Beyond the Surface: A Computational Exploration of Linguistic Ambiguity

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering by Research

by

Anmol Goel
2021701045
anmol.goel@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **"Beyond the Surface: A Computational Exploration of Linguistic Ambiguity"** by Anmol Goel, has been carried out under my supervision and is not submitted elsewhere for a degree.

June|28|2023

Date

Adviser: Prof. Ponnurangam Kumaraguru

To Mom and Dad

Acknowledgments

I would like to express my deepest gratitude and appreciation to all those who have contributed to the completion of this thesis. Their support, guidance, and encouragement have been invaluable throughout this academic journey.

First and foremost, I would like to extend my heartfelt gratitude to my thesis advisor, Prof. Ponnurangam Kumaraguru. His expertise, patience, and unwavering support have been instrumental in shaping this research project. I am grateful for his guidance, insightful feedback, and constant motivation that pushed me to strive for excellence. From a random email about an internship with him to this thesis, I am grateful for all the opportunities, trust, belief and support he has given. My sincere thanks go to the faculty and staff of IIIT. Their commitment to academic excellence and their tireless efforts to create a conducive learning environment have been pivotal to my growth as a researcher. I am grateful for the opportunity to learn from such distinguished scholars and for the resources and facilities made available to me throughout this journey. I immensely enjoyed all my interactions with Prof. Charu, Prof. Manish and Prof. Vinoo. I have also had the opportunity to interact with scholars from multiple institutions - Prof. Saptarshi, Prof. Ravindran, Prof. Ashutosh - thanks for all the interactions and conversations.

I would like to express my deep appreciation to my family for their unwavering love, support, and encouragement. Their belief in my abilities, their patience during my periods of stress, and their constant motivation have been invaluable. I am profoundly grateful for their sacrