed linguists to reject explanations based on accidental similarity, and instead ground comparison in recurring sound correspondences. If a proposed genealogical relationship between languages could not be supported by systematic correspondences, its historical validity is questioned. The Indo-European languages provided especially strong evidence for its regularity; famous example include Grimm's Law and Verner's Law, where even apparent exceptions were eventually explained through specific conditioning factors (Baldi, 2002).

At the same time, the Neogrammarians also recognized that sound change was not the sole driver of linguistic evolution. They acknowledged that new form could arise through analogy (or form association), a process where morphological, syntactic, or semantic patterns within a language shape its structure (Baldi, 2002; Clackson, 2022; Lehmann, 1993). This shows that linguistic change is regular in some respects, but also shaped by internal restructuring within the language system.

Following the methodological refinements of the Neogrammarians, Indo-European subgrouping came to rely primarily on shared innovations rather than shared retentions. A subgroup consist of languages that descend from a common ancestor more recent than the proto-language of the entire family (Clackson, 2022). The strongest evidence for such a relationship is found in shared morphological innovations. If two languages develop a unique morphological category or a new grammatical marker not found elsewhere in the family, it provides strong evidence that they belong to the same exclusive speech community (Clackson, 2022; Weiss, 2015). By contrast, shared retentions, which are features that several languages simply preserved from the parent language, are less decisive. These features do not prove a shared history after the initial split; they only prove that the languages did not change in that specific way (Clackson, 2007).

For this reason, morphological innovations are often treated as especially important evidence for subgrouping. While phonological and lexical similarities provide a starting point, they are often compromised by external factors such as borrowing, chance resemblance, or late-stage convergence (Weiss, 2015; Clackson, 2022). Complex morphological developments, on the other hand, are generally less likely to spread through casual contact than individual lexical items, and therefore, they provide stronger evidence. Thus, scholars mostly prioritize morphological innovations to identify a shared history among subgroups.

Subgrouping nevertheless remains methodologically complex. A major difficulty lies in distinguishing genuine shared innovations from "false positives", that is, similarities that arise through later contact or independent, parallel developments (Clackson, 2022). This challenge is especially relevant for lexical and semantic evidence, where a shared trait may reflect inherited material, borrowing, or a later semantic shift. Consequently, Indo-European subgrouping is best understood through cumulative evidence rather than through any single decisive feature. The most robust classifications combine regular sound correspondences, morphological innovations, and historically meaningful lexical or semantic developments.

2.2.3. Computational Approaches to Phylogenies

In recent decades, advances in statistical modeling and computing have made it possible to process much larger linguistic datasets than were previously feasible. As a result, the reconstruction of language relationships has increasingly been supplemented by computational approaches that model Indo-European subgrouping and chronology using explicit algorithms. In this context, linguists refer not only to family trees, but also to phylogenies and cladistics. These terms borrowed from evolutionary biology because some of the underlying methods are formally related to those studied in the genetic descent (Clackson, 2007; Olander, 2022). These approaches do not replace the comparative method but rather seek to formalize certain aspects of it through more systematic, large-scale analysis.

Within Indo-European studies, computational work has developed mainly along two lines [2]. One line, associated with a group of researchers from Pennsylvania, applied weighted maximum compatibility algorithms to a data set consisting of phonological, morphological, and lexical characters (Ringe et al., 2002; Nakhleh et al., 2005). The other approach, pioneered by a group from New Zealand, relies primarily on basic vocabulary lists and utilizes Bayesian phylogenetic methods adapted from evolutionary biology (Gray and Atkinson, 2003; Bouckaert et al., 2012). The contrast between these two lines of research is significant, as they differ not only in their computational techniques but also in the kinds of linguistic evidence they prioritize.

The Pennsylvania tree, associated with Ringe et al. (2002), is based on a dataset that combines 22 phonological, 13 morphological, and 333 lexical characters for 24 ancient and medieval Indo-European languages. This method is partly rooted in traditional Indo-European comparative method, since it seeks to identify shared innovations and gives greater computational weight to morphological and phonological developments than lexical ones (Olander, 2022; Clackson, 2022; Piwowarczyk, 2022).

For this reason, this study is often seen as more closely aligned with classical linguistics rather than vocabulary-based phylogenies. However, it revealed some important difficulties. Most notably, the position of Germanic could not be fitted neatly into the "perfect phylogeny" (Ringe et al., 2002). The authors treated Germanic as problematic in their main Maximum Compatibility analysis, and later discussion has interpreted this as evidence of a complex history involving contact and convergence with several other branches (Ringe et al., 2002; Nakhleh et al., 2005; Piwowarczyk, 2022). This highlighted the limitation of a strictly tree-based model when significant language contact is involved.

A different approach was proposed by Gray and Atkinson (2003), whose model relied exclusively on basic vocabulary data derived from the word lists and cognancy judgments of Dyen et al. (1992) for 87 Indo-European languages. Using Bayesian inference, Gray and Atkinson (2003) estimated both subgroupings and divergence dates. One advantage

[2]The discussion here primarily focuses on (Ringe et al., 2002; Nakhleh et al., 2005) and (Gray and Atkinson, 2003). For broader methodological discussion on computational tools, see also Gerhard Jäger & Johann-Mattis List, 2019, *Statistical and computational elaborations of the classical comparative method* (unpublished manuscript), https://lingulist.de/documents/papers/jaeger-list-2016-computational-elaborations-comparative-method.pdf, (accessed 06 February 2026).

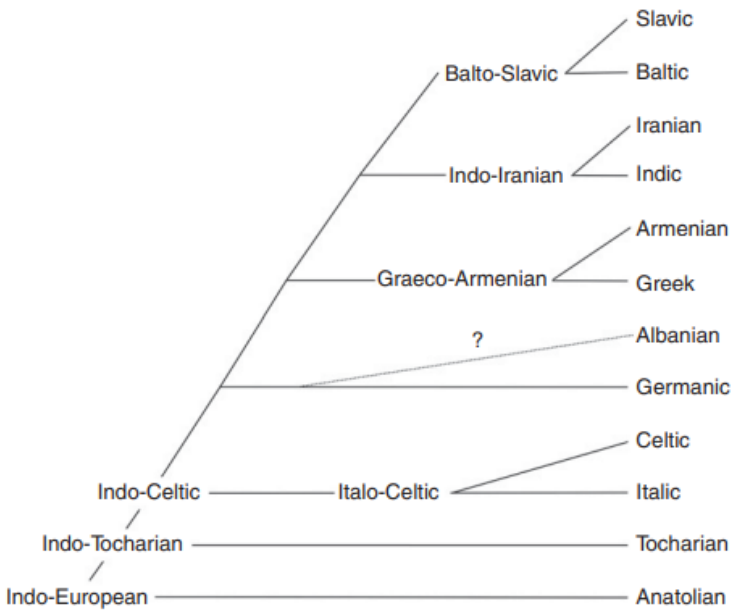

Figure 2.4.: The Pennsylvania model tree, showing the uncertain placement of Germanic and Albanian in the perfect pyhlogeny analysis. Reproduced from Olander (2022, p. 6), based on (Ringe et al., 2002) and (Nakhleh et al., 2005). Licensed under CC BY-NC-ND 4.0.

of this dataset is that lexical datasets are comparatively easy to process and allow for the inclusion of a much larger number of modern languages.

However, vocabulary-based phylogenies have often been treated with caution in Indo-European studies, as lexical evidence alone is typically considered insufficient to establish subgroups in the strict comparative sense. Furthermore, the chronological implications of the Gray and Atkinson (2003)'s model remain controversial[3]. Their results were especially influential because their dating was broadly compatible with the Anatolian hypothesis (Renfrew, 1987). Later refinements, such as the study by Chang et al. (2015), introduced ancestry constraints (e.g. the known descent of Romance from Latin), which produced a shorter chronology more compatible with Steppe hypothesis (Chang et al., 2015; Piwowarczyk, 2022). This shows that computational results are highly sensitive to the assumptions and constraints built into the model. Even with similar methods, different constraints and data treatments may produce significantly different trees and chronologies.

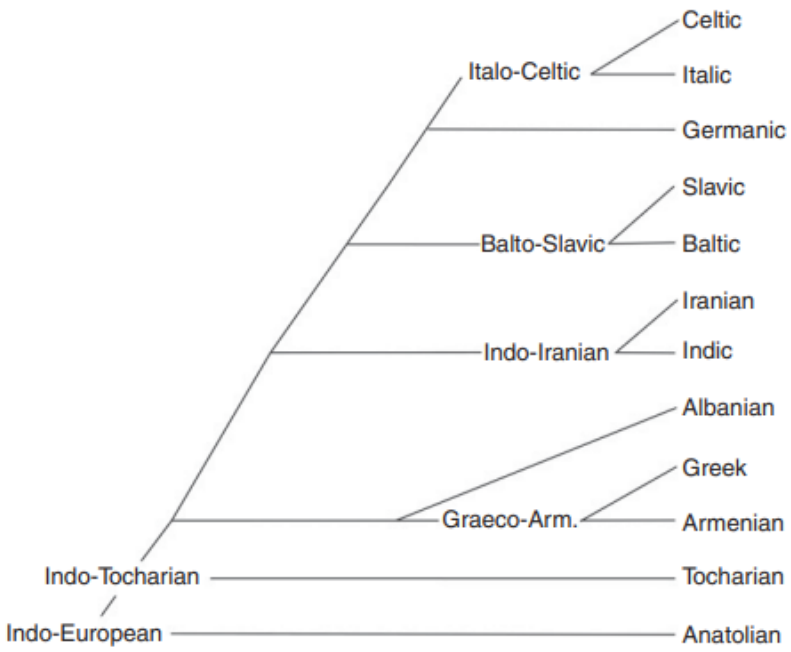

Figure 2.5.: The New Zealand tree based on the Bayesian method, utilizing lexical cognacy to estimate both subgrouping and divergence dates. Reproduced from Olander (2022, p. 6), based on (Gray and Atkinson, 2003) and (Chang et al., 2015). Licensed under CC BY-NC-ND 4.0.

[3]While Bayesian vocabulary based methods have been influential, their use for dating Indo-European divergence is still debated, especially with regard to model choice, lexical data selection, and the extent to which semantic or lexical evidence alone can support chronological inference. See Piwowarczyk (2022), and Ringe (2022); for a more favorable discussion of Bayesian method, see Greenhill and Gray (2012).

A comparison of Figures 2.4 and 2.5 shows that while these computational traditions agree on major points, they diverge in others. In both models, the first splits seperate Anatolian and Tocharian from the rest of the family. Beyond these early stages, however, the structures differ significantly. In the New Zealand model shown in Figure 2.5, which is based on lexical cognacy data, Albanian is placed close to Indo-Iranian, and Germanic appears relatively close to Italo-Celtic. By contrast, in the Pennsylvanian model shown in Figure 2.4, is not closely connected to Indo-Iranian, and Germanic is represented in a polytomy with Albanian, indicating that its phonological and morphological character do not allow for a clear binary placement.

These differences are not merely technical disagreements; they reflect the fact that the two approaches model different types of evidence. The New Zealand model captures patterns of lexical cognacy and overall vocabulary similarity, whereas the Pennsylvania model is shaped by phonological and morphological innovations. This shows that computational models may capture different dimensions of linguistic history rather than simply producing different versions of the same family tree. For the present thesis, this distinction is important because semantic network properties derived from distributional data may likewise reflect one particular dimension of linguistic similarity, rather than a complete model of genealogical descent.

2.3. Distributional Semantics

Distributional semantics shifts the focus from explicitly described lexical relations to patterns of language use. The basic idea is that words acquire meaning through the contexts in which they occur, and that these contexts can be used as evidence for semantic similarity. Rather than starting from predefined relations such as synonymy, antonymy, or membership in semantic fields, distributional approaches infer relations between words from their distribution across a corpus. In this way, distributional semantics provides a corpus-based and computational extension of the relational view of meaning discussed in the previous sections (Turney and Pantel, 2010; Lenci and Sahlgren, 2023; Boleda, 2020).

The core assumption underlying this framework is the distributional hypothesis, commonly associated with Harris (1954) and Firth (1957). According to this hypothesis, words that occur in similar linguistic contexts tend to have similar meanings. Harris (1954) argues that linguistic elements can be characterized by their distribution across contexts, while Firth (1957, p. 11) expresses this idea in his well-known formulation that "one should know a word by the company it keeps". Within this framework, the linguistic environments in which a word appears are treated as evidence of its meaning. Words with similar contextual distributions are therefore interpreted as semantically similar (Lenci and Sahlgren, 2023; Boleda, 2020; Turney and Pantel, 2010).

A key feature of distributional semantics is that semantic representations are learned directly from corpus data rather than being manually specified. Instead of defining meaning through a limited set of interpretable features, distributional models represent words through patterns distributed across many contextual dimensions. These patterns form a semantic space in which relationships between words can be measured geometrically.

Words that occur in similar contexts are positioned closer together, while words with less similar distributions are placed farther apart Baroni et al. (2014a); Turney and Pantel (2010); Boleda (2020).

This spatial representation makes it possible to model semantic similarity as a graded phenomenon. Rather than treating words as either related or unrelated, distributional semantics represents similarity as a matter of degree: words may be more or less similar depending on the extent to which their contextual distributions overlap (Boleda and Herbelot, 2016). Such relationships are commonly measured geometrically within the semantic space, for example, through cosine similarity. This makes distributional semantics especially useful for computational approaches to lexical meaning, because it allows large vocabularies to be represented, compared, and analyzed systematically.

2.3.1. Vector Space Models and Semantic Similarity

The distributional hypothesis explains meaning as a function of linguistic context, but in order to be used computationally, it requires a formal representation. ector space models provide such a representation by encoding words as numerical vectors derived from their patterns of occurrence in a corpus. In this framework, lexical meaning is modeled geometrically in a high-dimensional space, so that words with similar contextual distributions are located in similar regions of that space (Turney and Pantel, 2010; Baroni et al., 2014a; Lenci and Sahlgren, 2023).

A standard way to construct such representations is via a word-context matrix, also known as a distributional matrix. In this matrix, rows correspond to target words, and columns correspond to contextual features. These features may be words that occur near the target word, grammatical contexts, or other corpus-derived properties. Each cell records how strongly a target word is associated with a given context. In the simplest case, this value is a raw co-occurrence count, although distributional models often apply weighting schemes to make these associations more informative(Boleda, 2020; Lenci and Sahlgren, 2023; Turney and Pantel, 2010).

Formally, let $W = \{w_1, w_2, \ldots, w_n\}$ be the set of target words and $C = \{c_1, c_2, \ldots, c_m\}$ the set of contextual features. Here, w_i refers to the i-th word in the vocabulary, while c_j refers to the j-th context used to describe that word. From these target words and contexts, a word-context matrix can be defined as:

$$\mathbf{M} \in R^{n \times m}, \tag{2.1}$$

where each row represents a word and each column represents a context. The value in cell M_{ij} shows the association between word w_i and c_j. In a simple count-based model, this value corresponds to the number of times w_i occurs with c_j in the corpus(Turney and Pantel, 2010; Lenci and Sahlgren, 2023). The row corresponding to a word w_i is then is its distributional vector of that word:

$$\mathbf{v}_i = (M_{i1}, M_{i2}, \ldots, M_{im}) \tag{2.2}$$

This means that each word is represented by the full row of values associated with it. For example, the vector for w_i contains it association values with all contexts c_1 to c_m. Two words can then be compared by comparing their vectors. If their rows contain similar patterns of values, the words are considered distributionally similar. In this way, semantic similarity is modeled not through predefined categories, but through similarity between numerical profiles in the word-context matrix (Turney and Pantel, 2010; Lenci and Sahlgren, 2023).

	runs	**barks**
dog	1	5
hyena	1	2
cat	4	0

Table 2.1.: Co-occurrence count example for *dog*, *hyena*, and *cat*. Adapted from (Baroni et al., 2014a, p. 248). Licensed under CC BY 4.0.

This idea can be illustrated from a small example from Baroni et al. (2014a). Suppose the target words are *dog*, *hyena* and *cat*, and the contexts are **runs** and **barks**. As shown in Table 2.1, the word *dog* occurs once with **runs** and five times with **barks**. Even from this small example, it is possible to see that *dog* is distributionally closer to *hyena* than to *cat* because they both share similar profile with both contexts. By contrast, *cat* occurs more with **runs** and not with **barks.**

In Figure 2.6, these contextual distributions are visualized as vectors in a two-dimensional space, where the x and y coordinates correspond to the first and second vector components(Baroni et al., 2014a). This illustrates the central idea of vector semantics: words with similar contextual distributions point in similar directions in semantic space. Although real distributional models operate in much higher dimensional spaces, often with hundreds or thousands of dimensions, the same geometric intuition still applies.

To measure such similarity more precisely, vector space models typically use cosine similarity. For two distributional vectors $\mathbf{u}$ and $\mathbf{v}$, cosine similarity is defined as

$$\cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \tag{2.3}$$

Rather than calculating the absolute distance between two vectors, cosine similarity measures the angle between them. Its value is highest when two vectors point in the same direction. In count-based distributional models, where vector values are non-negative, cosine values usually range between 0 and 1. The more similar the direction of two vectors, the higher the cosine value, and the more similar their contextual distribution is assumed to be (Jurafsky and Martin, 2026; Lenci and Sahlgren, 2023).

Cosine similarity is generally preferred over raw distance measures because it is less dependent on vector length. Frequent words tend to have longer vectors simply because they co-occur more often, but greater length does not necessarily imply greater semantic similarity. Cosine normalization focuses on direction rather than magnitude, so that

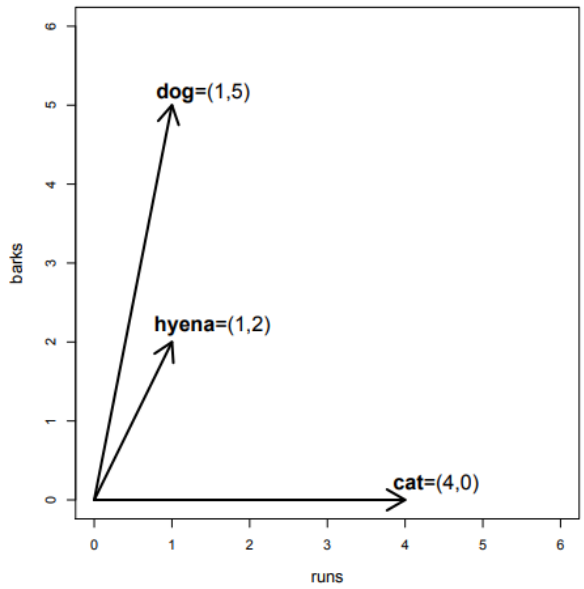

Figure 2.6.: Geometric representation of *dog*, *hyena*, and *cat* as distributional vectors in a two-dimensional semantic space. Reproduced from (Baroni et al., 2014a, p. 249). Licensed under CC BY 4.0.

similarity is captured through the relative orientation of vectors in semantic space (Baroni et al., 2014b; Turney and Pantel, 2010).

In the Figure 2.6, the cosine between *dog* and *hyena* is high (≈ 0.96) because their vectors are closely aligned, while the cosine between *dog* and *cat* is much lower (≈ 0.20) because their directions differ more strongly (Baroni et al., 2014a). The smaller the angle between two vectors, the greater the semantic similarity; the wider the angle, the lower the similarity. In this way, vector space models make semantic similarity computable as a geometric property.

However, these co-occurrence matrices are often high-dimensional, since languages contain a large number of possible contexts, and sparse, meaning most cells are equal to zero (Lenci and Sahlgren, 2023). As Lenci and Sahlgren (2023) notes, the high dimensionality and sparseness can introduce noise and make it difficult to distinguish between similar from dissimilar words in semantic space. This combination also creates computational difficulties, since larger vectors require more storage and make similarity calculations more costly. It can also affect the quality of representation itself, because similar contexts are treated as separate dimensions, and many plausible co-occurrences remain unobservable (Turney and Pantel, 2010; Lenci and Sahlgren, 2023). For this reason, later developments in distributional semantics shifted toward word embeddings, which aim to represent words in denser, lower-dimensional vector spaces.

2.3.2. Word Embeddings

Word embeddings emerged partly in response to the limitations of count-based vector models. In count-based models, words are represented through explicit co-occurrence information, with each dimension corresponding directly to a particular context. While this makes the representation interpretable, it also produces vectors that are typically high-dimensional and sparse (Lenci and Sahlgren, 2023). By contrast, word embeddings learn more compact vector representations from distributional patterns in corpora. Lenci and Sahlgren (2023) describe embeddings as implicit distributional vectors whose components no longer correspond to individual contexts but to latent features extracted from co-occurrences. In simpler terms, word embeddings are relatively short numerical vectors whose positions in semantic space reflect patterns of linguistic usage and preserve semantic relations between words that occur in similar contexts (Schnabel et al., 2015; Turian et al., 2010; Torregrossa et al., 2021; Ghannay et al., 2016).

A central feature of word embeddings is that they operate in a lower-dimensional latent semantic space. Rather than preserving each observed context as a separate dimension, the model compresses contextual evidence into a smaller number of abstract dimensions. This process is often described as feature extraction or dimensionality reduction (Turney and Pantel, 2010; Lenci and Sahlgren, 2023). Its purpose is not only to reduce the size of the representation, but also to reveal broader semantic structure by reducing noise, smoothing the effects of data sparsity, and capturing higher order co-occurrence patterns (Turney and Pantel, 2010). Word embeddings therefore do not simply shorten vectors; they aim to represent semantic structure more effectively.

Another important difference concerns how these representations are learned. Many embedding models are predictive rather than count-based. Count-based approaches begin with a co-occurrence matrix and derive vectors from global corpus statistics such as word counts and frequencies. Predictive models learn vectors by solving a language-related prediction task over running text Baroni et al. (2014b); Torregrossa et al. (2021); Lenci and Sahlgren (2023). Instead of directly storing how often a context word appears near a target word, the model learns by asking whether that context word is likely to occur near the target at all (Jurafsky and Martin, 2026). The goal is not the prediction in itself for its sake, but the fact that learned weights become the word vectors (Jurafsky and Martin, 2026; Baroni et al., 2014b). In this way, semantic similarity emerges through the training process: words that occur in similar contexts, or help predict similar contexts, tend to end up close to one another in the latent semantic space. This is one reason why word embeddings became especially useful in NLP and were quickly adopted for large-scale unsupervised learning from unlabelled corpora (Ghannay et al., 2016; Lenci and Sahlgren, 2023; Baroni et al., 2014b; Torregrossa et al., 2021).

The best known model in this line is word2vec, introduced by Mikolov et al. (2013a,b). The basic idea is to learn a dense vector for each word by training the model to make that vector useful for predicting neighboring words in a corpus. Mikolov et al. (2013a) describe the skip-gram model as an efficient model for learning high quality vector representations from large amounts of unstructured data and emphasize its training objective as learning representations that are good at predicting nearby words. One of

the reasons word2vec became so influential was its efficiency. Because the architecture avoided the heavier computations required by earlier models, it made training on very large corpora much more practical. At the same time, the resulting vectors were shown to capture many syntactic and semantic regularities. A well known example is the analogy relation vec(King) − vec(Man) + vec(Woman) yields a vector that is closest to vec(Queen), suggesting that the embedding space can encode certain relations in a geometrically meaningful way (Mikolov et al., 2013a).

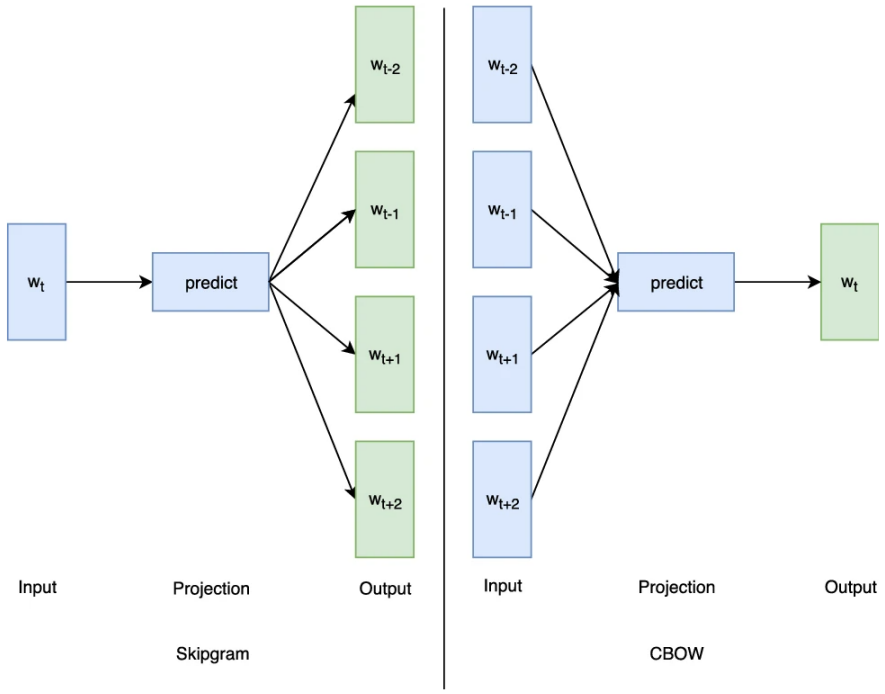

Figure 2.7.: Architecture of skip-gram and continuous bag of words (CBOW). In skip-gram, the target word is used to predict surrounding context words; in CBOW, the surrounding context words are used to predict the target word. Reproduced from (Torregrossa et al., 2021), based on (Mikolov et al., 2013a), with permission from Springer Nature.

Word2vec includes two main architectures, continuous bag of words (CBOW) and skip-gram, as shown in Figure 2.7. In CBOW, the model predicts a target word from its surrounding context words. In skip-gram, the direction is reversed, and the model uses the target word to predict surrounding context words (Mikolov et al., 2013a,b; Torregrossa et al., 2021; Ghannay et al., 2016). During training, word2vec keeps separate representations for target words and context words, since appearing as the center of a context window and appearing within that window are treated as distinct roles (Torregrossa et al., 2021). If the target word is written as w_t and a context word as w_c, then skip-gram learns by maximizing the probability $p(w_c \mid w_t)$ across the corpus (Torregrossa et al., 2021;

Ghannay et al., 2016). CBOW reverses this by maximizing the probability of the target word given the surrounding context. In practical terms, CBOW combines contextual evidence in order to recover the center word, whereas skip-gram uses the center word to recover its neighbors (Baroni et al., 2014b; Ghannay et al., 2016; Torregrossa et al., 2021). Because the CBOW task is somewhat simpler, it is generally faster and often performs well with frequent words, whereas skip-gram produces stronger representations from rarer words and is often better at capturing semantic relations (Mikolov et al., 2013a,b; Almeida and Xexéo, 2019).

Later improvements to word2vec introduced two important techniques: negative sampling and subsampling of frequent words (Mikolov et al., 2013b). Negative sampling replaces the costly calculation of a full probability distribution over the entire vocabulary with a simpler task in which the model learns to distinguish genuine target context pairs from randomly generated noise pairs (Mikolov et al., 2013b; Baroni et al., 2014b). Subsampling of frequent words reduces the influence of common words such as function words, which often contribute relatively little semantic information (Mikolov et al., 2013b). These techniques made training faster and often improved the quality of the learned vectors.

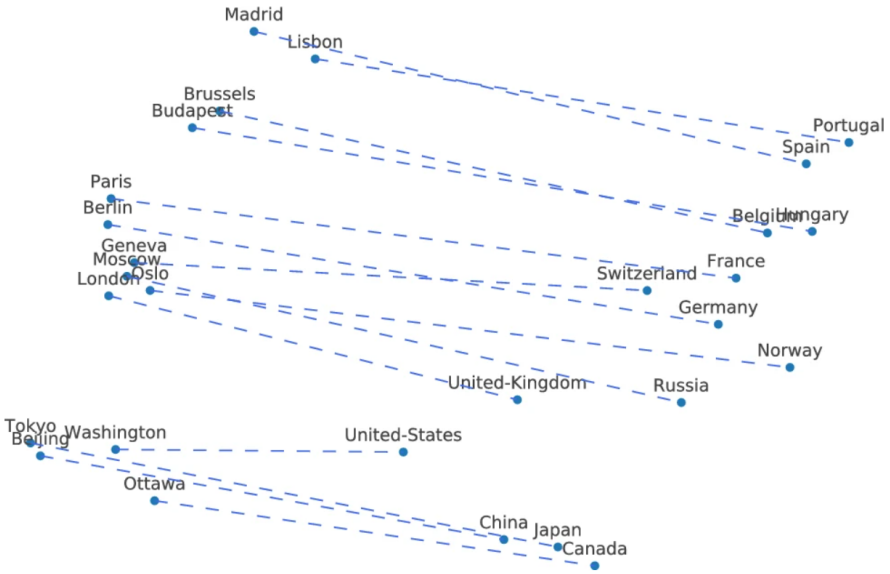

Figure 2.8.: Countries and capital cities in a 300-dimensional FastText embedding space, projected onto a two-dimensional plane using PCA. Reproduced from Torregrossa et al. (2021), based on Bojanowski et al. (2017), with permission from Springer Nature.

A useful illustration of the geometric structure of embeddings is provided in Figure 2.8. The figure shows countries and their capital cities in a 300-dimensional FastText embedding space using principal component analysis. Even in this reduced space, countries and their capital cities form roughly parallel relations, with dashed lines connecting each

country to its capital. The model was not provided with explicit information about what a capital city is; rather, this relation emerges from distributional learning. This shows that embedding models can organize lexical concepts in a geometrically meaningful way and capture certain semantic relations implicitly through the structure of the embedding space.

An important extension of word2vec is fastText, proposed by Bojanowski et al. (2017). The starting point of fastText is a limitation shared by word2vec and similar models like GloVe (Pennington et al., 2014). These models assign one vector to each whole word form and potentially ignore the internal structure of words (Bojanowski et al., 2017). This becomes especially problematic in morphologically rich languages, where many word forms are rare or absent from the training corpus. As Torregrossa et al. (2021) notes, subword units can carry meaningful information. For example, the suffix <*ed*> signals past, <*er*> can mark comparison and <*ly*> is associated with adverbs (Torregrossa et al., 2021). Bojanowski et al. (2017) identify this as a major limitation for languages such as Turkish or Finnish who are morphologically rich, and also note that in languages such as French or Spanish, many inflected forms occur too rarely to obtain reliable whole word vectors.

FastText addresses that issue by representing each word as a bag of character n-grams. Instead of learning only a vector for the whole word, it also learns vectors for character sequences and represents the word as the sum of these subword vectors. Bojanowski et al. (2017) describe this as an extension of the continuous skip-gram model that incorporates subword information. Let's take the example of the word *talked*. In fastText, the whole word form is taken into account, but so are its internal n-grams, such as <*ta*>, <*tal*>, <*alk*>, <*lke*>, <*ked*>, and <*ed*>, depending on the chosen n-gram size. The final vector for *talked* is computed as the sum of these subword representations together with the whole word vector. This allows the model to capture not only the lexical base *talk*, but also the morphological ending <*ed*>, which may signal past tense.

This subword design is one of the reasons fastText is especially relevant for cross-linguistic work. Bojanowski et al. (2017) report across nine languages and show that using character n-grams improves performance on similarity and analogy tasks, especially in morphologically richer languages. Their experiments also show stronger gains over skip-gram in languages such as Czech and Russian than in Spanish or French, which underlines the particular importance of subword information when morphology is more complex (Bojanowski et al., 2017). FastText is therefore not only an extension of word2vec, but also a model that is better suited to handling rare forms, inflectional variation, and out-of-vocabulary items across languages.

For the purposes of this thesis, word embeddings are relevant because they provide compact and comparable semantic representations across languages. FastText is especially suitable because it also handles morphological variation and data sparsity more robustly than whole-word models. Since the later analysis relies on comparing semantic structures across languages and identifying structural patterns, fastText-based embeddings offer a practical and theoretically appropriate basis for such comparison.

2.4. Semantic Networks

The previous section illustrated how semantic similarity can be modeled in a vector space using embeddings. A complementary way of representing semantic structure is through networks. From this perspective, the analytical focus shifts from pairwise similarities to the broader structural patterns of relations among lexical items. This shift is important for the present thesis because embeddings provide the semantic information from which networks can be constructed, while network analysis makes it possible to examine how this information is organized across the lexicon as a whole.

In its simplest form, a network is a graph composed of nodes or vertices, connected by links, also called edges or arcs (Newman, 2018; Steyvers and Tenenbaum, 2005). In linguistic and semantic applications, nodes usually represent words, concepts or lexical units, while links represent semantic relations between them (Steyvers and Tenenbaum, 2005). WordNet serves as an intuitive example, organizing meaning through synsets connected by relations such as synonymy, hypernymy , and meronymy (Fellbaum, 1998).

However, semantic networks are not limited to manually created resources such as WordNet. More generally, they can be understood as graph- based representations of semantic knowledge, where the meaning of a node is partly determined by its connections to other nodes. In this sense, semantic networks are closely related to the broader view of lexical meaning developed in the previous sections: words and concepts are not treated as isolated units, but as elements embedded in a larger relational structure. The exact form of this structure depends on how nodes and links are defined (Steyvers and Tenenbaum, 2005).

This flexibility is important because the term "semantic network" is used in different ways across linguistics, cognitive science, psychology, and computational modeling [4] . In some studies, links are based on co-occurrence, where words are connected because they appear near each other in a sentence, document, or corpus. In other cases, links are based on word association data. A third type is based on lexical-semantic relations, such as in the case of WordNet, which may be defined in lexical databases or extracted computationally (Budel et al., 2023). As Borge-Holthoefer and Arenas (2010) note, researchers must therefore decide both what counts as a node and what counts as a link, since these decisions determine the kind of semantic structure that the network represents.

A more recent development that is especially relevant for the present thesis is the connection between distributional semantic models and network-based representations. Network-based approaches and distributional models differ in their data sources and representational structure, but they can also be understood as complementary ways of modeling semantic knowledge (Kumar et al., 2022). Utsumi (2015), for example, applies network analysis to semantic networks created from distributional semantic models and shows that such networks exhibit properties similar to word-association networks, including small-world and scale-free tendencies. In this context, embedding-based networks can be understood as a way of transforming distributional similarity into graph structure.

[4] For a systematic review of how the term semantic network is defined across different fields, see (Pereira et al., 2022).

Such networks may not directly encode manually defined semantic relations such as synonymy or hyponymy; rather, they represent broader patterns of semantic proximity in distributional space.

For this reason, embeddings and semantic networks should not be understood as competing models of meaning. Embeddings provide continuous vector representations of lexical meaning, while networks transform these similarities into a graph structure that can be analyzed using network measures. This allows researchers to move from the level of individual word similarities to the level of larger structural patterns, such as connectivity, clustering, centrality, and community structure.

In the present thesis, the term "semantic network" is used specifically in this distributional sense. Nodes represent words, and links represent semantic similarity derived from fastText embeddings. The aim is therefore not to reconstruct manually defined lexical-semantic relations, but to examine whether the structure of embedding-based semantic similarity networks differs systematically across Indo-European languages. This definition provides the basis for the following section, which introduces the semantic network properties used to describe and compare these networks.

2.4.1. Semantic Network Properties

Once semantic relations are represented as a network, the methods applied to them can be classified at different level of granularity. Following Borge-Holthoefer and Arenas (2010), it is useful to distinguish between three primary scales of analysis:

- Micro level (Local): This level concerns the properties of individual nodes. Measures such as degree and local clustering coefficient describe the specific position and neighborhood of individual words within the network.
- Macro level (Global): This level concerns the structure of the network as a whole. Measures such as average degre, density, average path length, and diameter are used to describe the broader, system-wide structural properties of the lexical graph.
- Meso level (Community): This level concerns intermediate group structures. Measures such as modularity capture how the network is partitioned into clusters or subgroups that are more densely connected internally than to the rest of the network.

The formal behavior of these measures depends on the underlying graph topology.. In an undirected network, a link has no direction and simply connects two nodes, whereas in a directed network, the link has an orientation, so that one node points to another (Borge-Holthoefer and Arenas, 2010; Steyvers and Tenenbaum, 2005). Within this structure, two nodes connected by a link are referred to as neighbors. A path is a sequence of connected nodes, and the distance between two nodes is the length of the shortest path between them. A network is considered connected if there is a path between all pairs of nodes. In directed network, a graph is strongly connected if every node can be reached from every other node by following the direction of the links (Borge-Holthoefer and Arenas, 2010).

Furthermore, networks can be weighted or unweighted. In an unweighted network, all links are treated equally, whereas in a weighted network, each link carries a value that

expresses the strength of the relation (Borge-Holthoefer and Arenas, 2010; Newman, 2018). This distinction is especially relevant for semantic data, as similarity-based networks may, in principle, preserve graded information through edge weights. To clarify the mathematical foundations of these properties, the following table of notation defines the symbols used in the subsequent formulas.

Level	Property	Symbol	Meaning
Basic	Number of nodes	N	Number of words in the network.
	Number of edges	E	Number of semantic similarity links.
Micro	Node degree	k_i	Number of direct neighbors of a node.
	Local clustering coefficient	C_i	Connectivity among a node's neighbors.
	Shortest path distance	d_{ij}	Shortest path between two nodes.
Macro	Average degree	$\langle k \rangle$	Mean number of edges per node.
	Density	ρ	Share of possible edges that exist.
	Average path length	L	Mean shortest-path distance.
	Diameter	D	Longest short-path distance.
	Average clustering coefficient	$\bar{C}$	Mean local clustering across nodes.
	Transitivity	T	Overall tendency to form triangles.
	Degree Assortativity	r	Tendency to connect similar degree nodes.
Meso	Modularity	Q	Strength of community structure.

Table 2.2.: Summary of semantic network properties and their descriptions.

Average Degree and Density

One of the most basic local properties of a network is node degree. In an undirected network, the degree of a node is the number of edges attached to it. Newman (2018) defines the degree of node i as k_i, representing the total number of edges connected to that node. In directed networks, the degrees can be further distinguished as in-degree, referring to incoming links, and out-degree, referring to outgoing links (Steyvers and Tenenbaum, 2005; Borge-Holthoefer and Arenas, 2010). The degree of a word indicates how many direct semantic neighbors it has in the network. A high-degree node is well-connected, while a low-degree node occupies a more peripheral position in the lexicon (Newman, 2018).

At the network level, this property can be summarized by the average degree ($\langle k \rangle$), defined as:

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i \tag{2.4}$$

Here, the degree of every node in the network is summed and divided by the total number of nodes N. The average degree indicates the typical number of connections per node and therefore provides a general measure of network connectivity (Steyvers and Tenenbaum, 2005; Borge-Holthoefer and Arenas, 2010; Newman, 2018).

A related global property is density (ρ), which measures the proportion of potential links that are actually present in the network. For an undirected network with N nodes and E edges, density is defined as:

$$\rho = \frac{2E}{N(N-1)} \tag{2.5}$$

Density compares the observed number of edges to the maximum possible number of edges in a network of the same size. As (Newman, 2018) notes, this can be interpreted as the probability that any two randomly chosen nodes are connected. Density indicates how compact or sparse the lexical structure is overall. A high-density network contains many semantic links relative to its size, whereas a low-density network contains relatively few. Most real-world networks are sparse, meaning that each node is connected to only a small fraction of the total possible neighbors (Steyvers and Tenenbaum, 2005; Newman, 2018; Borge-Holthoefer and Arenas, 2010).

Together, degree, average degree, and density provide a basic description of semantic network connectivity. Degree captures the connectivity of individual words, average degree summarizes the typical number of semantic neighbors per word, and density describes the overall compactness of the lexical network.

Shortest Path and Diameter

Shortest path and diameter of a semantic network concern the distances between nodes. The shortest path length (d_{ij}) between two nodes i and j, is the minimum number of edges required to connect them (Steyvers and Tenenbaum, 2005; Borge-Holthoefer and Arenas, 2010). It expresses the relational proximity between words: a short distance indicates that two words are closely related within the network, whereas a long distance suggests they belong to more distant regions of the lexical structure (Steyvers and Tenenbaum, 2005; Newman, 2018; Borge-Holthoefer and Arenas, 2010).

To summarize these distances across the network as a whole, one can use the average shortest path length (L), which represents the mean shortest-path distance between all pairs of nodes for which a path exists:

$$L = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \tag{2.6}$$

This measure is calculated by summing the shortest-path distances between every pair of nodes and dividing by the total number of ordered pairs. Borge-Holthoefer and Arenas (2010) describe L as the average separation of node pairs, interpreting it as the effective "linear size" of the network. Alongside the previously discussed connectivity measures, L indicates how compactly the network is organized by showing how many steps are typically needed to connect one word to another.

A related macro-level property is the diameter (D) of the network, defined as:

$$D = \max_{i,j} d_{ij} \tag{2.7}$$

The diameter represents the longest shortest path between any two connected nodes in the graph (Newman, 2018; Borge-Holthoefer and Arenas, 2010). While the average shortest path length describes the typical separation between words, the diameter captures the maximum extent of the network. A small diameter indicates that even the most distant concepts in the lexicon remain relatively close to one another, while a large diameter suggests a more expansive or weakly connected structure (Steyvers and Tenenbaum, 2005).

Together, average shortest path length (L) and diameter (D) describe how compact or expansive a semantic network is in terms of connectivity and path distances.

Clustering Coefficient and Transitivity

Clustering coefficient (C_i) describes the extent to which the neighbors of a node are also connected to one another. It captures whether local semantic neighborhoods form tightly connected groups rather than loose sets of unrelated neighbors. A high clustering coefficient suggests that words connected to the same word are also likely to be semantically connected to each other (Steyvers and Tenenbaum, 2005; Borge-Holthoefer and Arenas, 2010).

Suppose a node i has a degree (k_i), meaning it is connected to k_i neighbors. Among these neighbors, a maximum of $\frac{k_i(k_i-1)}{2}$ edges can exist, which occurs if every neighbor is connected to every other neighbor. If the actual number of links among those neighbors is E_i, then the local clustering coefficient (C_i) of node i is defined as:

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \tag{2.8}$$

This formula compares the observed edges within a node's neighborhood to the theoretical maximum. The value is normalized between 0 and 1, where $C_i = 1$ indicates a fully connected neighborhood and $C_i = 0$ indicates that none of the neighbors are connected to one another (Steyvers and Tenenbaum, 2005; Borge-Holthoefer and Arenas, 2010).

At the macro-level, local clustering can be summarized using the average clustering coefficient ($\bar{C}$):

$$\bar{C} = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{2.9}$$

This measure described the overall tendency of nodes in the network to form locally cohesive neighborhoods. In semantic analysis, a high average clustering coefficient suggests that semantically related words tend to form dense local clusters (Steyvers and Tenenbaum, 2005).

A closely related concept is transitivity (T). Newman (2018, p. 183) explains this through the intuition that "the friend of my friend is also my friend." Transitivity measures the proportion of connected triples that are closed into triangles:

$$T = \frac{3 \times \text{number of triangles}}{\text{number of connected triples}} \tag{2.10}$$

Here, a connected triple consists of three nodes connected by at least two edges, while a triangle is a closed triple in which all the nodes are connected to one another. A higher value of T indicates a stronger tendency for semantic neighborhoods to form closed, mutualy connected groups.

Although average clustering coefficient and transitivity are closely related, they capture clustering from different perspectives. The average clustering coefficient ($\bar{C}$) summarized how strongly individual nodes tend to form cohesive neighborhoods, while transitivity (T) measures how often connected triples are closed into triangles across the network as a whole. For this thesis, the two measures are treated as complementary indicators of how strongly semantic neighborhoods are clustered.

Degree Assortativity

Degree assortativity (r) measures the tendency of nodes to connect to other nodes with a similar degree. Following Newman's formulation of assortative mixing, degree assortativity can be understood as a correlation between the degrees of nodes connected by an edge (Newman, 2002, 2003). Borge-Holthoefer and Arenas (2010) describes this in semantic network terms as a degree-degree correlation, commonly measures by the Pearson correlation coefficient between the degrees of nodes at both ends of an edge. If the coefficient is positive ($r > 0$), the network is assortative, meaning that highly connected nodes tend to link to other highly connected nodes. If the coefficient is negative ($r < 0$), the network is disassortative, meaning that highly connected hubs tend to connect to weakly connected, peripheral nodes (Borge-Holthoefer and Arenas, 2010).

This distinction is particularly meaningful for semantic networks, since different types of semantic relations can exhibit different assortativity patterns. Budel et al. (2023) note, for example, that many association-based networks are disassortative, while synonymy and antonymy networks often show a more assortative tendency. Degree assortativity (r) may indicate whether the lexical structure is organized through hub-like nodes connected to many low-degree nodes, or through more homogenous clusters in which nodes of similar connectivity tend to connect to one another.

Modularity

At the meso level, semantic networks can be analyzed in terms of community structure, that is, which refers to whether the network divides into groups of nodes that are more densely connected internally than externally. Borge-Holthoefer and Arenas (2010) describe this as an intermediate level of organization between local node properties and the network's global structure. They emphasize that communities must be identified by the graph's topology rather than by the pre-existing meaning of the nodes themselves (Borge-Holthoefer and Arenas, 2010).

A standard measure of community structure is modularity (Q). Newman (2006) defines modularity as a quantity that compares the observed number of edges within groups to the number that would be expected in an equivalent random network. In this sense, a meaningful division of a network is not simply one with few links between groups, but

one with significantly more within-group connectivity than would be expected by chance (Clauset et al., 2004). Modularity therefore evaluates how strongly a given partition separates the network into internally dense and externally sparse communities.

In its standard undirected form, modularity (Q) is defined as:

$$Q = \frac{1}{2E} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2E} \right) \delta(c_i, c_j) \tag{2.11}$$

where A_{ij} is the adjacency matrix, E is the total number of edges in the network, k_i and k_j are the degrees of nodes i and j, and $\delta(c_i, c_j)$ is the Kronecker delta, which equals 1 when the two nodes belong to the same community and 0 otherwise. This formulation explicitly calculates the difference between observed and expected within-community connectivity (Newman, 2006).

A higher value of modularity (Q) indicates that the network is divided into more clearly defined communities, whereas lower values suggest a weaker or less distinct community structure. In semantic terms, high modularity suggests that the lexical network is organized into clearly separated semantic neighborhoods or thematic domains (Newman, 2006; Borge-Holthoefer and Arenas, 2010).

2.4.2. State of the Art in Semantic Networks

Early work on semantic networks show that lexical and conceptual structure can be studied through graph-based measures, revealing non-random patterns of organization. Several of these studies interpret semantic networks in terms of broader topological patterns such as small-world and scale-free structure. In this context, short average path lengths together with high clustering were taken as evidence of small-world organization (Watts and Strogatz, 1998; Steyvers and Tenenbaum, 2005; Cong and Liu, 2014), while unequal degree distributions suggested the presence of highly connected hubs (Barabási and Albert, 1999; Steyvers and Tenenbaum, 2005). These concepts are discussed here because they shaped how early semantic network research interpreted measures.

Ferrer i Cancho and Solé (2001) constructed lexical co-ocurrence networks from the British National Corpus (BNC Consortium, 2007) and analyzed properties such as clustering coefficient, average path length, and degree distribution. Their results show that lexical co-occurrence networks combine short paths with high clustering, which they interpreted as a small-world structure. The degree distribution further suggested that lexical networks are not uniformly connected, but contain a relatively small number of highly connected words. Similarly, Motter et al. (2002) built a conceptual network from Moby English Thesaurus (Ward, 1996), defining connections between words on the basis of thesaurus entries expressing similar concepts. They found high clustering and very short average path lengths, again supporting a small-world interpretation, while the connectivity distribution suggested scale-free-like organization. In the same year, Sigman and Cecchi (2002) analyzed WordNet as a graph of word meanings, examining how different semantic relations contribute to its overall organization. Their results showed that semantic links follow scale-invariant patterns and that polysemy plays a particularly important role in

making the network more compact. Polysemous words connect multiple meanings and can therefore link otherwise distant regions of the semantic graph. This suggests that the structure of a semantic network depends not only on the number of nodes and edges, but also on the types of semantic relations through which meanings are connected.

While these early studies examined individual lexical or conceptual resources, later work began to compare different types of semantic networks within a shared methodological framework. A more systematic comparison was provided by Steyvers and Tenenbaum (2005), who analyzed three semantic networks within a single framework: free association norms (Nelson et al., 1998), WordNet(Miller, 1995), and Roget's Thesaurus (Roget, 1911). They applied a shared set of properties, including the number of nodes, average degree, average shortest path length, diameter, clustering coefficient, and degree distribution. Despite their different origins, all three networks exhibited sparse connectivity, short average path lengths, high local clustering, and highly unequal degree distributions. The authors interpreted these shared patterns as evidence that semantic networks tend to exhibit both small-world and scale-free-like organization. This comparison was important because it showed that graph-based measures can be used to compare semantic organization across resources that were constructed in very different ways.

A further step toward the approach used in this thesis is Utsumi's complex-network analysis of distributional semantic models. While earlier studies focused mainly on semantic networks derived from free association norms, WordNet, thesauri, or co-occurrence data, Utsumi (2015)constructed semantic networks directly from distributional semantic models trained on the British National Corpus (BNC Consortium, 2007). In these networks, words were represented as nodes and semantic similarity relations in distributional space were treated as edges. Utsumi then compared these networks with free-association norms (Nelson et al., 1998). The results showed that distributional semantic networks exhibit several properties also found in traditional semantic networks, including small-world structure, hierarchical organization, and truncated power-law degree distributions. More importantly for the present thesis, the study showed that different distributional model settings lead to different network structures, suggesting that graph-theoretic measures can be used to compare semantic spaces.

The use of network measures was later extended to cross-linguistic comparison through parallel, comparable, and multilingual data. Gao et al. (2014) constructed co-ocurrence networks from United Nations reports in six languages and showed that network properties can reflect interpretable linguistic contrasts. English, for example, exhibited denser connectivity, while French and Spanish showed similar values across several parameters, reflecting structural closeness. Sparser connectivity in Arabic and Russian was related to greater lexical variety and richer inflectional morphology. Similarly, Škrlj and Pollak (2019) compared token networks across nine languages. In these networks, nodes correspond to tokens, while edges encode textual relations between tokens in the corpus. Their results showed that different metrics capture different dimensions of similarity. The number of nodes separated morphologically rich languages such as Finnish and Estonian from more analytic languages such as English, while the clustering coefficient recovered clearer genealogical patterns, grouping Spanish with Portuguese and Slovene with Slovak.

These studies show that network measures can capture both typological and genealogical dimensions of linguistic structure, even when the networks are not semantic similarity networks in the strict sense used in the present thesis (Gao et al., 2014; Škrlj and Pollak, 2019).

More recent studies have moved closer to the type of question addressed in the present thesis by applying network measures to explicitly semantic and multilingual structures. Xu et al. (2023), for example, compared English and Mandarin Chinese networks built from feature norms and embeddings. Their results showed that Mandarin lexical networks have consistently higher clustering coefficients than English networks, which they interpret as evidence of a more tightly clustered lexical-semantic structure. Budel et al. (2023) similarly analyzed ConceptNet (Speer et al., 2017), a multilingual knowledge graph of words and phrases connected by labeled relations, across eleven languages and seven relation types. Their study considered properties such as the largest connected component, degree distribution, degree assortativity, and clustering coefficient, showing that different semantic relation types produce distinct topological patterns. They also argue that language-specific deviations in degree distributions may reflect grammatical inflection, since morphologically richer languages contain more distinct word forms, which in turn affect network structure.

Further work connects graph-based analysis more directly to multilingual embedding spaces. Fujinuma et al. (2019) propose graph modularity as a resource-free intrinsic measure for evaluating cross-lingual word embeddings. Their method constructs a graph from a shared multilingual embedding space and measures the extent to which words cluster by language. High modularity indicates that words remain separated by language, whereas lower modularity suggests better cross-lingual mixing. Casacuberta et al. (2021) develop a related measure, categorical modularity, based on a k-nearest-neighbor graph constructed from word embeddings. Instead of measuring whether words cluster by language, categorical modularity measures whether words belonging to the same semantic category tend to appear near one another in the graph. Using semantic categories across multiple languages and comparing models such as fastText, MUSE, and subs2vec, their study shows that graph-based properties of embedding spaces can provide useful information about semantic organization and downstream performance.

Another relevant line of work compares embedding-derived semantic networks with human-curated lexical resources. Veremyev et al. (2019) compare human-built semantic networks derived from lexical sources such as WordNet and Moby Thesaurus with machine-built semantic networks constructed from word2vec embeddings trained on Google News and Amazon Reviews. Their analysis considers both global and local network properties, including degree, path distance, clustering, and centrality. They show that human-built networks tend to display more intuitive global connectivity patterns, whereas embedding-based networks reveal richer local clusters that reflect contextual usage and perceived semantic similarity. This study is relevant because it demonstrates that semantic spaces derived from word embeddings can be transformed into networks and analyzed through graph-theoretic measures.

Despite these developments, embedding-based semantic networks have rarely been

used for broad cross-linguistic comparison or for examining whether semantic network structure reflects established genealogical subgroupings. Existing studies show that network properties can describe lexical organization, compare semantic resources, evaluate embedding spaces, and reveal structural differences between languages. However, less is known about whether the graph-theoretic properties of distributional semantic networks vary systematically across related languages. This gap motivates the present thesis, which compares semantic similarity networks derived from fastText embeddings across Indo-European languages and evaluates whether their structural properties correspond to established subgroupings.

3. Methodology

> "Essentially, all models are wrong, but some are useful."
>
> George Box

The previous chapter introduced the theoretical background of this thesis by discussing the Indo-European language family, word embeddings as distributional representations of lexical meaning, and the application of graph theory to semantic structure. Building on this theoretical foundation, the present chapter describes the methodological process used to compare semantic network structures across selected Indo-European languages. It outlines the data source, language selection, preprocessing of the embedding data, calculation of semantic similarity, construction of semantic networks, extraction of network properties for cross-linguistic comparison, and hierarchical clustering procedure.

This study applies a quantitative semantic network approach to investigate the structure of lexical meaning across languages. For each language, word embeddings are used as distributional representations of lexical items. Semantic similarity between words is calculated using cosine similarity and then transformed into language-specific semantic networks. In these networks, nodes represent words, and edges represent semantic similarity relations based on cosine similarity. The resulting networks are examined using semantic network properties to analyze whether historically related Indo-European languages exhibit similar patterns of semantic structure.

The methodological workflow consists of several stages. First, a subset of Indo-European languages is selected from the available subs2vec embeddings. The embedding data is then preprocessed to make the networks comparable across languages. Pairwise semantic similarities are calculated between word vectors. These similarity relations are transformed into semantic networks using two construction methods: a threshold-based approach and a mutual k- nearest-neighbor approach. The resulting networks are described through selected semantic network properties. Finally, hierarchical clustering is applied to these properties to examine whether the resulting language groupings correspond to established Indo-European subgroupings.

The purpose of this methodology is not to reconstruct a historical language tree in the strict phylogenetic sense. Rather, the aim is to test whether semantic structures derived from distributional data show patterns that correspond, at least partially, to known linguistic relationships among Indo-European languages. The study is therefore exploratory and comparative: it uses computational methods to examine whether structurally related languages also display similar organization in their semantic network properties. Figure 3.1 summarizes the methodological workflow of the present thesis. The code used for preprocessing, semantic network construction and analysis, hierarchical clustering, and

figure generation is available in the GitHub repository listed in Appendix A.1.

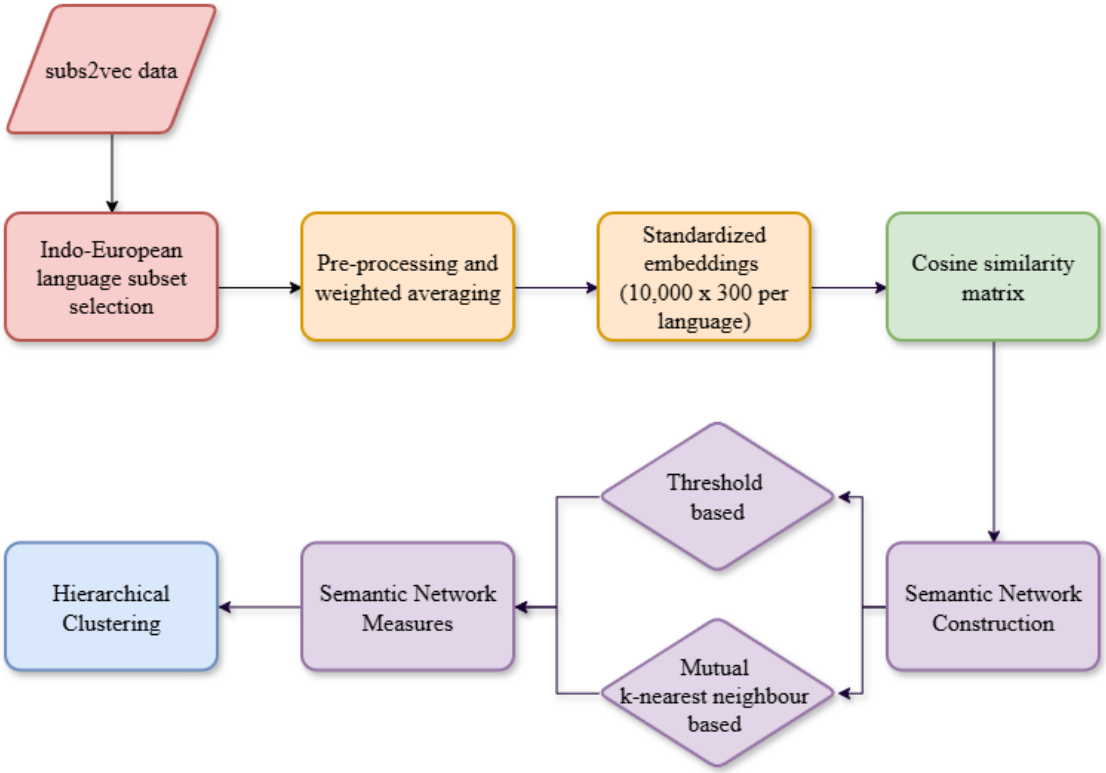

Figure 3.1.: Overview of the methodological workflow used in this thesis.

3.1. Data

This study uses subs2vec embeddings introduced by van Paridon and Thompson (2021). The dataset consists of pretrained word embeddings, trained and frequency information for 55 languages. The embeddings were trained on language-specific subsets of the OpenSubtitles corpus, which contains subtitles from films and television series. van Paridon and Thompson (2021) describe these subtitles as a form of pseudo-conversational language, since they are based on spoken dialogue, but are still written, edited, and often translated texts. For this reason, the corpus differs from more conventional written sources such as Wikipedia or newspaper corpora.

For the present study, the subtitle-based corpora have been selected because the aim is not to reconstruct historical language development from written historical texts, but to explore whether semantic network structures can be derived from modern distributional language data. Compared with encyclopedic or formal written corpora, subtitles provide a more dialogue-like source of language and therefore offer a useful basis for examining semantic organization in a pseudo-conversational context. This also fits the exploratory

nature of the study, which investigates whether modern distributional data can be used to compare semantic network structures across languages. However, this choice has limitations. Subtitles may include translated material, edited dialogue, genre-specific vocabulary, and language shaped by audiovisual context. They should therefore be understood as approximations of dialogue-like language rather than direct transcriptions of natural conversation.

The embeddings were trained using the fastText implementation of the skip-gram algorithm. Each word is represented as a 300-dimensional vector, which allows semantic similarity between lexical items to be calculated through vector-based measures. In addition to the embeddings themselves, van Paridon and Thompson (2021) provides related unigram, bigram, and trigram frequency resources, as well as code reproducing the models and analyses. These resources are available through authors' GitHub repository Van Paridon and Thompson (2020).

3.1.1. Language Selection

From the 55 languages available in the subs2vec dataset, this study selects Indo-European languages in order to compare semantic network structures against an established genealogical classification. The selection was based on two criteria. First, the language had to be available in the subs2vec dataset. Second, it had to belong to the Indo-European language family according to Glottolog. Glottolog provides a comprehensive catalogue of the world's languages, language families, and dialects, and was used here as the reference source for assigning each language to an Indo-European branch(Hammarström et al., 2026).

A metadata table was created for the selected languages. This table contains the language name, ISO 639-1 code, and Indo-European branch. The ISO codes were used for file handling and for matching language-specific frequency and embedding files,while the Glottolog-based branch labels were used for the interpretation of the clustering results.

A total of 37 Indo-European languages were initially identified in the subs2vec dataset. However, not all of these languages could be included in the final analysis. Since the network construction procedure required a fixed vocabulary size of 10,000 unigram tokens with valid embeddings for each language, languages that did not reach this threshold after preprocessing were excluded. Armenian and Breton were removed because, after filtering, fewer than 10,000 usable unigram embeddings were available. Afrikaans, Hindi, and Urdu were also excluded because the overlap between the frequency files and the available embedding vocabulary was too small to produce a final set of 10,000 valid tokens. The final analysis therefore includes only languages for which a comparable 10,000-token vocabulary with valid embeddings could be constructed.

Therefore, the final sample consists of 32 Indo-European languages. These languages represent several major branches of the family: Albanian, Greek [1], Balto-Slavic, Germanic, Italic, and Indo-Iranian. The selected languages and their Indo-European branch labels

[1] Glottolog classifies Greek within the Graeco-Phrygian grouping. In this thesis, the branch is referred to as Greek because Greek is the only representative of this grouping in the selected dataset.

are shown in Table 3.1.

Indo-European branch	**Language**	**Code**
Albanian	Albanian	sq
Balto-Slavic	Bosnian	bs
	Bulgarian	bg
	Croatian	hr
	Czech	cs
	Latvian	lv
	Lithuanian	lt
	Macedonian	mk
	Polish	pl
	Russian	ru
	Serbian	sr
	Slovak	sk
	Slovenian	sl
	Ukrainian	uk
Germanic	Danish	da
	Dutch	nl
	English	en
	German	de
	Icelandic	is
	Norwegian	no
	Swedish	sv
Greek	Greek	el
Indo-Iranian	Bengali	bn
	Farsi	fa
	Sinhala	si
Italic	Catalan	ca
	French	fr
	Galician	gl
	Italian	it
	Portuguese	pt
	Romanian	ro
	Spanish	es

Table 3.1.: Final Indo-European language sample used in the analysis, including ISO 639-1 codes and branch labels.

3.2. Preprocessing of the Data

After the Indo-European subset had been selected, the corresponding unigram frequency files and pretrained word embedding files were downloaded from the subs2vec GitHub repository (Van Paridon and Thompson, 2020). For each language, two files were used: the OpenSubtitles unigram frequency file, referred to as "word counts" in the repository, and the corresponding OpenSubtitles vector file containing the pretrained fastText embeddings. The frequency file contains unigram counts from the OpenSubtitles corpus and was used to rank tokens by frequency. The embedding files contain up to one million 300-dimensional word vectors per language.

The analysis was restricted to unigram tokens because the semantic networks in this thesis represent individual lexical items as nodes. Multiword expressions, such as bigrams and trigrams, were therefore not included in the network construction. This restriction makes the interpretation of nodes more direct, since each node corresponds to a single lexical item rather than to a phrase or multiword unit.

The aim of preprocessing was to create a comparable vocabulary and embedding matrix for each language. To do this, the frequency file and the embedding file were loaded separately and cleaned using the same token normalization procedure. All tokens were lowercased, Unicode normalization was applied, and typographic variants of apostrophes, hyphens, and spaces were standardized. Zero-width characters were also removed. This ensured that similar variants of the same written token were not treated as separate lexical items. After normalization, the frequency data were cleaned by removing rows with invalid or missing frequency values, and converting the remaining frequencies to integer counts. Duplicate unigram entries that became identical after normalization were grouped together, and their frequency counts were summed. This produced a single frequency value for each normalized token.

The embedding file was then processed in parallel. Each vector entry was checked to ensure that it contained exactly 300 dimensions, which corresponds to the dimensionality of the subs2vec fastText embeddings. Vector entries with an invalid dimensionality were removed. The remaining vector tokens were normalized using the same procedure as the frequency tokens.

Once both resources had been cleaned, the frequency table and the embedding table were merged. Only tokens that occurred in both files were retained. This ensured that every selected token had both a frequency value and an embedding vector. The merged table was then sorted by frequency, and the 10,000 most frequent tokens with valid embeddings were selected for each language. This fixed vocabulary size was used to make the initial network size comparable across languages, since each language contributes the same number of lexical nodes to the network construction stage.

Languages that did not reach the required number of 10,000 matched unigram tokens after preprocessing were excluded from the final analysis, as described in the language selection part. The final output of this preprocessing step was one vocabulary list and one $10{,}000 \times 300$ embedding matrix for each language. The preprocessing is summarized in Algorithm 1.

Data: Unigram frequency files and pretrained vector files for each selected language
Result: A vocabulary list and a 10,000 × 300 embedding matrix per language
foreach *selected language* **do**
  Load frequency file and vector file
  Normalize tokens in both files
  Clean frequency values and sum duplicate frequency entries
  Keep only valid 300-dimensional unigram vectors
  Merge duplicate vector entries for the same normalized token
  Join frequency and vector tables by normalized token
  Sort matched tokens by frequency
  if *at least 10,000 matched tokens are available* **then**
    Select the 10,000 most frequent tokens
    Save vocabulary list and embedding matrix
  else
    Exclude language from final analysis
  end
end

Algorithm 1: Preprocessing and vocabulary selection procedure.

3.2.1. Merging Multiple Embeddings

During normalization, some originally distinct vector entries became identical after lowercasing and token normalization. This can happen when capitalization differences or orthographic variants are removed. For example, German *Sie* and *sie* both become *sie* after lowercasing, although they may carry different grammatical or pragmatic meanings. Similarly, French forms such *l'ami* and *l'ami* may contain different apostrophe characters in the original data, but become identical after normalization.

Since each node in the semantic network requires one vector representation, these duplicate entries had to be merged. The aim was to produce one standardized embedding for each normalized token while retaining as much information as possible from the original vectors. This step also made the final vocabulary and embedding matrix consistent across languages. However, it may also remove distinctions that were originally expressed through capitalization or orthographic variation.

Three possible strategies were considered for merging multiple vector entries. The first option was to keep only the first available vector, but would ignore the remaining variants. The second option was to calculate a simple arithmetic average, but this would give equal weight to all variants, including variants that may be less representative of the group. The third option was to calculate a similarity-weighted average, which was used in the final preprocessing pipeline when more than two vectors were available for the same normalized token.

If only one vector was available, it was kept unchanged. If two vectors were available, their arithmetic mean was used. If more than two vectors were available, the similarity-

weighted average was calculated. For each group of vector entries associated with the same normalized token, pairwise cosine similarities were calculated, and each vector received a weight based on its total similarity to the other vectors in the group. Vectors that were more similar to the other variants therefore had more influence on the final representation, while less central variants had less influence. If the similarity-based weights were invalid, non-finite, or summed to zero, the process fell back to uniform averaging. This prevented missing or invalid values from being introduced into the final embedding matrix. The similarity-weighted averaging procedure is summarized in Algorithm 2.

```
Data: All vectors mapped to the same normalized token
Result: One merged vector for the token
Remove invalid vectors
if one vector remains then
    Return the vector unchanged
else
    if two vectors remain then
        Return the arithmetic mean
    else
        Compute pairwise cosine similarities between remaining vectors
        Set self-similarities to zero
        Compute vector weights from row sums of the similarity matrix
        if weights are invalid, non-finite, or sum to zero then
            Use uniform weights
        else
            Normalize weights to sum to one
        end
        Return the weighted average of the vectors
    end
end
```

Algorithm 2: Similarity-weighted averaging procedure for merging multiple embeddings assigned to the same normalized token.

3.3. Cosine Similarity Matrix

After preprocessing, each language was represented by a vocabulary of 10,000 unigram tokens and a corresponding 10,000 × 300 embedding matrix. The next step was to calculate the semantic similarity between all pairs of words within each language. This step was necessary because the semantic networks in this thesis are created on the basis of similarity relations between lexical items.

As discussed in the theoretical chapter, cosine similarity is commonly used in vector space models to measure the similarity between distributional vectors. In the present methodology, cosine similarity was applied to every pair of word vectors within each

language-specific embedding matrix, following Equation 2.3. Since each language contains 10,000 selected word vectors, each similarity matrix has the size of 10,000 × 10,000. Each row and column corresponds to one token in the selected vocabulary, and each cell represents the cosine similarity between two tokens. The diagonal values of the matrix represent the similarity of each word with itself and therefore are equal to 1.

The resulting matrices were saved as compressed NumPy files together with their corresponding word lists. This ensured that the same token order was preserved across the vocabulary list, the embedding matrix, and the similarity matrix during the later network construction stage.

3.4. Semantic Network Construction

After the cosine similarity matrices had been created, they were transformed into semantic networks. In these networks, unigram tokens are the nodes, and edges represent selected cosine-similarity links between tokens. Since the full cosine similarity matrix contains a value for every possible pair of tokens, retaining all pairwise similarities would produce a complete graph. This would not be useful for the present analysis, because even very weak similarities would be treated as meaningful connections. Therefore, only a subset of cosine-similarity values was retained as edges.

Two methods are used for network construction: a percentile-threshold method and a mutual k-nearest neighbor method. The threshold method retains similarities above a language-specific percentile cutoff, producing networks with a controlled global level of connectivity. The mutual k-nearest-neighbor method retains only reciprocal nearest-neighbor relations, producing networks based on local similarity structure. Using both methods makes it possible to examine whether the observed network patterns depend on the construction procedure or remain stable across different definitions of semantic relatedness.

The networks were constructed as undirected weighted graphs using igraph. In both methods, edge weights correspond to the original cosine similarity values between the connected tokens. The resulting graphs were saved in GraphML format, preserving both node labels and edge weights for the calculation of network properties.

3.4.1. Threshold-Based Network

In the threshold-based approach, an edge was created between two tokens if their cosine similarity was equal to or above a chosen percentile cutoff. This method retains only the strongest pairwise similarities and removes weaker connections. An initial trial used the 75th percentile as the threshold, but this produced networks that were too dense for the intended analysis. Since each language contains 10,000 nodes, even retaining 25% of all possible pairwise relations would result in millions of edges. Therefore, the final threshold-based networks used the 95th percentile as the cutoff, retaining only the strongest 5% of pairwise similarity values.

The threshold was calculated separately for each language using the off-diagonal values

of the upper triangle of the similarity matrix. This excluded self-similarities and avoided double-counting symmetric token pairs. The percentile threshold can be understood as a global construction rule because it is applied to the entire similarity matrix of a language, rather than to the neighborhood of each individual token. By retaining the same upper proportion of similarities for each language, the method controls the overall level of connectivity across networks. As a result, properties such as edge count, density, and average degree are expected to be very similar across languages.For every token pair above the 95th percentile cutoff, an undirected weighted edge was created, with the original cosine similarity value stored as the edge weight.

3.4.2. Mutual k-Nearest Neighbor Network

The second method used a mutual k-nearest neighbor approach. Unlike the percentile-threshold method, which applies one global cutoff to the similarity matrix, this method begins from the local similarity neighborhood of each token. For each token, the k most similar tokens were identified after self-similarities had been removed from the similarity matrix. In this thesis, $k = 1000$.

A standard k-nearest-neighbor graph would create an edge whenever one token appears in another token's nearest-neighbor list. This was not used as the final construction method because such one-sided relations can overemphasize hub tokens, which appear as nearest neighbors for many other tokens in high-dimensional spaces (Radovanović et al., 2010; Ozaki et al., 2011). This is especially relevant for high-dimensional language data, where distance concentration and hub effects can make standard or weighted k-nearest-neighbor graphs include noisy relations that do not accurately reflect local structure (Dalmia and Sia, 2021). For this reason, the standard k-nearest-neighbor approach was not used as the final graph construction method. Instead, a mutual criterion was applied: an edge between two tokens i and j was retained only if $i \in \text{kNN}(j)$ and $j \in \text{kNN}(i)$. In other words, two tokens were connected only when each appeared among the other's 1000 nearest neighbors.

This reciprocal criterion makes the mutual k-nearest-neighbor network more selective. Although every token starts with the same number of nearest-neighbor candidates, the final number of retained edges depends on how many of these relations are mutual. Unlike the percentile-threshold method, this approach therefore does not impose the same global density across languages. In the implementation, the diagonal of the similarity matrix was set to negative infinity so that a token could not be selected as its own nearest neighbor. The top k neighbors were sorted by descending cosine similarity, and each mutual edge was stored once as an undirected weighted edge. The original cosine similarity value was stored as the edge weight, and the reciprocal ranks of the two tokens in each other's nearest-neighbor lists were stored as additional edge attributes.

3.4.3. Semantic Network Properties

After the semantic networks had been constructed, a set of graph-theoretic properties was calculated for each language and for both network construction methods. The theoretical

definitions and formulas for these measures were introduced in Section 2.4.1. The present section therefore focuses on how these measures were applied to the constructed semantic networks and why they were selected for the cross-linguistic comparison.

The selected measures were chosen to capture different aspects of network structure. First, basic size and degree-based properties were calculated, includig number of nodes and edges, density, average degree, median degree, degree standard deviation, and maximum degree. Since all networks were constructured from an initial vocabulary of 10,000 tokens, the number of nodes was expected to remain constant across languages. Differences in edge count, density and degree-related measures reflect the way semantic similarity relations are distributed in each language and how these relations are retained by the network construction method.

Connectivity was examined through the number of connected components, the size of the largest connected component, and the largest component ratio. These measures were included because the constructed semantic networks are not guaranteed to be fully connected. Path-based measures, such as diameter and average shortest path length, require paths between node pairs. In disconnected graphs, some node pairs belong to different components and therefore have no connecting path (Newman, 2018). For this reason, diameter and average shortest path length were calculated on the largest connected component rather than on the full graph. This made the path-based measures comparable across languages while avoiding undefined values caused by disconnected nodes.

Clustering was measures using both the average local clustering coefficient and global transitivity. In this context, clustering refers to the extent to which neighboring tokens are also connected to one another. These measures were included because they describe whether semantic neighborhoods form locally dense groups. The average local clustering coefficient summarizes clustering around individual nodes, while transitivity measures triangle closure across the network as a whole. Together, they capture how strongly local semantic neighborhoods are clustered.

Community structure was measured using modularity and the number of detected communities. Community detection was carried out using the Leiden algorithm, which was chosen because it improves on the Louvain algorithm by producing well-connected communities and providing stronger guarantees about the quality of the resulting partition (Traag et al., 2019). In this thesis, modularity describes how strongly each semantic network can be divided into internally dense groups of tokens. The number of detected communities was recorded as an additional descriptive measure of how the network is partitioned.

Finally, degree assortativity was calculated to describe whether nodes tend to connect to other nodes with similar degrees. A positive value indicates that highly connected tokens tend to connect to other highly connected tokens, while a negative value indicates that highly connected tokens tend to connect to less connected, more peripheral tokens. This measure was included because it provides information about the degree-mixing pattern of the semantic network Newman (2003).

All measures were calculated using igraph and the full set of measure groups and network properties is summarized in Table 3.2. The results were saved in a summary

Measure group	Calculated properties
Basic properties	Number of nodes, number of edges, density, average degree, median degree, degree standard deviation, maximum degree
Connectivity	Number of components, largest component size, largest component ratio, diameter, average shortest path length
Clustering	Average local clustering coefficient, global transitivity
Community	Modularity, number of communities
Degree assortativity	Degree assortativity

Table 3.2.: Summary of the measure groups and network properties calculated for each semantic network.

table containing the language code, the network construction method, and the calculated network properties.

3.5. Hierarchical Clustering

Previous studies have shown that graph-theoretic properties of linguistic networks can be used as input for hierarchical clustering and language classification. Liu and Li (2010) constructed syntactic dependency networks for fifteen languages and used network parameters such as average degree, clustering coefficient, average path length, network centralization, diameter, and degree-distribution measures as clustering features. Their results showed that several language groupings, especially among Romance languages, could be partly recovered from network topology. Abramov and Mehler (2011) developed a more systematic version of this approach by calculating 21 topological coefficients from syntactic dependency networks and treating these coefficients as language-level vectors for automatic classification. Similarly, Liu and Cong (2013) used word co-occurrence networks based on parallel texts and applied cluster analysis to network parameters, showing that Slavic languages could be separated from non-Slavic languages and, in several cases, grouped into Eastern, Western, and Southern sub-branches.

More recently, Schrader and Gultepe (2023) applied hierarchical clustering to verse-level document vector representation based on Bible translations in 22 Indo-European languages . Their aim was to examine whether vector-based representation of parallel texts could recover Indo-European subgroupings. The results shows strong clustering performance for Indo-Iranian and Slavic languages, while Germanic and Romance languages were more difficult to separate clearly. These studies show that clustering can be used to examine whether quantitative representations of languages reflect known linguistic groupings. However, most previous studies are based on syntactic dependency networks, word co-occurrence networks, or document-level vector representations. The present thesis extends this line of work by applying hierarchical clustering to graph-theoretic properties derived from word embedding-based semantic networks, using a broader sample of Indo-European

languages.

Figure 3.2 provides a reference structure used for interpreting the clustering result in this thesis. It is a simplified overview of the Indo-European languages included in the dataset and shows their broad subgroup placement based on Glottolog. It should not be interpreted as a precise phylogenetic tree, and the visual distances between languages should not be read as exact historical or linguistic distances. Instead, the figure serves as a schematic reference for the major branches and sub-branches represented in the analysis.

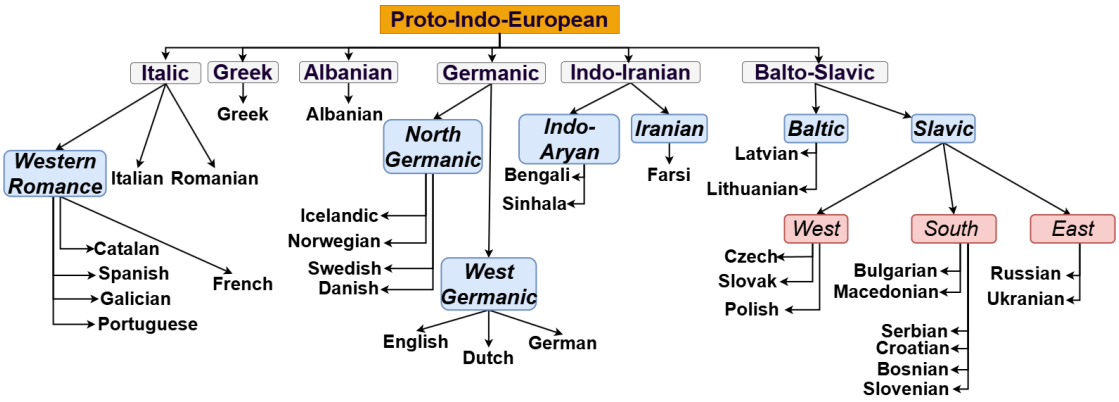

Figure 3.2.: Simplified Indo-European subgroup structure of the languages included in this study. Arrows indicate branch and sub-branch membership rather than exact historical or linguistic distance.

At the highest level, the languages are divided into six main branches: Italic, Greek, Albanian, Germanic, Indo-Iranian, and Balto-Slavic. The arrows in the figure indicate subgroup membership. They show how individual languages are placed under broader branches, and where relevant, under smaller sub-branches.

The Italic branch includes Italian, Romanian, French, and the Western Romance languages. In the figure, Catalan, Spanish, Galician, and Portuguese are grouped under Western Romance, indicating that they are closer to each other within the dataset than to the other Italic languages shown. French is also part of the Romance group, but it is represented as somewhat separate from the Western Romance cluster.

The Germanic branch is divided into North Germanic and West Germanic. North Germanic includes Icelandic, Norwegian, Swedish, and Danish. The figure shows closer pairings within this group, such as Icelandic with Norwegian and Swedish with Danish, while still keeping all four languages under the broader North Germanic branch. West Germanic includes English, Dutch, and German. These languages are grouped together under the same sub-branch, but they are not shown to form the same kind of close pairings as some North Germanic languages do.

The largest number of languages in the dataset belongs to the Balto-Slavic branch. This branch is divided into Baltic and Slavic. The Baltic group includes Latvian and Lithuanian. The Slavic group is further divided into West, South, and East Slavic. West

Slavic includes Czech, Slovak, and Polish, with Czech and Slovak represented as especially close to each other. South Slavic includes Bulgarian, Macedonian, Serbian, Croatian, Bosnian, and Slovenian. Within this group, Bulgarian and Macedonian are shown as a closer pair, while Serbian, Croatian, Bosnian, and Slovenian are grouped together as another closely related set. East Slavic includes Russian and Ukrainian, which are shown as a close pair.

The Indo-Iranian branch is divided into Indo-Aryan and Iranian. Bengali and Sinhala are placed under Indo-Aryan, while Farsi is placed under Iranian. Greek and Albanian are represented as separate single-language branches in the dataset.

The hierarchical clustering results are interpreted in relation to this reference figure. Ideally, if the semantic network properties reflect genealogical relationships, languages from the same branch or sub-branch would cluster together. For example, one might expect Germanic languages to group with other Germanic languages, and Slavic languages with other Slavic languages. However, the goal is not to reconstruct the Indo-European family tree directly. Rather, the figure provides a reference point for evaluating whether the graph-theoretic properties of the semantic networks produce groupings that partially resemble known Indo-European branches and sub-branches.

For the clustering step, each language was represented as a feature vector made up of selected graph-theoretic measures. However, not all calculated measures were used directly. Since hierarchical clustering is based on distances between languages, including several highly correlated measures could overrepresent the same structural dimension. For example, edge count, density, and average degree all describe closely related aspects of overall graph connectivity. To reduce this redundancy, Spearman's rank correlation was first calculated among the graph-theoretic measures separately for the threshold networks and the mutual k-nearest neighbor networks. Measures with an absolute Spearman correlation of 0.90 or higher were treated as strongly redundant.

After the reduced feature sets were selected, the values were standardized using z-score normalization. This was necessary because the graph measures are measured on different numerical scales. For example, number of communities and diameter may have much larger numerical ranges than density or modularity. Without standardization, measures with larger values could dominate the distance calculation.

Hierarchical clustering was then applied separately to the threshold-based networks and the mutual kNN networks. The selected measures were first extracted into a feature matrix, where each row represents one language and each column represents one network measure. The matrix was then standardized. Pairwise distances between languages were calculated using Euclidean distance on the standardized feature values. Ward linkage was then used as the linkage criterion, because it merges clusters by minimizing the increase in within-cluster variance at each step(Ward, 1963). This is useful for identifying compact groups of languages with similar overall network profiles. Finally, optimal leaf ordering was applied to make the dendrograms easier to read, without changing the underlying hierarchical clustering structure.

The resulting dendrograms were interpreted in relation to the Indo-European reference figure. The analysis focuses on whether major branches such as Germanic, Italic, Balto-

Slavic, and Indo-Iranian form recognizable clusters in semantic network space. The height at which languages or groups merge in the dendrogram indicates how dissimilar they are according to the selected standardized graph-theoretic measures. Therefore, the dendrograms should be interpreted as representations of similarity between semantic network profiles, not as direct representations of historical linguistic descent.

4. Results

> "Everything is related to everything else but near things are more related than distant things."
>
> Waldo Tobler

This section presents the results of the semantic network analysis. The findings are organized according to the main groups of graph measures used in the study: basic network properties, connected components, clustering and transitivity, community structure, and degree assortativity. For each group, the two graph-construction methods are compared in order to analyze both method-specific effects and cross-linguistic differences.

All networks were constructed from the same initial vocabulary size, with each language represented by 10,000 unigram tokens. This controls for differences in network size and makes it possible to compare the resulting graph structures across languages. However, the two construction methods impose different constraints on the networks. The percentile threshold method retains edges according to a fixed similarity threshold defined separately for each language, whereas the mutual k-nearest neighbor retains only reciprocal nearest neighbor relations. These methodological differences are therefore taken into account when interpreting the network properties.

The graph-theoretic measures are interpreted at both the individual language level and the Indo-European branch level in order to examine whether languages within the same branches display similar network properties. After the individual measures and their correlations have been examined, selected network properties are combined into language-level profiles for each construction method. Hierarchical clustering is then applied to these profiles in order to assess whether languages with similar semantic network properties form broader groups, and whether these groups correspond to established Indo-European subgroupings.

4.1. Basic Network Properties

These measures describe the general size and edge structure of the semantic networks. The main basic properties considered here are the number of edges, density, average degree, median degree, degree standard deviation, and maximum degree. These measures provide a first overview of how many semantic similarity relations are retained in each network and how these relations are distributed across nodes.

Because of the way the networks were constructed, some basic properties are strongly shaped by the construction method itself. In the percentile-threshold networks, the number of edges, density, and average degree are expected to be highly similar across

languages, because each network retains approximately the same proportion of the strongest similarity relations. The results confirm this expectation: the threshold networks contain approximately 2.5 million edges, have a density of about 0.05, and have an average degree of about 500 across all languages. These measures therefore mainly reflect the fixed thresholding procedure rather than substantial cross-linguistic differences.

In the mutual kNN networks, by contrast, the total number of edges, density, and average degree are not fixed by the construction method. Although $k = 1000$ limits the maximum possible degree of each node, an edge is only retained when the nearest-neighbor relation is reciprocal. As a result, the final edge structure can vary across languages depending on how mutually concentrated or dispersed the semantic neighborhoods are.

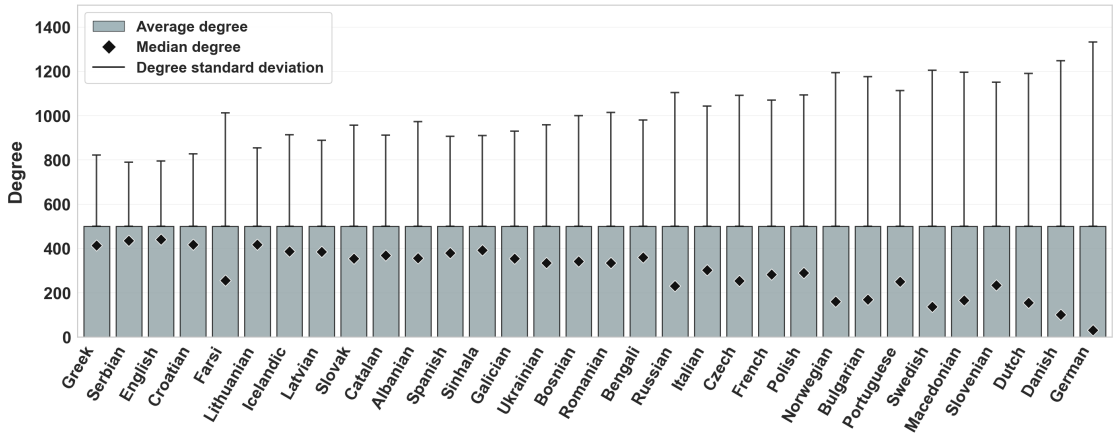

Figure 4.1.: Average degree, median degree, and degree standard deviation in the threshold networks.

Figure 4.1 shows the average degree, median degree, and degree standard deviation in the threshold networks. The bars represent average degree, the diamond markers represent median degree, and the vertical T-lines show the standard deviation of node degrees. The languages are ordered by decreasing average in the mutual kNN networks, and the same order is kept for the threshold networks to make the two figures easier to compare.

As expected from the threshold construction, the average degree is approximately 500 for all languages. This reflects the fact that each threshold network retains the same proportion of strongest similarity relations. However, the median degree and degree standard deviation vary across languages. Germanic languages show the clearest deviation from the overall patterns. German has the lowest median degree, at 31, followed by Danish at 102, Swedish as 138, and Dutch at 152. These languages also have high degrees of standard deviations, with values for German the highest at 833, followed by Danish, Swedish, and Dutch. This shows that retained edges are distributed unevenly across nodes in these languages.

By contrast, Greek, Serbian, and English have much higher median degrees with values 413, 436, and 442, respectively. Their median values are therefore much closer to the

fixed average degree of approximately 500, suggesting a more even distribution of edges across the network. Farsi is an interesting intermediate case. It has a comparatively low median degree together with high degree variation, indicating that its threshold network contains a more uneven degree structure despite having the same average degree as the other languages.

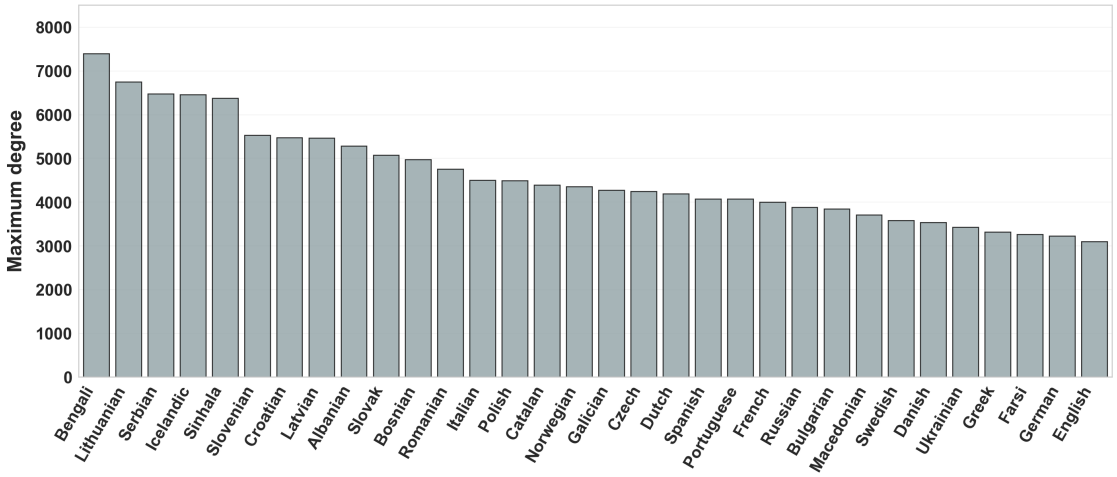

Figure 4.2.: Maximum degree in the threshold networks.

Figure 4.2 shows the maximum degree in the threshold networks. Maximum degree refers to the number of connections held by the single most connected node in each network. In this context, it indicates how strongly the thresholding procedure concentrates edges around highly connected words. Bengali is the clearest high outlier, with a maximum degree of approximately 7400, followed by Lithuanian, Serbian, and Icelandic. This means that in these networks, at least one word is connected to a very large proportion of all other words. At the lower end, English has the lowest maximum degree, at approximately 3098, followed by German, Farsi, and Greek. These lower values suggest that the most connected nodes in these networks are less dominant, even though the total number of edges and average degree remain almost the same across languages.

Figure 4.3 shows the average degree, median degree and degree standard deviation in the mutual kNN networks. In contrast to the threshold networks, the average degree is not fixed across languages. German is the clearest low outlier, with an average degree of 356 and a median degree of 304. This means that German words retain fewer reciprocal nearest-neighbor connections on average than words in the other language networks. Its degree standard deviation of 246 further indicates that these connections are unevenly distributed across nodes, with some words having substantially more connections than others. Danish and Dutch follow German at the lower end of the distribution, suggesting similarly sparse mutual kNN networks.

At higher end, Greek, Serbian, and English show the largest degree values. Greek has an average degree of approximately 757 and a median degree of 780, with a degree standard deviation of 150. Serbian has an average degree of approximately 743 and a

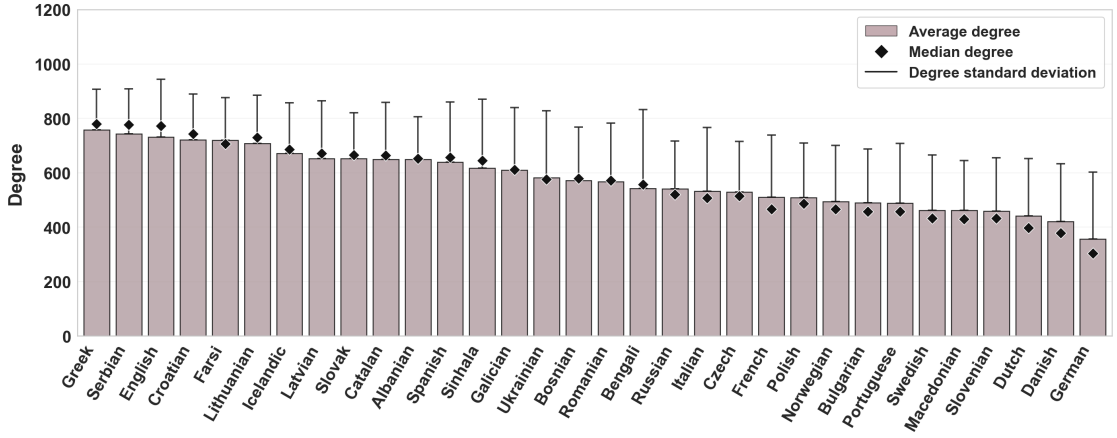

Figure 4.3.: Average degree, median degree, and standard deviation degree in the mutual kNN networks.

median degree of 777, with a standard deviation of 164, while English has an average degree of approximately 730 and a median degree of 773, with a standard deviation of 212. These standard deviation values indicate some variation in node degrees, but relative to the high average degree values they suggest comparatively balanced degree distributions. In other words, these networks retain many reciprocal nearest-neighbor connections overall, without this connectivity being concentrated only in a small number of highly connected nodes.

Across most languages, the median degree is relatively close to the average degree. This suggests that the connectivity of a typical node is close to the overall network mean, rather than being dominated by only a small number of very highly connected nodes. The degree standard deviation values are not identical across languages, but they remain moderate compared with the threshold networks. This indicates that the mutual kNN method produces more balanced degree distributions while still allowing cross-linguistic differences in overall connectedness.

Figure 4.4 shows the number of edges and density in the mutual kNN networks. Edge count and density express the same overall pattern in absolute and normalized form. Languages with more retained edges necessarily also have higher density.

This figure confirms the pattern observed in the average degree results. Greek has the highest number of edges, with more than 3.8 million retained edges, corresponding to a density of approximately 0.078. Serbian and English also have high edge counts and densities, reflecting the large number of reciprocal nearest neighbor relations retained in these networks. German is again the clearest low outlier, with approximately 1.78 million edges and a density of 0.036. Danish and Dutch follow German at the lower end of the distribution.

These results show that the mutual kNN construction produces substantial variation in overall network density across languages. This variation is expected, because $k = 1000$ only sets an upper limit on the possible degree of each node; the final number of edges

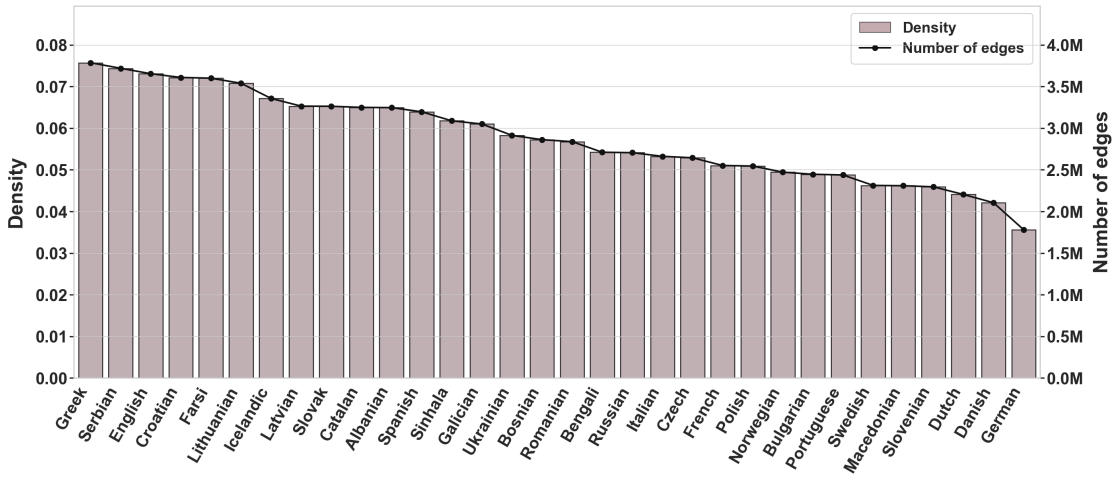

Figure 4.4.: Number of edges and density for the mutual kNN networks.

depends on how many nearest-neighbor relations are reciprocal. The density and edge-count results therefore support the average-degree pattern: Greek, Serbian, and English form denser mutual kNN networks, whereas German, Danish, and Dutch form sparser ones.

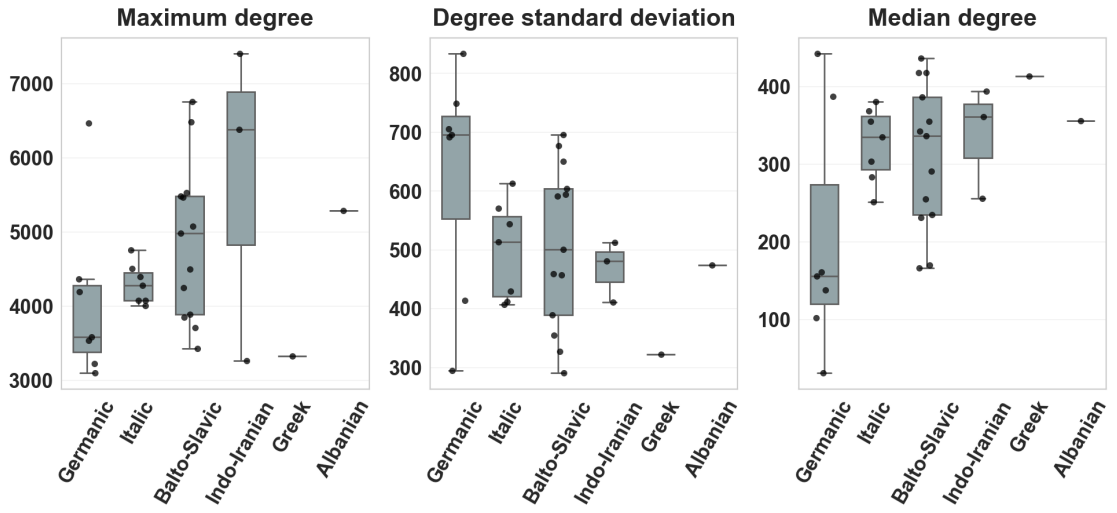

Figure 4.5.: Indo-European branch level distribution of selected basic network properties in the threshold networks. Points represent individual languages, while boxplots summarize the distribution within each branch.

Figure 4.5 and 4.6 compare selected basic network properties across Indo-European branches for the threshold and mutual kNN networks. Albanian and Greek are represented by only one language each, and therefore serve as individual reference points rather than branch-level distributions. The Indo-Iranian branch also contains only three languages,

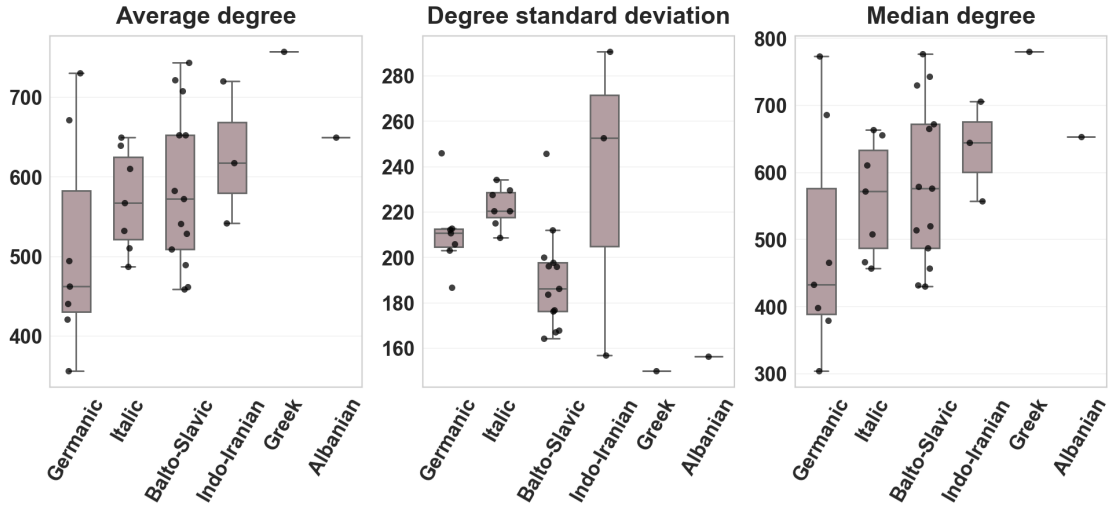

Figure 4.6.: Indo-European branch level distribution of selected basic network properties in the mutual kNN networks. Points represent individual languages, while boxplots summarize the distribution within each branch.

so its branch-level patterns should be interpreted cautiously.

In the threshold networks, the branch-level differences are most visible in median degree, degree standard deviation, and maximum degree, since average degree is fixed by the construction method. The Germanic branch shows the clearest internal variation. It includes some of the lowest median-degree values and some of the highest degree standard deviations, indicating that threshold edges are distributed unevenly across nodes in several Germanic languages. This pattern is consistent with the individual-language results, where German, Danish, and Dutch appear as particularly uneven networks. By contrast, the Italic branch is more compact, especially in median degree, suggesting more similar degree distributions within this group. Balto-Slavic languages show a broader spread across all three measures, which is partly expected because this is the largest branch in the dataset. Indo-Iranian languages also show high maximum-degree values, but the small number of languages makes this pattern less stable as a branch-level result.

The maximum-degree plot further shows that some threshold networks contain highly connected hub nodes. Indo-Iranian and Balto-Slavic languages include some of the highest maximum-degree values, while English, German, Farsi, and Greek appear at the lower end in the individual-language results. This means that, although the threshold networks have similar average degree overall, some languages concentrate many edges around a small number of highly connected nodes, whereas others distribute edges more evenly.

In the mutual kNN networks, the branch-level pattern is different because average degree is no longer fixed. Here, Germanic again shows the widest internal spread, ranging from low-degree networks such as German, Danish, and Dutch to more highly connected cases such as English and Icelandic. This suggests that the Germanic branch is not structurally uniform in terms of mutual nearest-neighbor relations. Italic languages are

more compact, with average and median degree values concentrated in a narrower range. Balto-Slavic languages show moderate to broad variation, while Indo-Iranian languages tend to occupy relatively high ranges but also show noticeable spread in degree standard deviation.

Overall, the branch-level plots suggest that the two construction methods emphasize different aspects of basic network structure. The threshold method highlights differences in how evenly edges are distributed, especially through median degree, degree standard deviation, and maximum degree. The mutual kNN method instead highlights differences in overall connectedness, visible through average degree and median degree. Across both methods, the Germanic branch shows the strongest internal variation, mainly because German, Danish, and Dutch differ from English and Icelandic. Italic languages appear more internally compact, while Balto-Slavic languages show broader variation without the same extreme pattern as Germanic. These results suggest that some branch-level tendencies are visible, but they are often driven by specific languages rather than by entire Indo-European branches as uniform groups.

4.2. Connectivity

The connectivity analysis combines component-based and path-based measures. The component-based measures include the number of connected components, the size of the largest connected component, and the largest component ratio. These measures indicate whether the networks remain connected as one large structure or whether some nodes are separated into smaller disconnected subgraphs. Since all networks contain 10,000 nodes, largest component size and largest component ratio express the same information in absolute and normalized form. The path-based measures, average shortest path length and diameter, then describe how far apart nodes are within the connected part of the network.

The data shows that the component-based measures show that the mutual kNN networks are fully connected for all languages. Each mutual kNN network consists of a single connected component, with all 10,000 nodes included in the largest component and a largest component ratio of 1.0. The threshold networks are also largely connected, but some languages show minor fragmentation. German is the clearest one, with 112 connected components and a largest component ratio of 0.9882, meaning that 9,882 nodes belong to the largest component, while a small number of nodes are separated into smaller disconnected components. Danish also shows some fragmentation, with 15 components and a largest component ratio of 0.9986. Swedish and Portuguese each have 6 components, while Dutch and Bengali have 3 components. These results suggest that the threshold method can produce small disconnected areas in some language networks, although the largest component still contains almost all nodes in every case.

Figures 4.7 and 4.8 compare the path-based connectivity measures for the threshold and mutual kNN networks. Average shortest path length describes the typical distance between nodes in the largest connected components, while diameter described the longest shortest path between any two nodes. In both figures, languages are ordered by decreasing

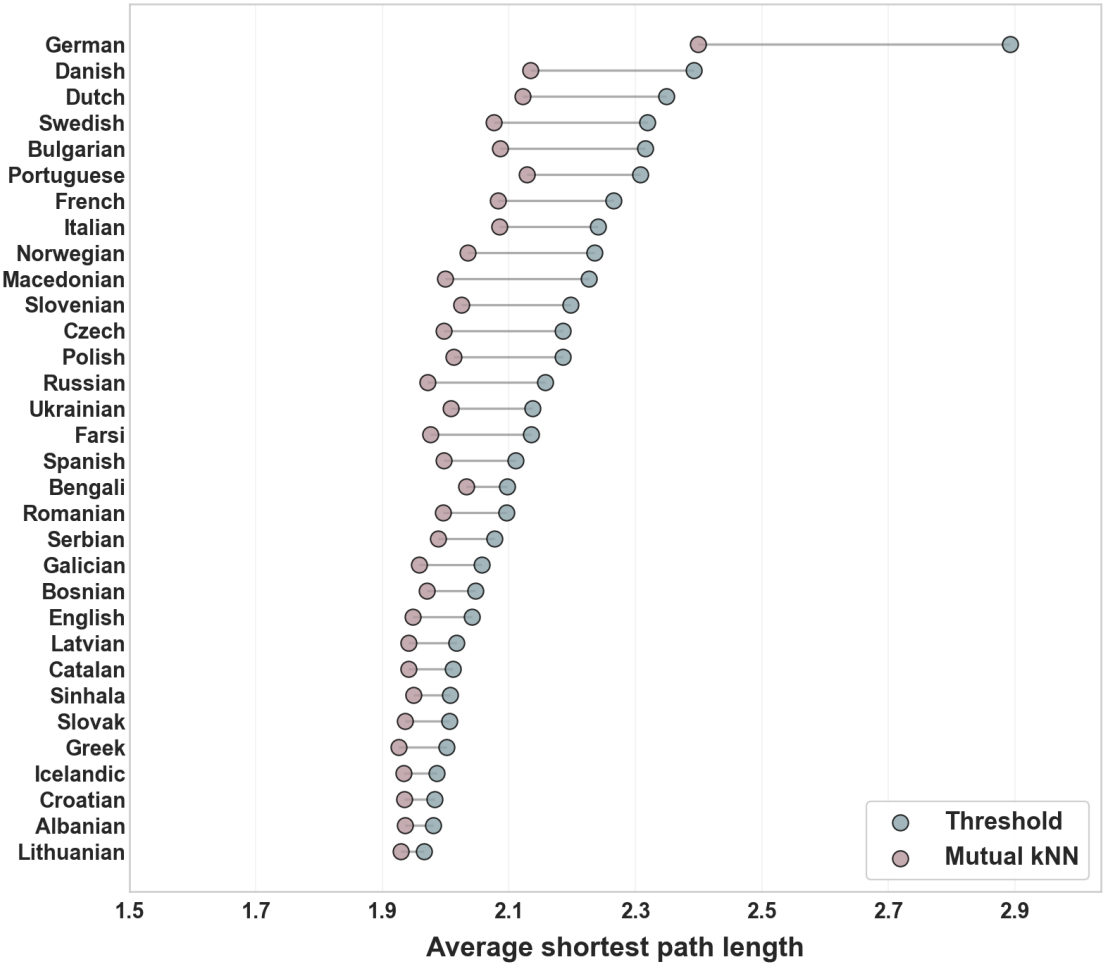

Figure 4.7.: Average shortest path length in threshold and mutual kNN networks. Languages are ordered by decreasing threshold values.

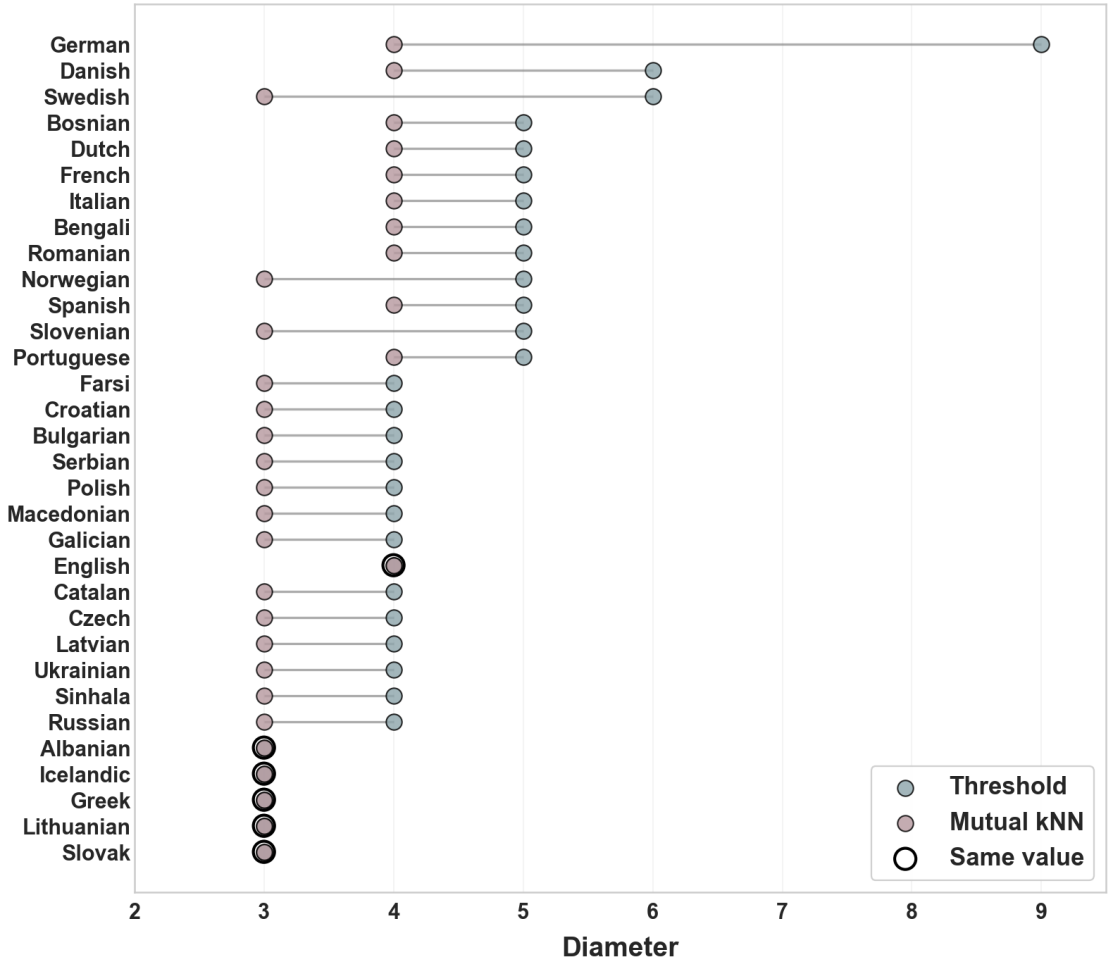

Figure 4.8.: Diameter in threshold and mutual kNN networks. Languages are ordered by decreasing threshold values.

threshold values.

Figure 4.7 shows that average shortest length path is generally higher in the threshold networks than in the mutual kNN networks. German is the clearest outlier, with an average shortest path length of approximately 2.89 in the threshold network, compared with approximately 2.40 in the mutual kNN network. This indicates that nodes in the German threshold network are, on average, farther apart than in the other language networks. Danish, Dutch, and Swedish also show relatively high threshold values, although their average shortest path lengths remain below 2.5.

The differences between the two construction methods are more compressed for average shortest path length than for diameter. In the mutual kNN networks, most languages cluster between approximately 1.9 and 2.3, suggesting that these networks have relatively compact path structures across languages. Lithuanian, Albanian, and Croatian show some of the lowest average shortest path length values, and their threshold and mutual kNN values are very close together. This suggests that, for these languages, the choice of construction method has little effect on distances.

Figure 4.8 shows whether the same patterns also appears in the most extreme path distances. The diameter results confirm that threshold networks generally produce longer paths than mutual kNN networks. German is again the clearest outlier, with a threshold diameter of 9 compared with a mutual kNN diameter of 4. Danish and Swedish also stand out, with threshold diameters of 6, while Norwegian increases from 3 in the mutual kNN network to 5 in the threshold network. These cases indicate that some Germanic threshold networks contain longer peripheral path structures, even though their largest connected components still include almost all nodes.

By contrast, most mutual kNN diameters are concentrated between 3 and 4. This supports the average shortest path length results: mutual kNN networks are generally more compact and more consistent across languages. Several languages, including English, Albanian, Icelandic, Greek, Lithuanian, and Slovak, have identical diameters under both construction methods, suggesting that their overall path structure remains stable regardless of the network construction method.

Figures 4.10 and 4.9 show the Indo-European family branches with regard to average shortest path length and diameter in threshold and mutual kNN networks. Albanian and Greek were excluded from these visualizations because both languages show relatively compact path structures, with a diameter of 3 and low average shortest path length values under both construction methods. Excluding them allows a clearer comparison betwen the four larger Indo-European branches.

Average shortest path length results show that the mutual kNN networks generally produce shorter paths than the threshold networks. The clearest branch-level variation appears in the Germanic languages. German is the strongest outlier, with the largest separation between the threshold and mutual kNN values, followed by Danish, Dutch, and Swedish. This suggests that several Germanic threshold networks are less compact than their mutual kNN counterparts. By contrast, the Balto-Slavic languages are more tightly grouped, with most languages showing relatively similar average shortest path lengths. The Italic languages also form a compact group, although French and Portuguese show

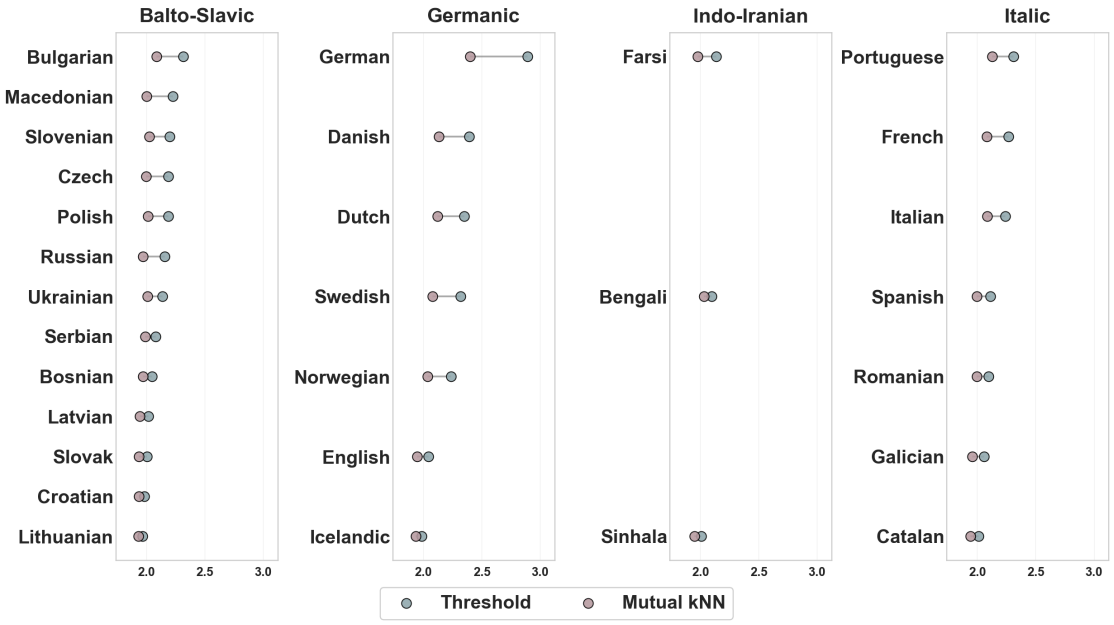

Figure 4.9.: Average shortest path length by Indo-European branch in threshold and mutual kNN networks.

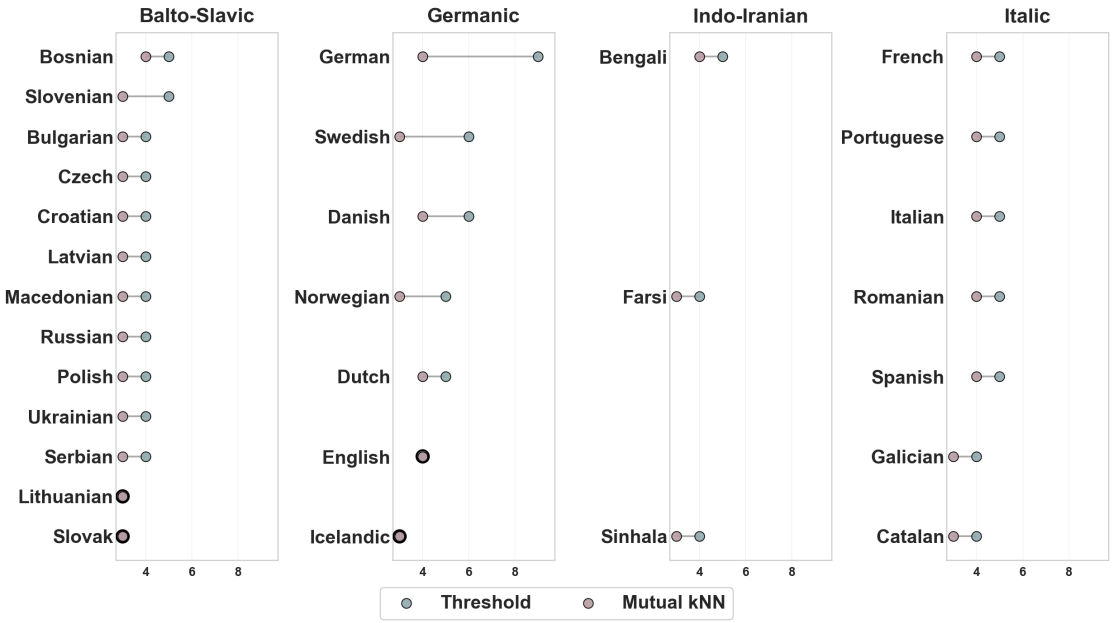

Figure 4.10.: Diameter by Indo-European branch in threshold and mutual kNN networks.

slightly higher threshold values than the other Romance languages. Within Indo-Iranian, Bengali has a higher threshold value than Farsi and Sinhala, but the small number of languages means that this pattern should be interpreted cautiously.

The diameter results in Figure 4.10 show a similar pattern in the longest shortest paths on a branch level. German is again the most extreme case, with a much larger threshold diameter than mutual kNN diameter. Swedish, Danish, and Norwegian also show larger threshold diameters, reinforcing the pattern that Germanic languages contain the most pronounced method-based differences in path structure. In Balto-Slavic, most languages have similar mutual kNN diameters, while threshold diameters vary slightly more. The Italic branch remains comparatively compact, with only moderate differences between the two methods. Indo-Iranian again shows some internal variation, especially because Bengali has a larger threshold diameter than Farsi and Sinhala.

Overall, the connectivity results suggest that both methods produce largely connected semantic networks, but the mutual kNN method yields more homogeneous path structures across languages. The threshold method reveals stronger cross-linguistic variation, especially in the Germanic branch. German, Danish, Dutch, Swedish, and Norwegian show longer path structures under the threshold construction, whereas Balto-Slavic and Italic languages are generally more compact and less strongly affected by the choice of construction method.

4.3. Clustering

The clustering structure of the semantic networks is evaluated using two related measures: average local clustering coefficient and global transitivity. The average local clustering coefficient summarizes how strongly the neighbors of individual nodes are connected to one another. This indicates whether the local neighborhoods of words form cohesive groups, rather than loose sets of unrelated neighbors. Global transitivity is closely related, but it measures triangle closure across the whole network by comparing the number of closed triples to all connected triples. Thus, while the average local clustering coefficient summarizes local neighbourhood density across nodes, transitivity captures triangle closure at the network level.

Figures 4.11 and 4.12 show the relationship between the average local clustering coefficient and global transitivity for the threshold and mutual kNN networks. Each point represents one language, and colors indicate Indo-European branch membership. The transparent ellipses are used as visual aids to highlight groups of languages from the same branch that occupy similar regions of the clustering space. They should be interpreted descriptively rather than as formal statistical cluster boundaries.

In the threshold networks, average local clustering coefficient ranges from approximately 0.28 to 0.54, while transitivity ranges from approximately 0.24 to 0.72. German is the clearest high outlier in both measures, with an average local clustering coefficient of 0.53 and transitivity of 0.72. Dutch and Danish also show high average local clustering values, while Norwegian, Danish, and Swedish show high transitivity values. This indicates that several Germanic threshold networks contain strong triangle closure, although the branch

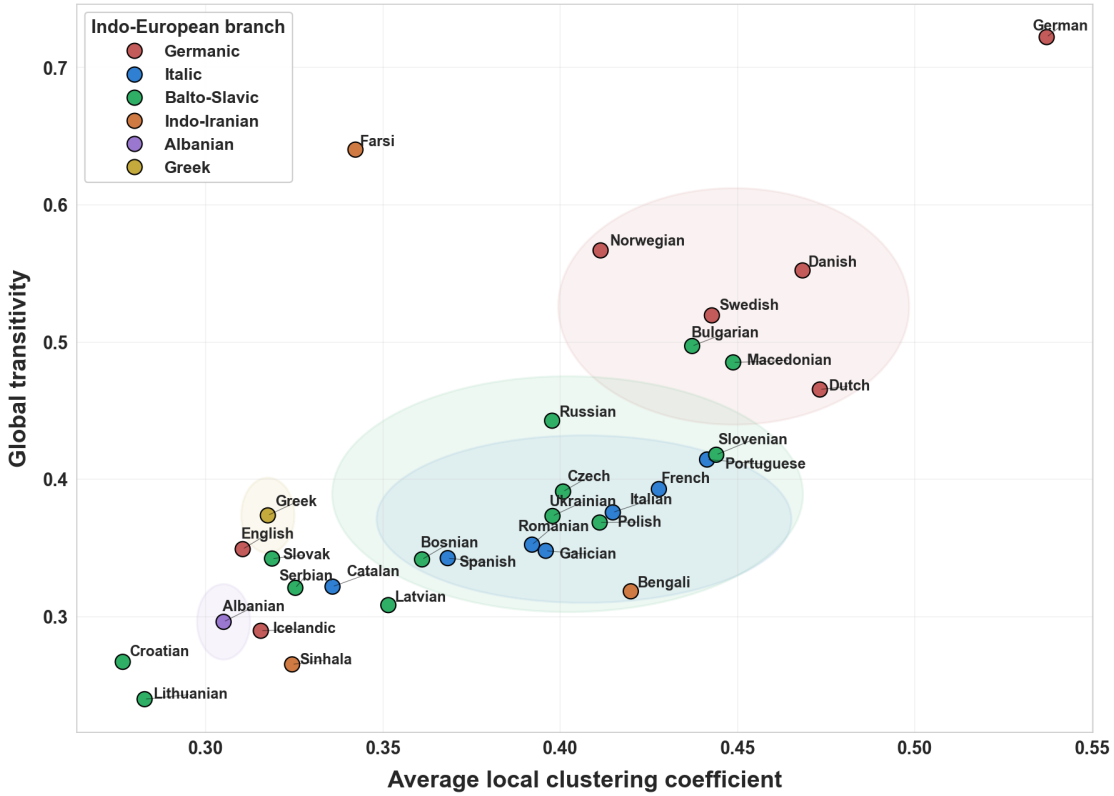

Figure 4.11.: Average local clustering coefficient and global transitivity for threshold networks.

is not uniform, since English and Icelandic lie much lower in the plot.

The Balto-Slavic languages are more dispersed. Bulgarian and Macedonian occupy the higher region of the threshold plot, with relatively high values for both clustering measures. By contrast, Croatian and Lithuanian lie at the lower end, with average local clustering coefficients of 0.27 and 0.28 and transitivity values of 0.26 and 0.24, respectively. Most other Balto-Slavic languages, including Polish, Czech, Ukrainian, and Russian, occupy a more central range. This suggests substantial internal variation within the Balto-Slavic branch.

The Italic languages form a more compact middle range. Their average local clustering coefficients range from approximately 0.34 to 0.44, and their transitivity values range from approximately 0.32 to 0.41. Portuguese and French are located toward the higher end of the Italic group, while Catalan has the lowest values. Overall, the Italic branch shows a more coherent clustering pattern than Germanic or Balto-Slavic in the threshold networks.

The Indo-Iranian languages do not form a tight group. Farsi is particularly notable because it combines moderate average local clustering, at 0.34, with very high transitivity, at 0.64. This suggests that Farsi has strong global triangle closure that is not fully reflected by the average local clustering coefficient alone. Bengali occupies a more central position, while Sinhala lies in the lower range of both measures. Albanian and Greek also appear in the lower-to-middle part of the threshold plot.

In the mutual kNN networks, the overall range of values is narrower. Average local clustering coefficient ranges from approximately 0.21 to 0.30, while transitivity ranges from approximately 0.22 to 0.47. This indicates that the mutual kNN construction produces more compressed clustering patterns across languages than the threshold construction.

The Italic languages occupy the higher range for average local clustering in the mutual kNN networks. Spanish has the highest average local clustering coefficient overall, at 0.30, followed by Portuguese, Galician, Italian, and French. German is again a notable case: although it does not have the highest average local clustering coefficient, it has the highest transitivity value, at 0.46. This means that German again shows especially strong global triangle closure. Danish, Bulgarian, Portuguese, Swedish, and Dutch also occupy relatively high positions in transitivity.

The Balto-Slavic languages are again spread across a broad range. Serbian and Bulgarian appear toward the higher end for average local clustering, while Lithuanian, Slovak, and Croatian occupy the lower end. For transitivity, Bulgarian is the highest Balto-Slavic language, while Lithuanian and Croatian are among the lowest. Indo-Iranian languages occupy mostly middle-to-lower positions, with Farsi and Bengali above Sinhala. Albanian is the lowest language for average local clustering in the mutual kNN networks.

Overall, the clustering results show that the two construction methods emphasize different patterns. The threshold networks produce a wider spread of clustering and transitivity values, with German and Farsi standing out as strong high-transitivity cases. The mutual kNN networks are more compressed, suggesting more homogeneous clustering structures across languages. At the branch level, Italic languages are comparatively compact, especially in the mutual kNN networks, while Germanic and Balto-Slavic

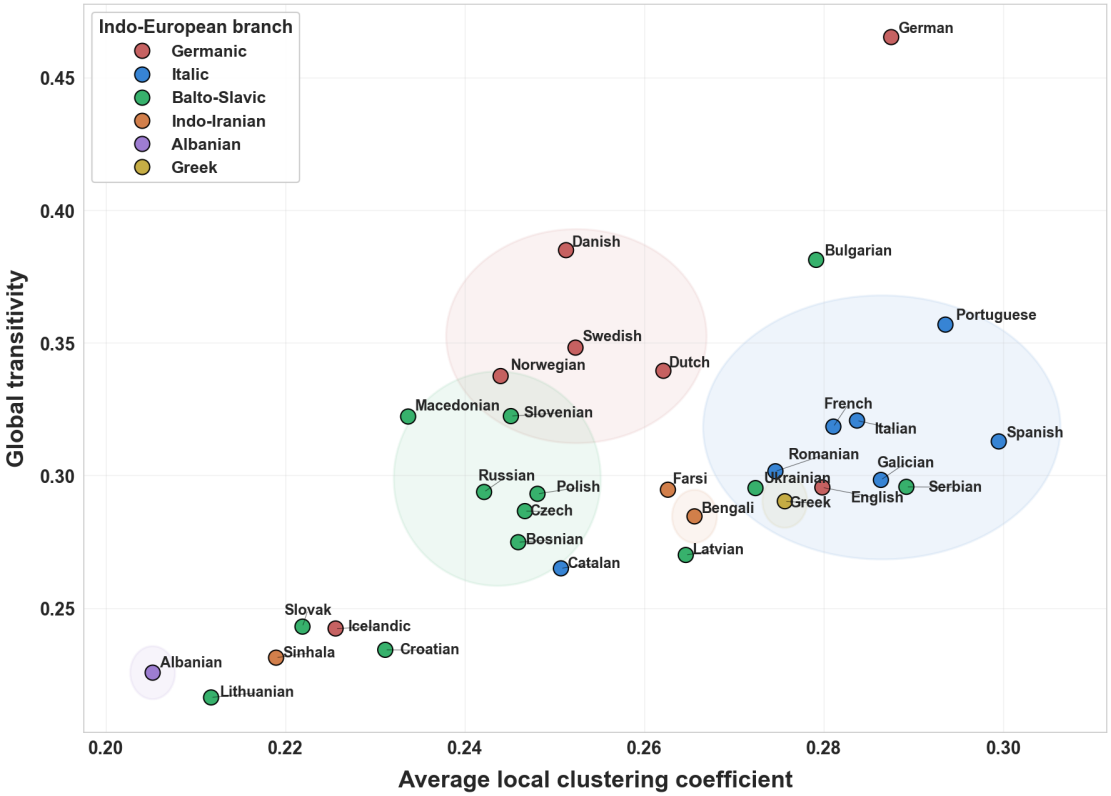

Figure 4.12.: Average local clustering coefficient and global transitivity for the mutual kNN networks.

languages show stronger internal variation. However, these branch-level patterns are often driven by individual languages rather than by entire branches behaving uniformly.

4.4. Community

Community structure is examined through modularity and the number of detected communities with Leiden algorithm. Modularity compares the concentration of edges within detected communities to the number of such edges expected under a comparable null model. Higher modularity indicates that the graph can be divided into more clearly separated groups of nodes, with relatively many connections within communities and fewer connections between them. The number of communities complements this measure by indicating how fine-grained the detected partition is.

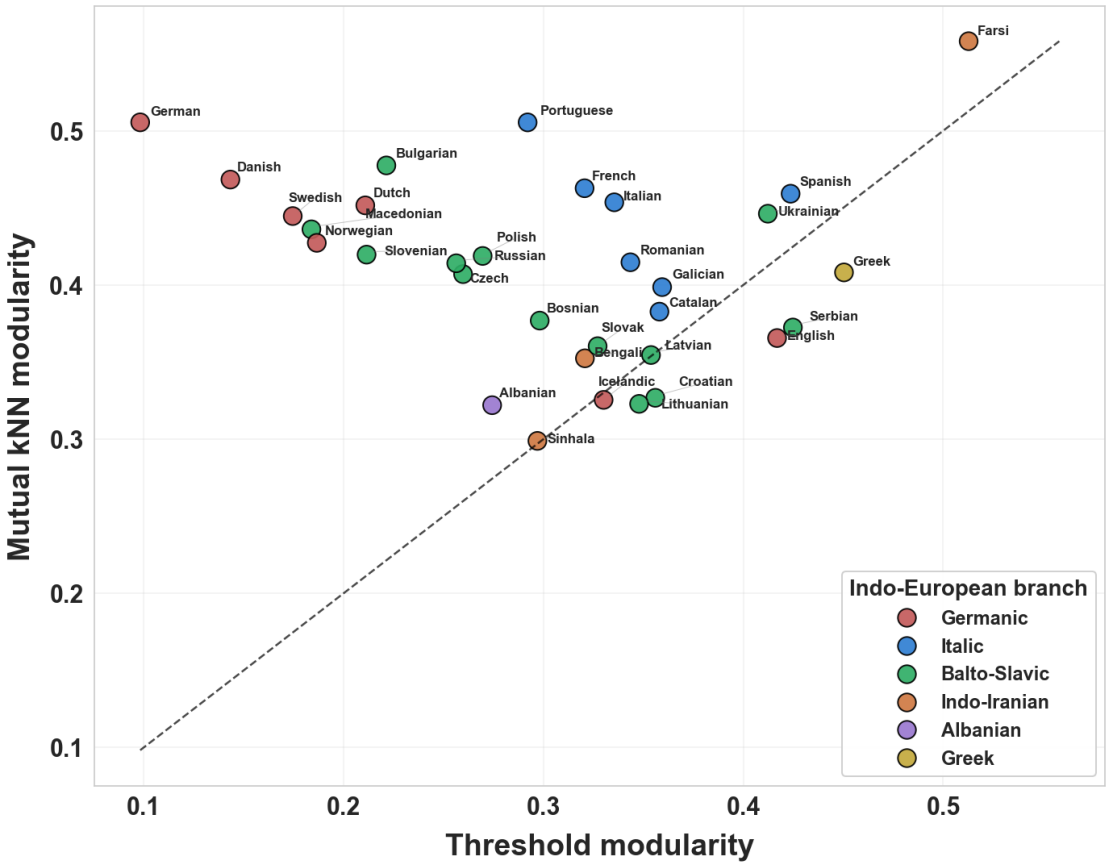

Figure 4.13.: Comparison of modularity values in threshold and mutual kNN networks.

Both modularity and the number of communities were calculated on the whole graph rather than only on the largest connected component. This preserves information about both community organization and graph fragmentation. However, the number of detected communities should therefore not be interpreted as a direct count of semantic categories,

since in fragmented networks some communities may correspond to small disconnected components rather than meaningful subdivisions within the main network.

Figure 4.13 compares modularity values for the threshold and mutual kNN networks. The diagonal line indicated equal modularity in both methods. Languages above this line have higher modularity in the mutual kNN network, while languages below it have higher modularity in the threshold network. Overall, most languages lie above the diagonal: 26 out of 32 languages have higher modularity in the mutual kNN networks than in the threshold networks. . German is the clearest outlier, with modularity increasing from 0.09 in the threshold network to 0.50 in the mutual kNN network. Portuguese also shows a large increase, with modularity increasing from 0.29 in the threshold network to 0.50 in the mutual kNN network, indicating a much stronger community structure under mutual kNN. By contrast, Farsi has high modularity under both construction methods, with 0.51 in the threshold network and 0.56 in the mutual kNN network, making it the strongest example of stable high community structure across construction methods. Languages such as Sinhala, Bengali, Latvian, Croatian, Lithuanian, Albanian, and Greek lie closer to the diagonal, indicating that their modularity values are less strongly affected by the choice of construction method.

Figures 4.14 and 4.15 relate modularity to the number of detected communities for each method separately. In both figures, the dashed lines indicate the median values and divide the plot into four descriptive regions: higher modularity with fewer communities, higher modularity with more communities, lower modularity with fewer communities, and lower modularity with more communities. These regions are used as visual aids rather than formal cluster boundaries. German and Danish are excluded from the median calculation for the threshold network due to their significant number of communities.

In the mutual kNN networks, the number of detected communities is relatively compact, with most languages positioned close to the median of 6 communities. The median modularity is approximately 0.41. Farsi is again the clearest high-modularity case, with modularity of about 0.56 and only 3 detected communities. This suggests a network divided into a small number of clearly separated communities. Portuguese also stands out with high modularity, around 0.51, although it lies closer to the median number of communities. German, Bulgarian, Danish, Spanish, and Swedish are also positioned around or above the modularity median, suggesting relatively strong community structure without extreme fragmentation. By contrast, Lithuanian, Icelandic, Latvian, Serbian, Bosnian, and Sinhala fall into the lower-modularity region, indicating weaker community separation.

In the threshold networks, the relationship between modularity and the number of communities is more uneven. The median modularity is lower, at approximately 0.32, and the number of communities varies more strongly. German and Danish are the clearest outliers in terms of community count. German has 127 detected communities but very low modularity, below 0.1, while Danish has 20 communities and also relatively low modularity. This combination suggests fragmentation rather than strong semantic community structure. In other words, the high number of communities in these cases should not be interpreted as many meaningful semantic categories, but rather as an

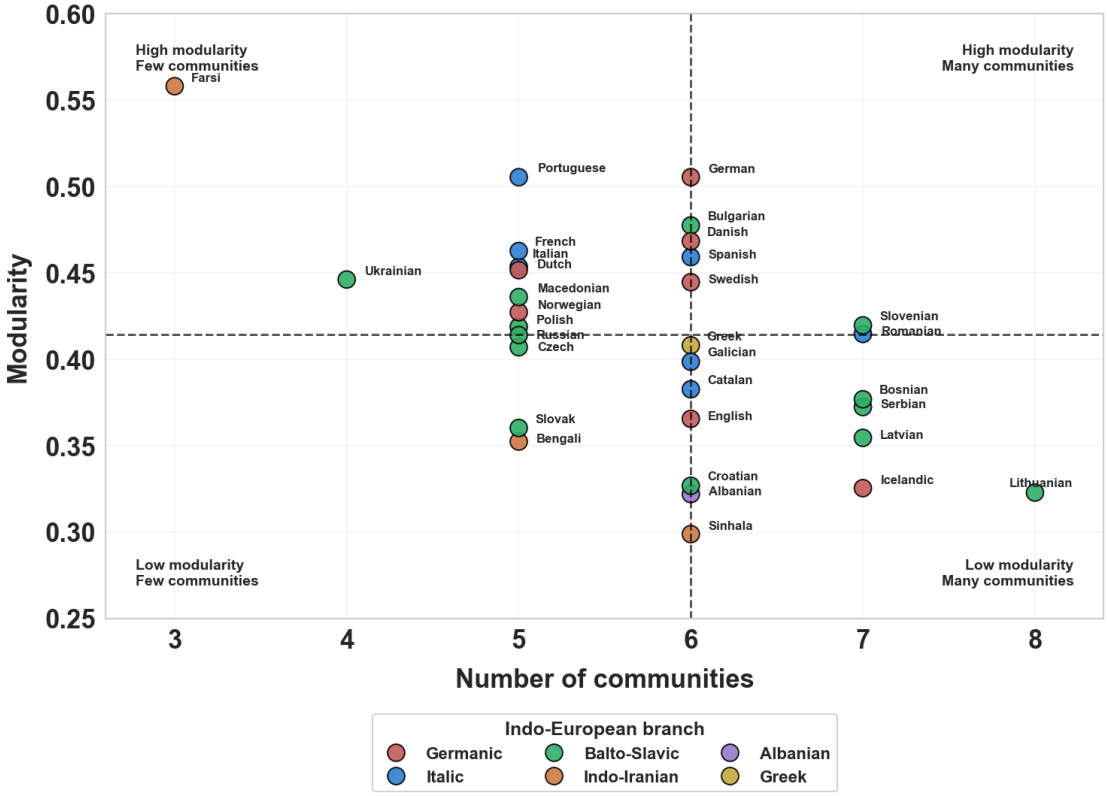

Figure 4.14.: Comparison of modularity and number of detected communities across languages in mutual kNN networks. The dashed diagonal lines indicate the median values.

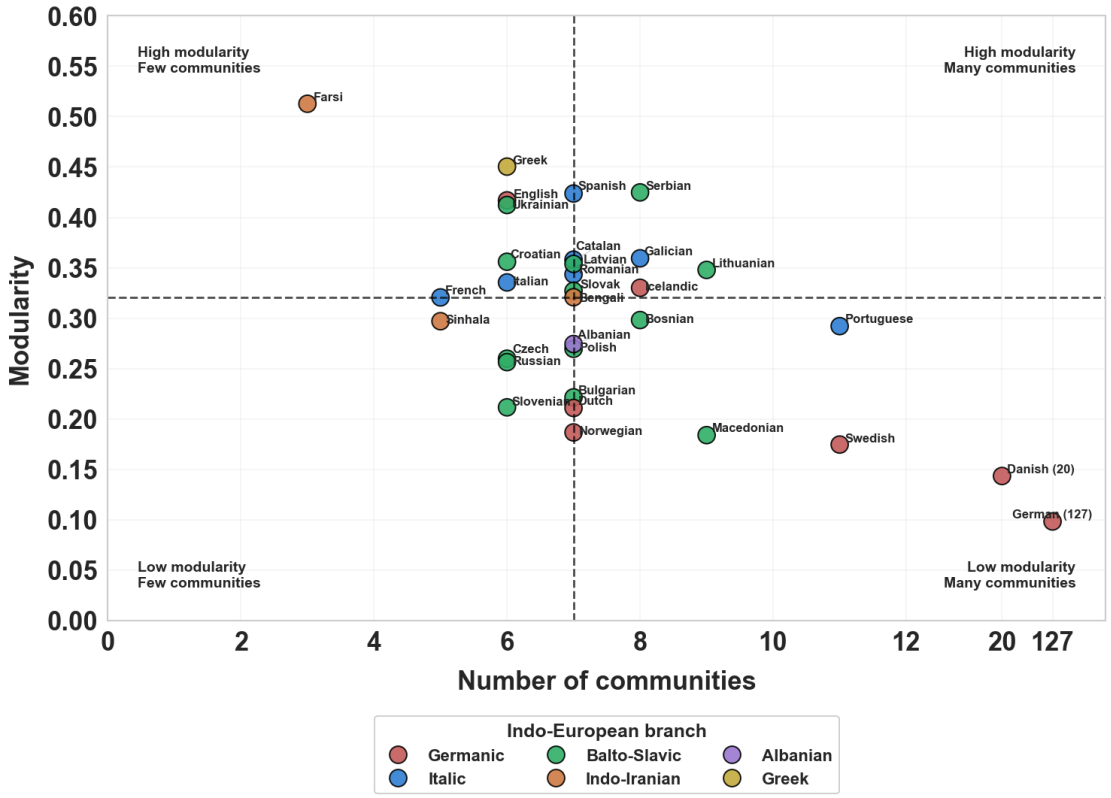

Figure 4.15.: Comparison of modularity and number of detected communities across languages in threshold networks. German and Danish are excluded from the median calculations.

effect of the threshold construction producing many weakly integrated or disconnected parts of the graph[1]. Farsi again stands out as a stable high-modularity case, with threshold modularity of about 0.51 and only 3 detected communities. English, Greek, and Ukrainian also appear in the higher-modularity region with few communities, while Czech, Russian, Slovenian, and Sinhala occupy lower-modularity positions with fewer communities. Portuguese, Swedish, and Macedonian have more communities but lower modularity, indicating a weaker version of the German/Danish pattern.

At the branch level, the figures show tendencies rather than clearly separated Indo-European subgroupings. Several Germanic languages show stronger modularity under mutual kNN than under thresholding, but the branch-level pattern is strongly shaped by German and Danish because of their unusually high community counts in the threshold networks. Italic languages generally occupy intermediate-to-high modularity positions, although Portuguese differs from the other Italic languages, showing especially high mutual kNN modularity. Balto-Slavic languages are distributed across the middle and lower regions, especially in the threshold plot, suggesting internal variation rather than a single branch-level profile. Indo-Iranian languages are also internally mixed, with Farsi clearly separated from Bengali and Sinhala as the only consistently high-modularity case across both methods.

Overall, the community results show that the mutual kNN networks tend to produce stronger and more stable modular structure, while the threshold networks are more sensitive to fragmentation and produce stronger outliers. The comparison also shows why modularity and number of communities must be interpreted together: high modularity indicates stronger community separation, but a high number of communities can also result from graph fragmentation, as seen most clearly in the German and Danish threshold networks.

4.5. Degree Assortativity

Degree assortativity is used to examine whether nodes with similar degrees tend to be connected to each other. Positive degree assortativity means that highly connected nodes tend to connect to other highly connected nodes, while low-degree nodes connect to other low-degree nodes. Negative values indicate the opposite pattern, where high-degree nodes tend to connect to low-degree nodes. Values close to zero suggest little systematic degree-based mixing.

Figure 4.16 compares degree assortativity for the threshold and mutual kNN networks, with languages grouped by Indo-European branch. The clearest overall pattern is that assortativity is much higher in the mutual kNN networks than in the threshold networks.

[1]Since German shows a strong outlier, modularity and the number of communities were also calculated in the largest connected component. The modularity value remained the same, but the number of communities dropped to 15. This suggests that the very high whole-graph community count is largely due to fragmentation outside the largest component. To inspect this pattern, a separate network visualization was created for German comparing the whole graph with the largest connected component, which supports this interpretation.

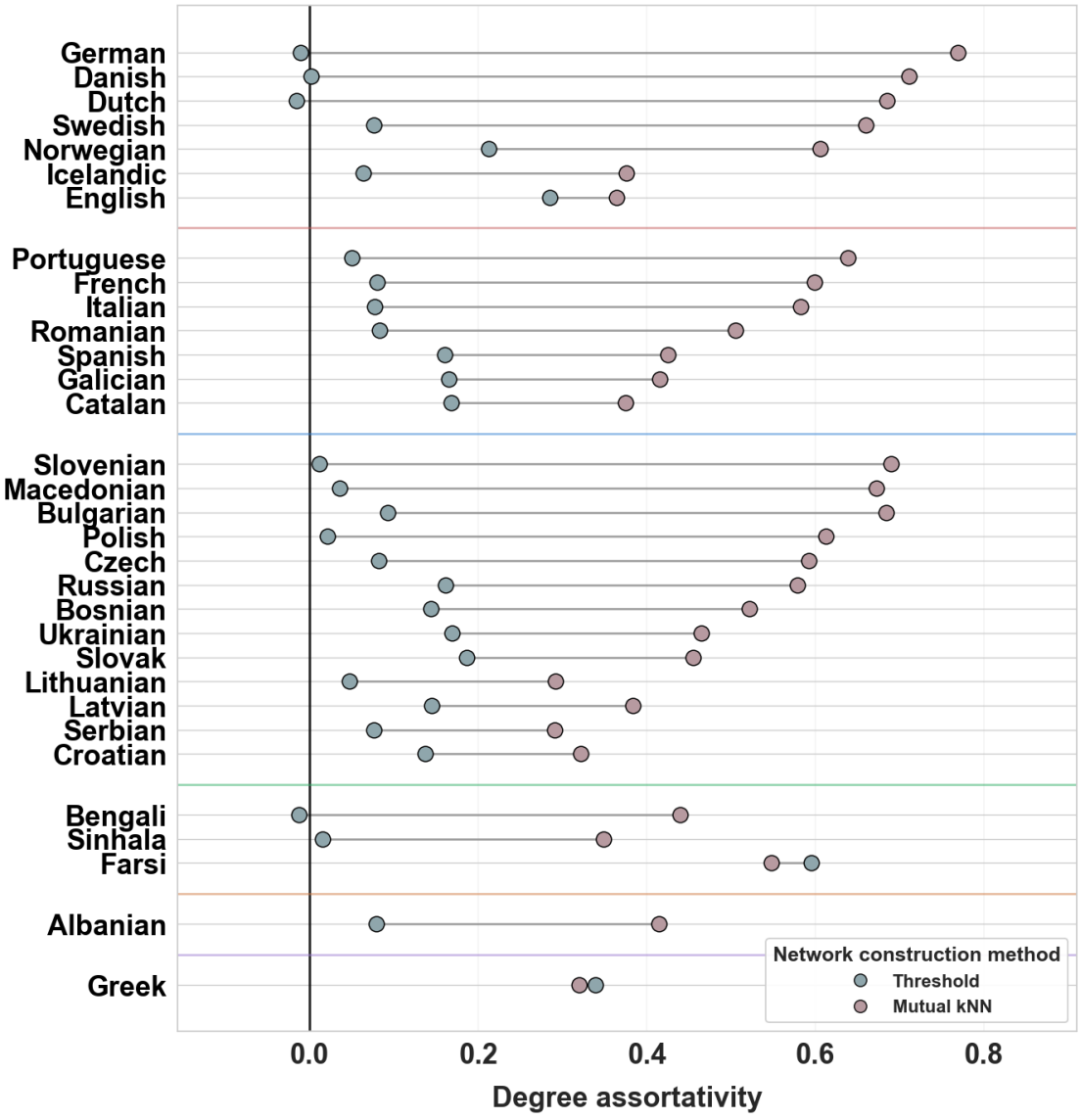

Figure 4.16.: Degree assortativity values for threshold and mutual kNN networks across languages, grouped by Indo-European branch.

In the threshold networks, values range from approximately -0.02 to 0.60, with a median of about 0.08. Most languages therefore show weak or near-zero assortative mixing under the threshold construction. Dutch, Bengali, German, Danish, Slovenian, Sinhala, Polish, and Macedonian all have values close to zero or slightly below zero. This suggests that, in these threshold networks, highly connected words are not especially likely to connect to other highly connected words. Farsi is the main exception, with a much higher threshold assortativity value of approximately 0.60. Greek and English also show comparatively higher threshold values, at approximately 0.34 and 0.28, respectively.

In the mutual kNN networks, assortativity is consistently higher, ranging from approximately 0.29 to 0.77. German has the highest value, at approximately 0.77, followed by Danish at 0.71, Slovenian at 0.69, Dutch at 0.69, Bulgarian at 0.68, Macedonian at 0.67, and Swedish at 0.66. These values indicate a much stronger tendency for nodes with similar degrees to connect to one another under the mutual kNN construction. The lowest mutual kNN values are found for Serbian and Lithuanian, both around 0.29, followed by Greek at 0.32 and Croatian at 0.32. Even these lower values are still clearly positive, showing that mutual kNN generally produces more degree-assortative networks than the threshold method.

The largest method effects occur in languages that move from near-zero threshold assortativity to high mutual kNN assortativity. German increases from approximately -0.01 to 0.77, Danish from 0 to 0.71, and Dutch from -0.02 to 0.69. Similar increases are visible for Slovenian, Macedonian, Bulgarian, Polish, Portuguese, Swedish, French, Czech, and Italian. Farsi and Greek are the only cases where assortativity is slightly lower in mutual kNN than in the threshold networks: Farsi decreases from approximately 0.60 to 0.55, while Greek decreases from approximately 0.34 to 0.32.

At the branch level, the Germanic languages show the strongest method effect. German, Danish, Dutch, Swedish, and Norwegian all shift from low or moderate threshold assortativity to high mutual kNN assortativity. The Italic languages also show consistently positive mutual kNN assortativity, with Portuguese reaching approximately 0.64. Their threshold values are mostly low to moderate, ranging from about 0.05 to 0.17. Balto-Slavic languages are more internally varied: Serbian, Lithuanian, and Croatian have the lowest mutual kNN assortativity values, around 0.29-0.32, while Slovenian, Bulgarian, Macedonian, Polish, Czech, and Russian are much higher, mostly between about 0.59 and 0.69. Indo-Iranian languages are also mixed, with Farsi showing high assortativity in both methods, while Bengali and Sinhala move from near-zero threshold values to moderate mutual kNN values.

Overall, degree assortativity is shaped more strongly by the network construction method than by Indo-European branch membership. The threshold networks mostly show weak degree-based mixing, whereas the mutual kNN networks show consistently positive assortativity. However, some branch-level tendencies remain visible. Germanic languages show the strongest increase under mutual kNN, Italic languages occupy a relatively consistent middle-to-high mutual kNN range, and Balto-Slavic languages show substantial internal variation.

4.6. Hierarchical Clustering

Correlation Analysis

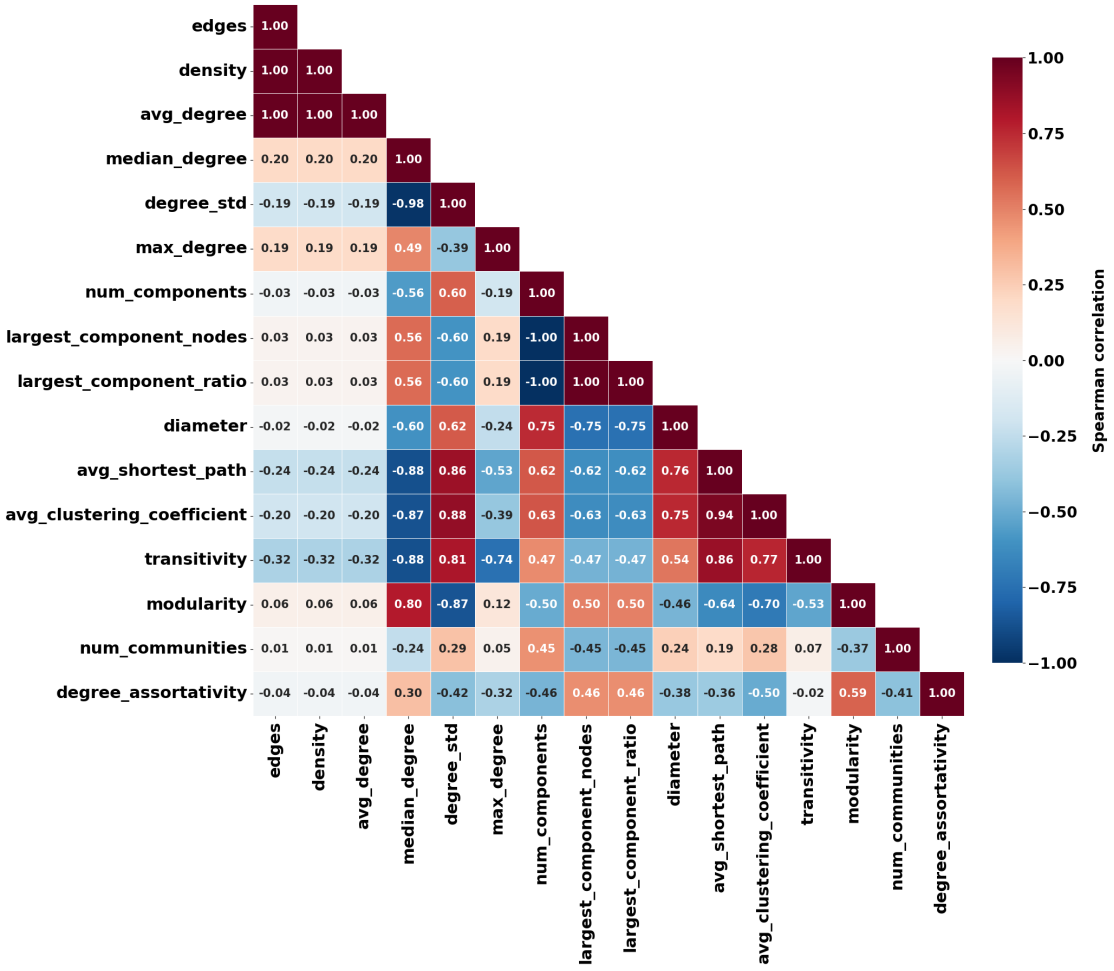

Figure 4.17.: Correlation analysis between semantic network properties for threshold networks.

Before carrying out hierarchical clustering, the graph theoretic measures are examined for redundancy using Spearman's rank correlation coefficient. This step is necessary because hierarchical clustering is based on distances between languages. If several highly correlated measures are included, the same structural dimension may be overrepresented in the clustering and may therefore have a disproportionate influence on the results. Correlations with an absolute value of $|\rho| \geq 0.90$ were treated as strongly redundant, while measures below this threshold were retained.

Figure 4.17 shows the correlation heatmap for the threshold networks. Within the basic network measures, edges, density, and average degree were perfectly correlated with

each other ($\rho = 1.00$). Since these measures all describe the overall edge density of the graphs and the number of nodes is fixed across languages, density was retained as the representative measure from this group. In the degree-related group, median degree and degree standard deviation showed a very strong negative correlation ($\rho = -0.98$). Since degree standard deviation captures heterogeneity in the degree distribution, it was retained. In the connectivity group, the largest component nodes and the largest component ratio were perfectly positively correlated ($\rho = 1.00$), while the number of components was perfectly negatively correlated with both largest component measures ($\rho = -1.00$). Therefore, the largest component ratio was retained because it expresses the relative size of the largest component and is easier to compare across networks. One additional strong correlation occurred between average shortest path length and average local clustering coefficient ($\rho = 0.94$). From this pair, the average local clustering coefficient was kept. Based on this procedure, the reduced set of threshold-network properties retained for hierarchical clustering consisted of density, degree standard deviation, maximum degree, largest component ratio, diameter, average local clustering coefficient, transitivity, modularity, number of communities, and degree assortativity. This selection preserves one or more representatives from the main measure groups considered in the thesis.

The same correlation-based selection procedure was applied to mutual kNN networks. As shown in Figure 4.18, basic network properties, which are edges, density, average degree, and median degree, were perfectly correlated with each other ($\rho = 1.00$). Therefore, density was retained as the representative measure for this group, as in the threshold networks. In addition, degree assortativity was very strongly negatively correlated with the same edge-related measures ($\rho = -0.94$). However, degree assortativity was retained because it captures degree mixing rather than edge density itself. The heatmap also shows that the maximum degree, number of components, largest-component nodes, and largest-component ratio are displayed as NaN in the mutual kNN. These values do not indicate missing data, they indicate that these properties are constant across all languages in the mutual kNN networks, which makes Spearman correlations undefined. Since constant measures cannot be distributed to distance calculations in hierarchical clustering, none of these measures was retained for the mutual kNN set. The reduced set of mutual kNN properties retained for hierarchical clustering consisted of density, degree standard deviation, diameter, average shortest path length, average local clustering coefficient, transitivity, modularity, number of communities, and degree assortativity.

Hierarchical Clustering after Measure Selection

After the correlation analysis, hierarchical clustering was carried out using the reduced set of graph-theoretic measures. The aim of this analysis was to examine whether languages with similar semantic network profiles cluster in ways that resemble the Indo-European family tree. In the figures, the leaf labels are colored according to Indo-European branch membership, which makes it possible to compare the clustering output with the expected genealogical groupings.

Figure 4.19 the hierarchical clustering result for the threshold networks using the reduced set of graph-theoretic measures. Overall, the dendrogram shows only partial cor-

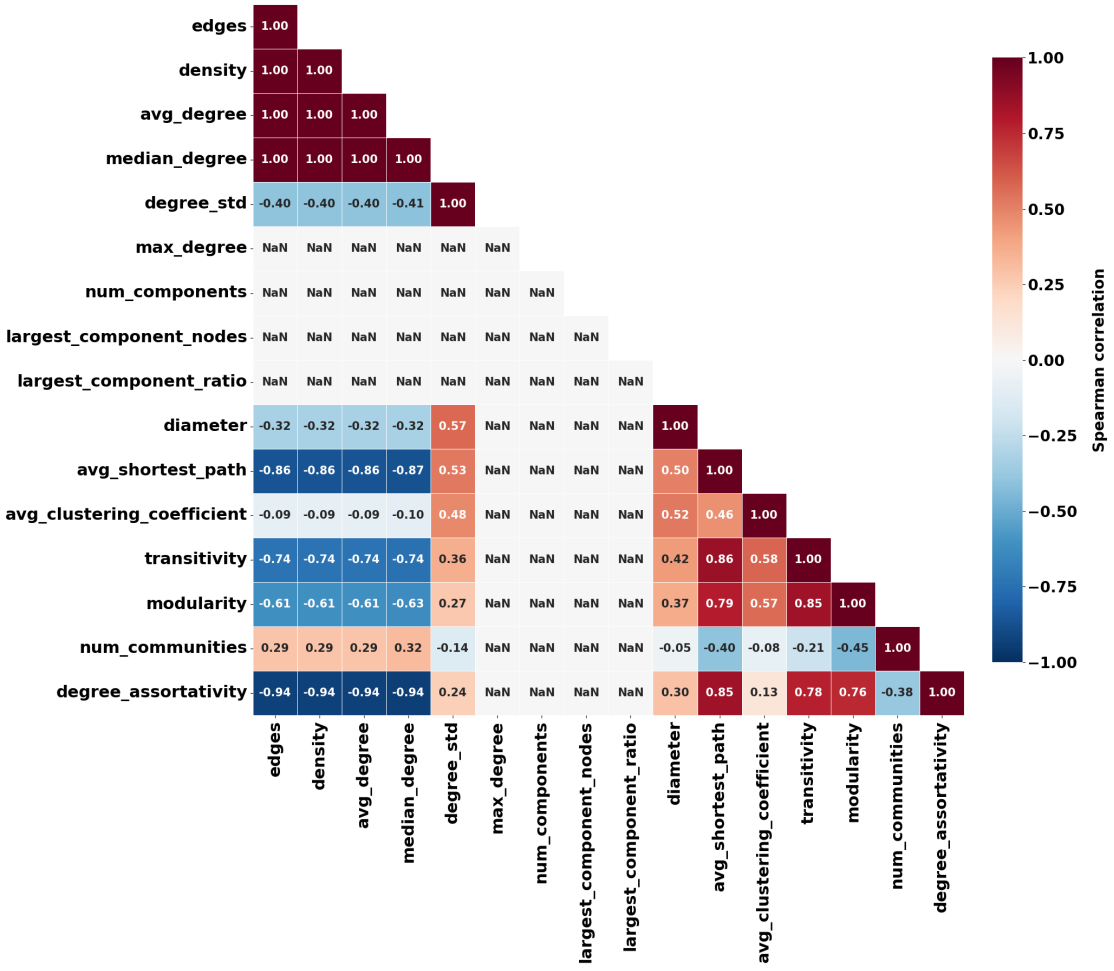

Figure 4.18.: Correlation analysis between semantic network properties for mutual kNN networks.

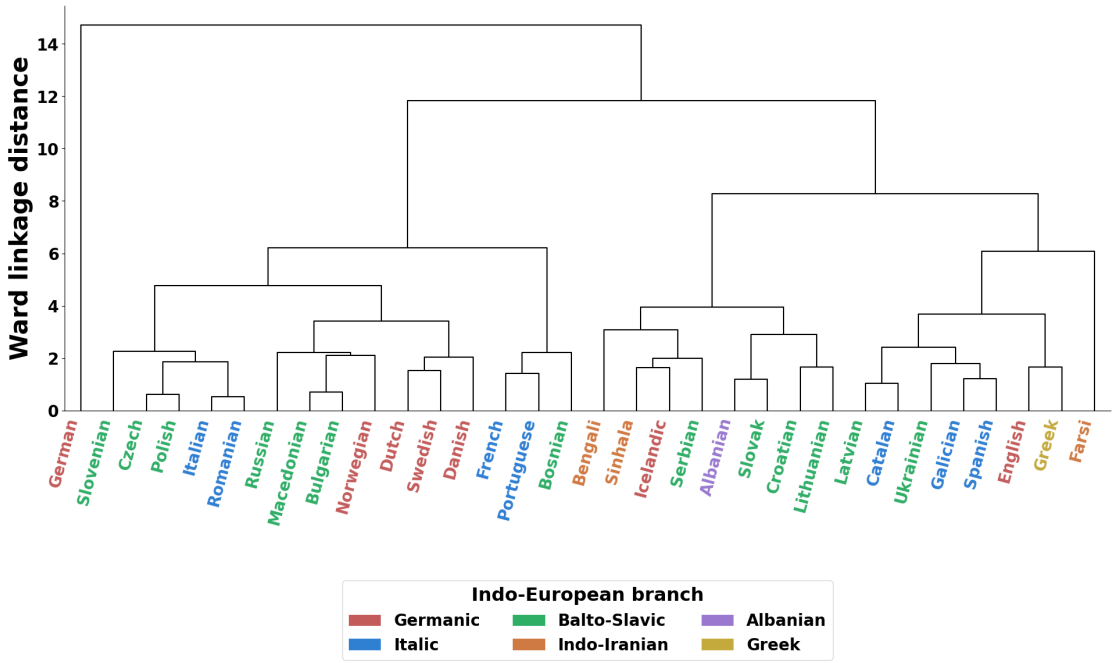

Figure 4.19.: Hierarchical clustering of threshold network properties using the reduced measure set. Measures were standardized before clustering, and leaf labels are colored according to Indo-European branch membership.

respondence with the Indo-European reference structure. Some closely related languages form local pairings at low linkage distances, but these pairings usually do not develop into clean higher-level branch clusters.

Czech and Polish form a close pair, which is compatible with their shared West Slavic affiliation. Italian and Romanian also form a close pair, which reflects their shared placement within the Italic branch. Macedonian and Bulgarian cluster together, matching their close relationship within South Slavic. Swedish and Danish form another close pair, which is consistent with their shared North Germanic affiliation. French and Portuguese also appear as a Western Romance pair, as well as Galician and Spanish. These local groupings suggest that the threshold-network measures capture some small-scale genealogical signal.

However, these local pairings usually do not develop into clean higher-level branch clusters. Czech and Polish are joined by Slovenian, which keeps the cluster broadly Balto-Slavic, but not specifically West Slavic. The Italian-Romanian pair is later absorbed into a larger mixed cluster rather than forming part of a complete Italic branch. Macedonian and Bulgarian are joined by Norwegian, which makes the larger cluster genealogically mixed. Swedish and Danish are joined by Dutch, creating a partially Germanic region, but one that mixes North and West Germanic languages rather than following the expected sub-branch structure. French and Portuguese are later joined by Bosnian, and the Galician-Spanish pair is later joined by Ukrainian. These additions show that although the dendrogram contains meaningful local correspondences, the broader clusters often combine languages from different Indo-European branches.

One notable feature is that German appears as a clear outlier. It is separated from all other languages at the highest level of the tree, suggesting that its threshold network profile differs substantially from the other languages according to the reduced graph measures. This should not be interpreted genealogically, but rather as a property of the graph-theoretic profile produced by the threshold construction.

Other parts of the dendrogram are also difficult to interpret genealogically. Sinhala is grouped with Icelandic, even though Sinhala belongs to Indo-Aryan and Icelandic to North Germanic. Bengali appears nearby at a higher level, but the resulting grouping still does not correspond to the Indo-Iranian branch. Croatian, Lithuanian, Slovak, Catalan, and Latvian also occur in a broader mixed region of the tree. Although several of these languages are Balto-Slavic, the cluster does not cleanly separate Baltic, Slavic, or Italic languages.

Figure 4.20 shows the hierarchical clustering result for the mutual kNN networks using the reduced set of graph-theoretic measures. As in the threshold dendrogram, some local groupings correspond partly to the Indo-European reference structure. Dutch and Danish form a close pair, and German joins them at a slightly higher level. This creates a small Germanic-related region of the tree. French and Italian form another close pair, and Portuguese joins them nearby, producing a partial Italic grouping. Czech and Polish also appear close together, with Russian joining this group at a higher level. This creates a local Slavic-related region of the dendrogram. These examples suggest that the mutual kNN network measures also capture some local similarities between related languages.

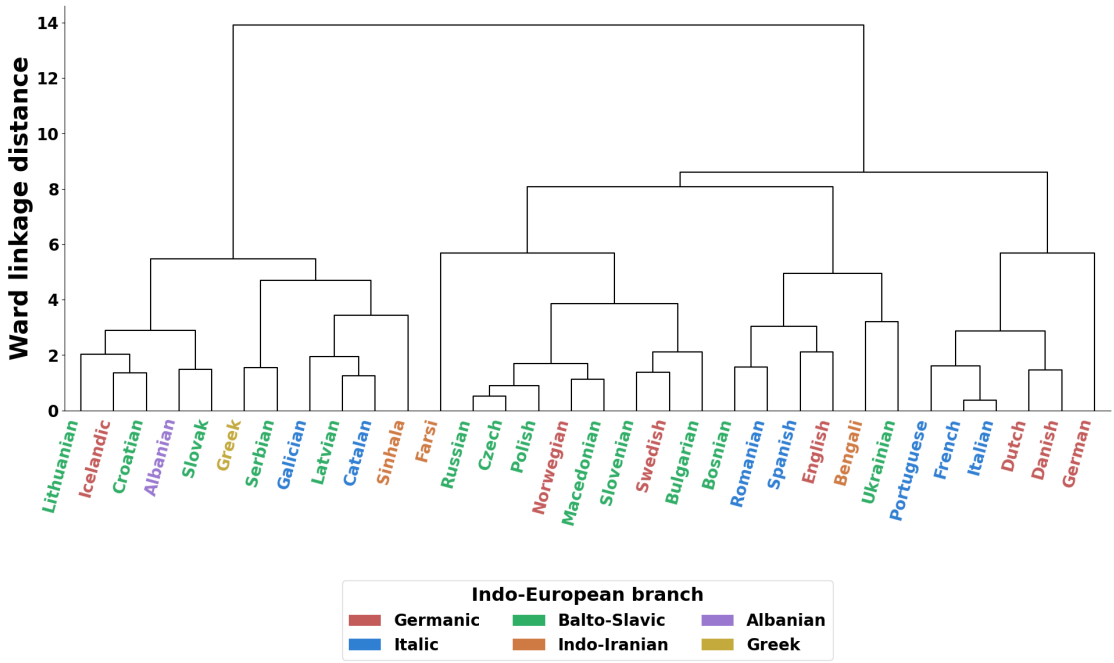

Figure 4.20.: Hierarchical clustering of mutual kNN network properties using the reduced measure set. Measures were standardized before clustering, and leaf labels are colored according to Indo-European branch membership.

However, as with the threshold networks, these groupings do not extend into clear higher-level branches. The Dutch-Danish-German grouping is only partially consistent with the Germanic reference structure, because Danish is North Germanic, while Dutch and German are West Germanic. Other Germanic languages, such as English, Swedish, Icelandic, and Norwegian, are located elsewhere in the dendrogram. Similarly, the French-Italian-Portuguese group captures part of the Italic branch, but the other Italic languages, including Spanish, Galician, Catalan, and Romanian, do not join the same cluster. The Czech-Polish-Russian grouping is Slavic-related, but it does not develop into a coherent Balto-Slavic branch, since other Slavic and Baltic languages are distributed across different parts of the tree.

Several broader regions of the mutual kNN dendrogram are genealogically mixed. Lithuanian, Icelandic, Croatian, Albanian, Slovak, and Greek are placed within the same broad cluster, although they belong to different Indo-European branches. Nearby, Serbian, Galician, Latvian, Catalan, and Sinhala form another mixed group. Bengali appears in a broader region with English and Ukrainian, which again does not correspond to the reference structure. These mixed groupings indicate that the mutual kNN graph measures capture structural similarities in the semantic networks that are not directly aligned with genealogical classification.

Overall, the threshold and mutual kNN dendrograms suggest that the reduced semantic network measures contain some genealogical signal, but this signal is limited. In the threshold networks, several close pairs correspond to expected genealogical relationships, such as Czech-Polish, Macedonian-Bulgarian, Swedish-Danish, French-Portuguese, Galician-Spanish, and Italian-Romanian. In the mutual kNN networks, the strongest partial correspondences appear in small Germanic, Romance, and Slavic-related regions, such as Dutch-Danish-German, French-Italian-Portuguese, and Czech-Polish-Russian. However, in both methods, these local groupings are often absorbed into larger mixed clusters that include languages from different Indo-European branches. This means that the reduced graph-theoretic measures do not reproduce the Indo-European family tree in a systematic way. Instead, they appear to capture some small-scale similarities between related languages, while also reflecting semantic network structures that cut across genealogical boundaries.

5. Discussion

> "I don't understand what you mean, but it sounds good!"
>
> My dad, during a phone conversation about this thesis

The central aim of the thesis was to investigate whether semantic network properties can reveal meaningful similarities and differences between Indo-European languages. More specifically, the analysis asked whether languages from the same subgroup show similar network properties, and whether these properties can recover groupings that resemble the established Indo-European family tree. The discussion therefore focuses on how the observed network patterns should be interpreted. It first considers the language and branch-level patterns, and then discusses whether these patterns provide evidence for the established Indo-European family tree.

To address these questions, semantic networks derived from word embeddings were used to compare the structural organization of meaning across 32 Indo-European languages. The languages were grouped into six branch categories: Germanic, Italic, Balto-Slavic, Indo-Iranian, Greek, and Albanian. For each language, semantic similarity relations between words were transformed into networks using two construction methods: a percentile threshold method and a mutual k-nearest-neighbor method. Graph-theoretic properties were then calculated and compared across individual languages and branches, and were also used in hierarchical clustering to examine whether semantic network profiles resemble established Indo-European subgroupings.

5.1. Language and Branch Level Patterns

Germanic

The Germanic languages show the clearest internal variation among the branches examined in this thesis. Across several measures, they are more spread out, especially in the threshold networks. German, Danish, and Dutch often appear structurally different from other Germanic languages such as English and Icelandic. In the basic network properties, these languages show lower median degree or lower mutual kNN connectivity. This suggests that their semantic similarity relations are either less evenly distributed across the network or less often reciprocal. The connectivity results show a similar pattern. German, Danish, Dutch, Swedish, and Norwegian have longer path structures under the threshold construction, with German being the strongest outlier.

This does not mean that the Germanic languages lack shared structural tendencies. Rather, the branch is not uniform. Some Germanic languages show similar patterns, but

these similarities are interrupted by strong language-specific effects, especially in German and Danish. In this sense, the Germanic results show internal variation within a branch more clearly than they show a single coherent Germanic profile.

German is the most important outlier in the analysis, but its outlier behavior differs between the two construction methods. In the threshold network, German repeatedly shows signs of unevenness and fragmentation. It has a very low median degree, high degree variation, more connected components, longer average shortest path length, larger diameter, and an unusually high number of detected communities, combined with low modularity. These patterns suggest that the German threshold network is less evenly integrated than most other language networks. Many words are only weakly connected to the rest of the network, while some regions appear more separated from the main structure. The high number of communities should therefore not be interpreted as evidence for many meaningful semantic categories. Since modularity is low and the component structure is fragmented, some of these communities are more likely to reflect disconnected or weakly connected parts of the graph.

German also behaves distinctively in the mutual kNN networks, but in a different way. It has the lowest average and median degree in the mutual kNN networks, which means that German words retain fewer reciprocal nearest-neighbor connections on average. However, the mutual kNN version of German is fully connected and shows high modularity and a very high degree of assortativity. This means that German is still relatively sparse in the mutual kNN network, but the connections that remain are more clearly organized into local groups. In simpler terms, German appears fragmented under global thresholding, but locally structured under mutual kNN. This contrast suggests that German's semantic network profile is especially sensitive to how semantic similarity is transformed into edges.

There are several possible explanations for German's outlier behaviour, although they must remain speculative. One possible explanation is German's lexical and morphological structure. German is known for highly productive noun compounding, which is often regarded as one of the most productive word-formation processes in the language (Schlücker and Hüning, 2009). Recent work on German compound semantics also notes that compound interpretation can be ambiguous at the level of the constituents (Miletić et al., 2025). Such compounds may create semantically specific lexical items that are strongly connected to a limited set of related terms. In network terms, this could create dense local neighborhoods while leaving other parts of the graph less integrated. This would fit the observed pattern in German, where the threshold network appears more uneven, while the mutual kNN network reveals a more locally organized structure.

Another possible explanation is German's richer inflectional and morphosyntactic system. German articles vary according to gender and case, and definite article forms include distinctions such as *der, die, das, den, dem,* and *des*. These grammatical distinctions may influence the distributional contexts in which words appear, especially if high-frequency function words, article forms, or inflected variants are included in the vocabulary. Compared with English, where the article system is reduced to *a*, *an* and *the*, German may show more visible grammatical variation. If grammatical markers, compounds, or inflected forms create repeated local patterns, they may strengthen connections within

certain lexical neighborhoods. Therefore, German may contain strongly connected local groups of words, but these groups may not connect evenly to one another.

At the same time, this explanation should remain cautious. The present analysis does not directly test morphology, article frequency, or compounding. Moreover, other Germanic languages also share some morphological or grammatical similarities with German, but they do not show the same pattern to the same extent. Therefore, German's outlier position cannot be attributed to these factors with certainty.

Danish shows a similar but less extreme profile. It also appears as a relatively fragmented or uneven threshold network, especially in the community and connectivity results. However, Danish does not behave as consistently as German across all measures. It is better interpreted as part of a broader Germanic pattern of internal variability rather than as an equally strong outlier. Dutch also frequently appears near German and Danish in degree-related measures and mutual kNN connectivity, suggesting that some Germanic languages share similar network profiles. Nevertheless, the Germanic branch as a whole cannot be characterized by this pattern, since English and Icelandic often behave differently.

Italic

The Italic languages show one of the clearer branch-level patterns in the analysis. Their network properties are generally closer to one another than would be expected if the branch were entirely internally scattered. This is especially visible in the basic network properties and connectivity measures. In the threshold networks, the Italic languages show broadly similar edge distribution, while in the mutual kNN networks, they remain relatively close in average and median degree. This suggests that reciprocal semantic-neighborhood relations are organized in a fairly comparable way across the branch.

This similarity is especially visible among the Western Romance languages. Catalan, Spanish, and Galician show similar profiles in several mutual kNN properties. They have relatively high reciprocal connectivity and compact path structures, meaning that many words keep mutual nearest-neighbor relations and remain closely connected within the network. Their semantic networks are relatively well integrated locally and do not show strong fragmentation. French, Italian, and Portuguese show a related but slightly different pattern, with similarities appearing more clearly in clustering, modularity, and degree assortativity. This suggests that their networks are not only connected but also organized into cohesive local neighborhoods and communities.

The branch is still not completely homogeneous. Portuguese is the most distinctive case because it shows especially high modularity and degree assortativity in the mutual kNN network. This suggests that Portuguese has more clearly separated local communities and a more regular degree-based structure when only reciprocal nearest-neighbor relations are considered. However, Portuguese does not disrupt the overall Italic profile in the same way that German disrupts the Germanic branch. It appears distinctive within the branch, but not isolated from it.

Overall, the Italic branch provides one of the stronger examples of subgroup level similarity in the analysis. The languages are relatively close in basic properties, connectivity,

and clustering, and the Western Romance languages show especially visible similarity in local clustering.

Balto-Slavic

The Balto-Slavic languages show a more internally varied pattern. They do not form one compact branch-level profile across the network properties. Instead, the results suggest partial structure at lower subgroup levels. This means that some closely related languages show similar semantic network properties, but these similarities do not extend to the entire Balto-Slavic branch.

Some of the clearest similarities appear within the Slavic languages. Czech and Polish show comparable profiles in several degree-related and mutual kNN properties, suggesting a similar organization of reciprocal semantic neighborhoods. Russian and Ukrainian also show similarities in connectivity and community structure, although they differ in some local organization measures such as assortativity and modularity. Bulgarian and Macedonian are another important pair, especially because both show relatively strong local organization in clustering, modularity, and mutual kNN assortativity. These patterns suggest that semantic network properties may capture similarities between close Slavic languages or smaller Slavic subgroups.

At the same time, the branch is not uniform. The Baltic languages, especially Lithuanian and Latvian, do not consistently follow the same structural tendencies as the Slavic languages. South Slavic is also internally mixed. Bulgarian and Macedonian appear close in several local-structure measures, but Serbian and Croatian show a different profile, with denser reciprocal connectivity but weaker local organization. This suggests that Balto-Slavic contains several different semantic network profiles rather than one shared branch-wide structure.

One possible explanation is that Balto-Slavic is too broad as a category for the graph-theoretic properties used here. The measures summarize general structural tendencies of the networks, such as degree distribution, connectivity, clustering, and community organization. They may therefore be sensitive enough to detect similarities between close language pairs, such as Czech–Polish, Russian–Ukrainian, or Bulgarian–Macedonian, but not specific enough to represent all Baltic and Slavic languages as one coherent group. This may be because the Balto-Slavic branch contains substantial internal linguistic diversity. Baltic and Slavic languages are historically related, but they are not expected to behave identically in modern distributional semantic spaces. Even within Slavic, the West, East, and South Slavic languages have developed under different conditions of language contact, standardization, and lexical change. These differences may influence vocabulary, semantic associations, and the way semantic neighborhoods are represented in the embeddings.

Language contact may be especially relevant here. Several Balto-Slavic languages may have been shaped by contact with different neighboring languages and linguistic areas. This may help explain why similarities appear more clearly between closer pairs, such as Czech–Polish or Bulgarian–Macedonian, rather than across the whole branch. In the South Slavic group, for example, Bulgarian and Macedonian are commonly associated

with the Balkan linguistic area, which has been shaped by long-term contact between languages (Friedman, 2022). It is therefore possible that contact-related developments contribute to some of the differences observed within the South Slavic languages.

Corpus and preprocessing differences may also contribute to the variation. The Balto-Slavic languages in the dataset differ in script and morphology. Some languages are written in the Latin script, while others use Cyrillic, and this may affect preprocessing depending on how the embeddings were created. Some languages also have richer inflectional systems or more complex token forms, which may influence how many related forms appear as separate tokens in the embedding vocabulary. If these forms are represented differently across languages, this could affect the distribution of cosine similarities and therefore the resulting network structure.

These explanations are nevertheless speculative. The present analysis does not directly test language contact, lexical borrowing, standardization, morphology, or script effects. Therefore, the observed variation cannot be attributed to any single factor with certainty.

Overall, the Balto-Slavic results should be interpreted as evidence for partial subgroup-level similarity rather than as a uniform Balto-Slavic branch. The network properties seem to capture some meaningful similarities among closer languages, especially within Slavic subgroups, but they do not reduce the entire branch to one shared structural pattern. This is important for the research question because it suggests that semantic network properties can reveal cross-linguistic similarities, but these similarities may appear more clearly at the level of close language pairs or smaller subgroups than at the level of broad Indo-European branches.

Indo-Iranian

The Indo-Iranian languages are difficult to interpret as a single branch-level group because only three languages are included. The group is also internally uneven because Farsi belongs to the Iranian branch, while Bengali and Sinhala belong to the Indo-Aryan branch. Therefore, the Indo-Iranian results should be treated cautiously. They can show language-specific patterns, but they are not enough to define a stable branch-level pattern.

Within this small group, Farsi is the clearest distinctive case. Unlike German, which mainly stands out because of fragmentation and strong method effects, Farsi shows a more stable structural profile across both construction methods. It has high modularity in both the threshold and mutual kNN networks, high degree assortativity in the threshold network, and only a small number of detected communities. This suggests that Farsi's network is organized into a few larger semantic regions that are internally well-connected and clearly separated from one another. Farsi also shows a contrast between moderate average local clustering and very high threshold transitivity. This suggests that dense local neighborhoods are not equally present around every word, but that the connected regions that do form tend to show strong triangle closure at the broader network level.

Bengali and Sinhala do not show the same stable high-modularity profile as Farsi. In this sense, they appear more similar to each other than to Farsi, which may partly reflect the fact that both belong to the Indo-Aryan branch, while Farsi belongs to the Iranian branch. Both Bengali and Sinhala have lower modularity and lower threshold

assortativity than Farsi, suggesting that their networks are less clearly divided into a small number of strongly separated semantic regions. However, they are not identical. Bengali is distinctive because it has a very high maximum degree in the threshold network, suggesting the presence of a strongly connected hub. Sinhala appears more moderate and often lower in clustering, modularity, and assortativity. Therefore, the three Indo-Iranian languages do not form one shared network profile. Instead, Farsi stands apart as the clearest distinctive case, while Bengali and Sinhala show a more moderate Indo-Aryan tendency with their own language-specific differences.

Greek and Albanian

Greek and Albanian cannot be discussed in terms of branch-level patterns because each language is the only representative of its branch in the dataset. They are therefore better interpreted as individual reference cases. Greek appears as a relatively well-integrated network, with high mutual kNN degree values, compact path structure, and stable connectivity across the two construction methods. It also shows moderately high threshold assortativity, suggesting that its semantic network is relatively compact and less strongly affected by the choice of construction method. Albanian also shows a generally compact profile. Its networks are fully connected in both methods, its diameter remains low, and its mutual kNN network has a relatively high average degree, meaning that many reciprocal nearest-neighbor relations are retained. However, Albanian does not show especially high local clustering or modularity, which suggests that these connections do not form unusually strong local clusters or sharply separated communities.

5.2. Genealogical Patterns in Network Properties

The hierarchical clustering results address the second research question of whether semantic network properties can be used to recover subgroupings that resemble the established Indo-European family tree. Overall, the answer is only partial. The dendrograms show some meaningful local groupings, especially between closely related languages. However, these local similarities do not develop into higher-level branch clusters. This suggests that the graph-theoretic properties contain some genealogical signal, but not enough to reconstruct the Indo-European family tree in a systematic way.

In the threshold dendrogram, several close language pairs correspond to expected genealogical relationships. These are mostly paired groupings at low linkage distances, such as Czech–Polish within West Slavic, Macedonian–Bulgarian within South Slavic, Swedish–Danish within North Germanic, Galician–Spanish within Western Romance, and Italian–Romanian within the Italic branch. These pairings suggest that the threshold-network properties are not arbitrary. They capture some structural similarities between closely related languages, especially at the level of close pairs. However, the threshold clustering does not preserve these relationships at higher levels. The close pairs do not develop into complete branch clusters, but are instead absorbed into broader mixed regions of the tree. German is especially important here because it appears as a complete

outlier, separated from the other Germanic languages. This fits the property level results, where German repeatedly showed a distinctive threshold network profile. Its position in the dendrogram should therefore not be interpreted as a genealogical distance, but rather as the result of its unusual graph-theoretic structure under the threshold method.

The mutual kNN dendrogram shows a similar pattern, although the local groupings are slightly broader. Instead of producing pairs, it produces some small three-language regions. Dutch, Danish, and German form a Germanic group, French, Italian, and Portuguese form an Italic group, and Czech, Polish, and Russian form a Slavic group. These groupings again suggest that related languages can have similar semantic network profiles. However, the broader dendrogram remains mixed and does not recover complete Indo-European branches. German is also no longer a complete outlier in this dendrogram, which fits the earlier interpretation that mutual kNN makes its local structure more organized than the threshold method.

The most important interpretation is that the clustering seems more sensitive to close language pairs or small local groups than to deep genealogical branches. This is consistent with the branch-level discussion. Italic languages showed a relatively compact profile, and some of this appears in the clustering. Slavic languages showed lower-level similarities, such as Czech–Polish and Bulgarian–Macedonian, and these also appear in the dendrograms. Germanic languages showed strong internal variation, and the clustering reflects this by failing to group all Germanic languages together. It confirms that semantic network properties capture partial similarities, but these similarities are not organized cleanly enough to reproduce the family tree.

One reason for the partial clustering pattern is that the semantic network approach captures a different level of linguistic organization from traditional phylogenetic classification. The analysis does not directly model morphological distance, sound correspondences, or shared historical innovations. Instead, it compares the structural organization of semantic similarity in modern embedding spaces. This means that the dendrograms should not be expected to reproduce the Indo-European family tree exactly. Rather, they show whether semantic network structure overlaps with genealogical relatedness. The fact that several close language pairs are recovered suggests that such an overlap exists to some extent. However, the mixed higher-level clusters suggest that this overlap is not strong enough to organize the languages into full Indo-European branches. Instead, the network properties seem to capture a broader kind of similarity, combining genealogical relatedness with language-specific semantic organization and possible corpus effects.

The network construction method also strongly affects what kind of similarity becomes visible. The threshold method captures broader global similarity structure, but it is more sensitive to unevenness and fragmentation. This explains why German becomes a strong outlier in the threshold dendrogram. The mutual kNN method captures local reciprocal neighborhoods and produces more stable structures, but it can also emphasize methodological regularities rather than genealogical relationships. As a result, the two methods produce different clustering structures and highlight a different aspect of semantic network similarity.

Overall, semantic network properties can be used for cross-linguistic comparison, and

they can reveal some genealogically meaningful similarities. However, they do not recover the Indo-European family tree as a whole. This answers both research questions cautiously: languages from the same subgroup can show similar semantic network properties, and these properties can recover some meaningful local patterns, but they cannot by themselves reconstruct the established Indo-European tree.

5.3. Limitations and Future Work

Several limitations should be considered when interpreting the results of this thesis. One important limitation concerns the data used to construct the semantic networks. The word embeddings used in this study were derived from subs2Vec, which is based on subtitle data. This made it possible to compare a relatively large number of Indo-European languages, but subtitle data mainly reflects modern, pseudo-conversational language. It is also unclear whether the subtitle corpora for all languages are balanced in terms of time period, genre, films, series, or translation practices. As a result, the vocabulary may be genre-specific or shaped by media-related language use. Translation practices may also influence which expressions occur frequently and how certain semantic relations are represented. Therefore, the semantic relations captured by the embeddings may partly reflect the characteristics of subtitle data to a certain extent. This is especially relevant for hierarchical clustering, because established genealogical classifications are normally based on historical linguistic evidence rather than modern corpus-based distributional patterns. Future work could address this by comparing embeddings trained on different types of corpora, such as subtitles, newspapers, literature, or historical texts. A mixed corpus design could show whether the observed network patterns remain stable across different kinds of language data.

A second limitation concerns the vocabulary used to construct the networks. The networks were built from the most frequent words in each language rather than from a shared multilingual vocabulary or an aligned set of concepts. As a result, the networks do not necessarily represent the same semantic content across languages. While the analysis compares graph-theoretic properties rather than individual words, the structure of a network is still influenced by the vocabulary from which it is constructed. Therefore, some of the observed differences may reflect differences in lexical composition in addition to differences in semantic organization. Future work could address this issue by constructing networks from aligned concept lists, translation dictionaries, or multilingual lexical resources. This would make it possible to examine whether similar structural patterns emerge when the comparison is based on more directly comparable semantic content.

A further challenge concerns the representation of some Indo-European branches in the dataset. Some branches, such as Germanic, Italic, and Balto-Slavic, are represented by several languages, while others are much more limited. Indo-Iranian includes three languages, while Greek and Albanian are each represented by a single modern language. This is not simply a sampling problem, since Greek and Albanian are isolated branches in the modern Indo-European family. However, from a data-driven perspective, it makes the branch-level comparison uneven. Branches represented by several languages can show

internal variation, while branches represented by one language can only provide a single language profile.Future work could address this challenge in several ways. For Indo-Iranian, the most direct solution would be to include more Iranian and Indo-Aryan languages. For Greek and Albanian, where modern branch-internal comparison is more limited, future work could compare dialectal varieties or historical stages if suitable embeddings and corpora are available. Another option would be to use resampling or bootstrapping methods. For example, multiple networks could be constructed from different vocabulary samples or corpus subsets for the same language. This would not create new branch members, but it would help test whether the observed Greek or Albanian network profile is stable or strongly dependent on the selected vocabulary and corpus data.

Future work could also apply the same general approach to questions beyond Indo-European classification. The combination of word embeddings, semantic networks, and graph-theoretic properties could be used to study semantic change over time if comparable corpora are available for different historical periods. For example, separate networks could be constructed from earlier and later stages of the same language in order to examine whether properties such as clustering, modularity, connectivity, or community structure change over time. The method could also be applied to the study of language contact by comparing languages that experienced prolonged contact with closely related languages that developed under different contact conditions. This could show whether semantic organization reflects not only shared ancestry, but also long-term interaction between speech communities.

Another promising direction would be to focus on specific semantic domains rather than the full vocabulary. Networks could be constructed for domains such as time, emotion, motion, color, kinship, or body terms in order to compare how closely languages organize particular areas of meaning. This could be done within Indo-European, but also across other language families, such as Turkic, Uralic, or Semitic languages. Such a comparison could show whether some semantic domains are more stable across related languages, while others show greater variation due to cultural, historical, or contact-related factors. In this way, the method could move beyond genealogical classification and be used more broadly to study how semantic structure is shaped across languages.

Despite these limitations, the findings suggest that the combination of word embeddings, semantic networks, and graph-theoretic analysis offers a useful approach for exploring semantic structure across languages. Future research can build on this framework to investigate a wider range of linguistic questions concerning language relatedness, contact, and semantic change.

6. Conclusion

This central idea in lexical semantics is that meaning is relational and words do not gain meaning in isolation, but through their relations to other words, concepts and semantic fields. Within this concept, meaning is understood through relations such as similarity, contrast, association, and hierarchy. If meaning is structured relationally, then the vocabulary of a language can also be understood as a system of connections rather than as a list of separate lexical items. Distributional semantics provides a computational way of modeling this relational view of meaning. In word embedding models, words that appear in similar contexts are positioned closer to one another in a high-dimensional vector space, making semantic similarity measurable through methods such as cosine similarity. Network analysis adds another layer to this approach by transforming these semantic relations into networks, where words are represented as nodes and similarity relations as edges. Semantic network properties can then be used to describe how semantic neighborhoods are formed, how strongly words are connected, and how broader patterns of semantic organization emerge within a language.

The Indo-European language family provides a useful context for applying this approach because its subgroupings are well established through historical comparative research, while the family itself contains considerable internal diversity. This makes it possible to compare the semantic network profiles of related languages against an existing genealogical reference structure. Building on this theoretical background, the thesis asked whether embedding-derived semantic networks can be used for cross-linguistic comparison within the Indo-European language family. More specifically, it examined whether languages from the same Indo-European subgroup exhibit similar semantic network properties, and whether these properties can recover subgroupings that resemble the established Indo-European family tree.

To answer these questions, the thesis combined distributional semantics with graph-based analysis. The empirical basis of the study was the Subs2Vec word embedding data, from which 32 Indo-European languages were selected. These languages represented six branches or branch categories,which are Germanic, Italic, Balto-Slavic, Indo-Iranian, Greek, and Albanian. For each language, the embeddings were used to calculate semantic similarity between words through cosine similarity. These similarity values were then transformed into two semantic networks using a percentile threshold method and a mutual k-nearest-neighbor method. In the prior, the edges were retained between word pairs whose cosine similarity exceeded the 95% threshold; in the latter, the edges were retained only when two words occurred in each other's nearest neighbor lists. This made it possible to compare a more globally oriented network construction method with a more locally oriented one.

For each resulting network, a set of graph-theoretic properties was calculated. These

included degree-related measures, connectivity measures, clustering and transitivity, community structure, and degree assortativity, to see whether languages that are in the same groupings or subgroupings exhibited any similarities in the group. These properties were used to describe different aspects of semantic network structure, including how evenly words are connected, how compact or fragmented the networks are, how strongly local semantic neighborhoods are formed, and whether larger communities of words emerge. They were then compared across individual languages and Indo-European branches in order to examine whether languages from the same group or subgroup showed similar structural profiles. Finally, the graph-theoretic properties were used for hierarchical clustering to test whether languages with similar semantic network profiles correspond to known Indo-European subgroupings.

The results for the first research question show that languages from the same subgroup can share similar semantic network properties, but that this similarity is not consistent across all branches. The clearest group-level similarity appears in the Italic languages, which show one of the more compact structural profiles across both network construction methods and are less internally varied than some other branches. Balto-Slavic shows a more mixed pattern. It does not form one uniform branch-level profile, but some lower-level similarities are visible, especially among close Slavic pairs such as Czech–Polish and Bulgarian–Macedonian. Germanic languages show the strongest internal variation, with German emerging as the clearest outlier, especially in the threshold network. Indo-Iranian, Greek, and Albanian are harder to interpret at branch level because they are represented by fewer languages or by only one language in the dataset. Thus, subgroup membership is reflected in some semantic network properties, but only partially and most clearly in compact branches or close language pairs.

The second research question can also be answered only partially, since the hierarchical clustering recovers some genealogical patterns but not the Indo-European family tree as a whole. In the threshold network, the clearest groupings are mostly close language pairs that correspond to known lower-level relationships, such as Czech–Polish within West Slavic, Macedonian–Bulgarian within South Slavic, Swedish–Danish within North Germanic, Galician–Spanish within Western Romance, and Italian–Romanian within Italic. In the mutual kNN network, the recovered groupings are slightly broader and often appear as small three-language regions, such as Dutch–Danish–German for Germanic, French–Italian–Portuguese for Italic, and Czech–Polish–Russian for Slavic. However, in both methods, these local or small-group similarities do not develop into interpretable higher-level branch clusters. Thus, semantic network properties can reflect some genealogical similarity, especially among close language pairs and smaller subgroups, but they are not sufficient to reconstruct the established Indo-European family tree.

The main contribution of this thesis is that it shows how semantic network analysis, combined with word embeddings, can offer an additional perspective on language-relatedness. The approach does not replace traditional historical-comparative methods, which remain essential for reconstructing genealogical relationships through phonology, morphology, and shared innovations. Instead, it provides a different way of comparing languages by focusing on the structure of semantic similarity in embedding-derived networks. In

this sense, the thesis demonstrates that network properties can reveal patterns that are relevant for cross-linguistic comparison, even when these patterns do not fully correspond to established family-tree structures.

More broadly, the method is useful for exploring semantic organization across languages when it is interpreted as a structural and corpus-based perspective rather than as direct genealogical evidence. The results suggest that semantic networks can capture how words are connected, how semantic neighborhoods are formed, and how larger patterns of meaning emerge within different languages. This makes the approach valuable not only for questions of language relatedness, but also for future research on semantic change, language contact, and cross-linguistic variation in semantic domains. The thesis therefore contributes to digital humanities and computational linguistics by showing that semantic network structure can be used as a meaningful tool for comparing languages, while also clarifying the limits of what this kind of comparison can show.

Bibliography

Abramov, O. and Mehler, A. (2011). Automatic Language Classification by means of Syntactic Dependency Networks. *Journal of Quantitative Linguistics*, 18(4):291–336.

Almeida, F. and Xexéo, G. (2019). Word Embeddings: A Survey. Version 2 updated 2023.

Baldi, P. (2002). *The Foundations of Latin*. De Gruyter, Berlin.

Barabási, A.-L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512.

Baroni, M., Bernardi, R., and Zamparelli, R. (2014a). Frege in Space: A Program for Compositional Distributional Semantics. *Linguistic Issues in Language Technology*, 9:241–346.

Baroni, M., Dinu, G., and Kruszewski, G. (2014b). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In Toutanova, K. and Wu, H., editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Berlin, B. and Kay, P. (1991). *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley.

BNC Consortium (2007). The British National Corpus, Version 3 (BNC XML Edition). http://www.natcorp.ox.ac.uk/.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*, 6(1):213–234.

Boleda, G. and Herbelot, A. (2016). Formal distributional semantics: Introduction to the special issue. *Computational Linguistics*, 42(4):619–635.

Bond, F., Vossen, P., McCrae, J., and Fellbaum, C. (2016). CILI: The Collaborative Interlingual Index. In *Proceedings of the 8th Global WordNet Conference*, pages 50–57, Bucharest. https://omwn.org/, Accessed: 20 May 2026.

Borge-Holthoefer, J. and Arenas, A. (2010). Semantic Networks: Structure and Dynamics. *Entropy*, 12:1264–1302.

Bouckaert, R. R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., and Atkinson, Q. D. (2012). Mapping the Origins and Expansion of the Indo-European Language Family. *Science*, 337:957–960.

Bowern, C. and Evans, B. (2015). *The Routledge Handbook of Historical Linguistics*. Routledge, 1st edition.

Brenndoerfer, M. (2025). WordNet - A Semantic Network for Language Understanding. `https://mbrenndoerfer.com/writing/history-wordnet-semantic-network`. Accessed: 20 April 2026.

Budel, G., Jin, Y., Van Mieghem, P., and Maksim, K. (2023). Topological Properties and Organizing Principles of Semantic Networks. *Scientific Reports*, 13:11728.

Casacuberta, S., Halevy, K., and Blasi, D. E. (2021). Evaluating Word Embeddings with Categorical Modularity. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1982–1993, Online. Association for Computational Linguistics.

Chang, W., Cathcart, C., Hall, D., and Garrett, A. (2015). Ancestry-constrained Phylogenetic Analysis supports the Indo-European Steppe Hypothesis. *Language*, 91:194–244.

Clackson, J. (2007). *Indo-European Linguistics: An Introduction*. Cambridge University Press, Cambridge.

Clackson, J. (2022). Methodology in Linguistic Subgrouping. In Olander, T., editor, *The Indo-European Language Family: A Phylogenetic Perspective*, pages 18–32. Cambridge University Press, Cambridge.

Clauset, A., Newman, M., and Moore, C. (2004). Finding Community Structure in Very Large Networks. *Physical Review E*, 70(6):066111.

CLICS (2026). CLICS: Database of Cross-Linguistic Colexifications. https://clics.clld.org/.

Cong, J. and Liu, H. (2014). Approaching Human Language with Complex Networks. *Physics of Life Reviews*, 11(4):598–618.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press, Cambridge.

Dalmia, A. and Sia, S. (2021). Clustering with UMAP: Why and How Connectivity Matters.

Dellert, J., Daneyko, T., Münch, A., et al. (2020). NorthEuraLex: A Wide-Coverage Lexical Database of Northern Eurasia. *Language Resources and Evaluation*, 54:273–301.

Dyen, I., Kruskal, J., and Black, P. (1992). An Indoeuropean Classification: A Lexicostatistical Experiment. *Transactions of the American Philosophical Society*, 82:1–132.

Eberhard, D. M., Simons, G. F., and Robinson, A. J. (2026). Ethnologue: Languages of the World. https://www.ethnologue.com/insights/largest-families/ Accessed: 20 April 2026.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Ferrer i Cancho, R. and Solé, R. V. (2001). The Small World of Human Language. *Proceedings of the Royal Society B: Biological Sciences*, 268(1482):2261–2265.

Firth, J. R. (1957). Applications of General Linguistics. *Transactions of the Philological Society*, 56(1):1–14.

Friedman, V. A. (2022). The Balkans. In Mufwene, S. S. and Escobar, A. M., editors, *The Cambridge Handbook of Language Contact: Volume 1: Population Movement and Language Change*, Cambridge Handbooks in Language and Linguistics, pages 189–231. Cambridge University Press, Cambridge.

Fujinuma, Y., Boyd-Graber, J., and Paul, M. J. (2019). A Resource-Free Evaluation Metric for Cross-Lingual Word Embeddings Based on Graph Modularity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4952–4962, Florence, Italy. Association for Computational Linguistics.

Gao, Y., Liang, W., Shi, Y., and Huang, Q. (2014). Comparison of Directed and Weighted Co-occurrence Networks of Six Languages. *Physica A: Statistical Mechanics and its Applications*, 393(C):579–589.

Geeraerts, D. (2006). Prospects and Problems of Prototype Theory. In Geeraerts, D., editor, *Words and Other Wonders - Papers on Lexical and Semantic Topics*. Mouton de Gruyter, Berlin.

Ghannay, S., Favre, B., Estève, Y., and Camelin, N. (2016). Word Embedding Evaluation and Combination. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 300–305, Portorož, Slovenia. European Language Resources Association (ELRA).

Gray, R. D. and Atkinson, Q. (2003). Language Tree Divergence Times Support the Anatolian Theory of Indo-European Origin. *Nature*, 426:435–39.

Greenhill, S. J. and Gray, R. D. (2012). Basic vocabulary and Bayesian phylolinguistics: issues of understanding and representation. *Diachronica*, 29(4):523–537.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2026). Glottolog 5.3. https://glottolog.org.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Jones, W. (1786). *The Third Anniversary Discourse, Delivered 2 February, 1786: "On the Hindús"*. Manuel Cantopher, Calcutta.

Jurafsky, D. and Martin, J. H. (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd edition. `https://web.stanford.edu/~jurafsky/slp3/`.Online manuscript released January 6, 2026.

Kumar, A. A., Steyvers, M., and Balota, D. A. (2022). A Critical Review of Network-Based and Distributional Approaches to Semantic Memory Structure and Processes. *Topics in Cognitive Science*, 14:54–77.

Lehmann, W. P. (1993). *Theoretical Bases of Indo-European Linguistics*. Routledge, London.

Lehrer, A. (1974). *Semantic fields and lexical structure*. American Elsevier, New York.

Lenci, A. and Sahlgren, M. (2023). *Distributional Semantics*. Studies in Natural Language Processing. Cambridge University Press, Cambridge.

Leskien, A. (1876). *Die Deklination im Slavisch-Litauischen und Germanischen*. S. Hirzel, Leipzig.

Lewis, M., Cahill, A., Madnani, N., and Evans, J. (2023). Local Similarity and Global Variability Characterize the Semantic Space of Human Languages. *Proceedings of the National Academy of Sciences*, 120(51).

List, J.-M., Cysouw, M., and Forkel, R. (2016). Concepticon: A Resource for the Linking of Concept Lists. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2393–2400, Portorož, Slovenia. European Language Resources Association.

Liu, H. and Cong, J. (2013). Language Clustering with Word Co-occurrence Networks Based on Parallel Texts. *Chinese Science Bulletin*, 58(10):1139–1144.

Liu, H. and Li, W. (2010). Language Clusters Based on Linguistic Complex Networks. *Chinese Science Bulletin*, 55(30):3458–3465.

Lyons, J. (1977). *Semantics: Volume I*. Cambridge University Press, Cambridge.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Miletić, F., Schmid, A., and Schulte im Walde, S. (2025). Probing BERT for German Compound Semantics. In Gerber, J., Cieliebak, M., Tuggener, D., and Hürlimann, M., editors, *Proceedings of the 10th edition of the Swiss Text Analytics Conference*, pages 81–88, Winterthur, Switzerland. Association for Computational Linguistics.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Motter, A. E., De Moura, A. P. S., Lai, Y.-C., and Dasgupta, P. (2002). Topology of the Conceptual Network of Language. *Physical Review E*, 65(6):065102.

Murphy, M. L. (2003). *Semantic Relations and the Lexicon: Antonymy, Synonymy, and Other Paradigms*. Cambridge University Press, Cambridge.

Murphy, M. L. (2010). *Lexical Meaning*. Cambridge University Press,, Cambridge.

Nakhleh, L., Warnow, T., Ringe, D., and Evans, S. N. (2005). A Comparison of Phylogenetic Reconstruction Methods on an Indo-European Dataset. *Transactions of the Philological Society*, 103(2):171–192.

Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (1998). The University of South Florida Word Association, Rhyme, and Word Fragment Norms. `http://w3.usf.edu/FreeAssociation/`, Accessed: 20 April 2026.

Newman, M. (2006). Modularity and Community Structure in Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–8582.

Newman, M. (2018). *Networks*. Oxford University Press, Oxford, 2nd edition.

Newman, M. E. J. (2002). Assortative Mixing in Networks. *Physical Review Letters*, 89(20):208701.

Newman, M. E. J. (2003). Mixing Patterns in Networks. *Physical Review E*, 67(2):026126.

Nunberg, G. (1978). *The Pragmatics of Reference*. Indiana University Linguistics Club, Indiana.

Olander, T. (2022). Introduction. In Olander, T., editor, *The Indo-European Language Family: A Phylogenetic Perspective*, pages 1–17. Cambridge University Press, Cambridge.

Osthoff, H. and Brugmann, K. (1878). Vorwort. *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*, 1:iii–xx.

Ozaki, K., Shimbo, M., Komachi, M., and Matsumoto, Y. (2011). Using the Mutual k-Nearest Neighbor Graphs for Semi-supervised Classification on Natural Language Data. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 154–162, Portland, Oregon, USA. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global Vectors for Word Representation. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Pereira, H. B. d. B., Grilo, M., Fadigas, I. d. S., Souza Junior, C. T. d., Cunha, M. d. V., Barreto, R. S. F. D., Andrade, J. C., and Henrique, T. (2022). Systematic review of the "semantic network" definitions. *Expert Systems with Applications*, 210:118455.

Piwowarczyk, D. (2022). Computational Approaches to Linguistic Chronology and Subgrouping. In Olander, T., editor, *The Indo-European Language Family: A Phylogenetic Perspective*, pages 33–51. Cambridge University Press, Cambridge.

Rabinovich, E., Ordan, N., and Wintner, S. (2017). Found in Translation: Reconstructing Phylogenetic Language Trees from Translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, BC, Canada.

Radovanović, M., Nanopoulos, A., and Ivanović, M. (2010). Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. *Journal of Machine Learning Research*, 11:2487–2531.

Renfrew, C. (1987). *Archaeology and Language: The Puzzle of Indo-European Origins*. Penguin, London.

Ringe, D. (2022). What We Can (and Can't) Learn from Computational Cladistics. In Olander, T., editor, *The Indo-European Language Family: A Phylogenetic Perspective*, pages 52–62. Cambridge University Press, Cambridge.

Ringe, D., Warnow, T., and Taylor, A. (2002). Indo-European and Computational Cladistics. *Transactions of the Philological Society*, 100(1):59–129.

Roget, P. M. (1911). *Roget's Thesaurus of English Words and Phrases*. T. Y. Crowell Company.

Rosch, E. (1978). Principles of Categorization. In Rosch, E. and Lloyd, B. B., editors, *Cognition and Categorization*, pages 27–48. Lawrence Elbaum Associates.

Schleicher, A. (1861). *Compendium der vergleichenden Grammatik der indogermanischen Sprachen*. Hermann Böhler, Weimar.

Schlücker, B. and Hüning, M. (2009). Compounds and Phrases: A Functional Comparison between German A+N Compounds and Corresponding Phrases. *Italian Journal of Linguistics*, 21(1):209–234.

Schnabel, T., Labutov, I., Mimno, D., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.

Schrader, S. R. and Gultepe, E. (2023). Analyzing Indo-European Language Similarities Using Document Vectors. *Informatics*, 10(76):1–21.

Serva, M. and Petroni, F. (2008). Indo-European Languages Tree by Levenshtein Distance. *EPL (Europhysics Letters)*, 81:68005.

Sigman, M. and Cecchi, G. A. (2002). Global Organization of the WordNet Lexicon. *Proceedings of the National Academy of Sciences*, 99:1742–1747.

Speer, R., Chin, J., and Havasi, C. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 4444–4451, San Francisco, California, USA. AAAI Press.

Steyvers, M. and 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.

Thompson, B., Roberts, S., and Lupyan, G. (2018). Quantifying Semantic Alignment Across Languages. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 40, pages 2554–2559.

Torregrossa, F., Allesiardo, R., Claveau, V., Kooli, N., and Gravier, G. (2021). A Survey on Training and Evaluation of Word Embeddings. *International Journal of Data Science and Analytics*, 11:85–103.

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Scientific Reports*, 9(5233).

Turian, J., Ratinov, L.-A., and Bengio, Y. (2010). Word Representations: A Simple and General Method for Semi-Supervised Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Utsumi, A. (2015). A Complex Network Approach to Distributional Semantic Models. *PLoS ONE*, 10(8):e0136277.

Van Paridon, J. and Thompson, B. (2020). subs2vec: Tools for training and evaluating word embeddings based on subtitles. `https://github.com/jvparidon/subs2vec`. Accessed: 25 April 2026.

van Paridon, J. and Thompson, B. (2021). subs2vec: Word Embeddings from Subtitles in 55 Languages. *Behavior Research Methods*, 53:629–655.

Veremyev, A., Semenov, A., Pasiliao, E. L., and Boginski, V. (2019). Graph-based Exploration and Clustering Analysis of Semantic Spaces. *Applied Network Science*, 4(1):104.

Škrlj, B. and Pollak, S. (2019). Language Comparison via Network Topology. In *Statistical Language and Speech Processing: 7th International Conference, SLSP 2019, Ljubljana, Slovenia, October 14–16, 2019, Proceedings*, page 112–123, Berlin, Heidelberg. Springer-Verlag.

Wallace, A. F. C. (1970). A Relational Analysis of American Kinship Terminology. *American Anthropologist*, 72(4):841–845.

Ward, G. (1996). Moby Thesaurus List. http://www.gutenberg.org/ebooks/3202 Accessed: 20 April 2026.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244.

Watts, D. J. and Strogatz, S. H. (1998). Collective Dynamics of 'Small-World' Networks. *Nature*, 393:440–442.

Weiss, M. (2015). The Comparative Method. In Bowern, C. and Evans, B., editors, *The Routledge Handbook of Historical Linguistics*, pages 127–145. Routledge, London, 1st edition.

Xu, Q., Peng, Y., and Li, P. (2023). Large-scale Network Analyses Reveal Cross-Language Differences in Semantic Structures: A Comparative Study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, pages 3484–3491. Cognitive Science Society.

A. Appendix

A.1. Code Availability

The code used for the analysis in this thesis is available in the following GitHub repository:

`https://github.com/laletuver/semantic-network-indoeuropean-languages`

The repository contains the code used for the computational analysis, including preprocessing, semantic network construction, the analysis of graph-theoretic properties, correlation analysis, hierarchical clustering, and the generation of figures. The original Subs2Vec by van Paridon and Thompson (2021) is an external data resource subject to separate access and licensing conditions. The repository therefore provides the scripts needed to reproduce the analysis, while the embedding files must be obtained separately from the original Subs2Vec source.

A.2. Disclosure of AI Use

Artificial intelligence tools were used during the preparation of this thesis. Grammarly was used to support proofreading, grammar correction, and stylistic revision. ChatGPT Plus was also used to support the revision and the improvement of academic wording, and the improvement of structure and flow of selected passages. OpenAI Codex was used to support coding-related tasks, including debugging, restructuring scripts, and improving the implementation of parts of the computational workflow. All code improvements were reviewed, further edited, and tested by the author. The content, data analysis, interpretation of results, and scholarly argumentation represent the author's own work and intellectual contribution.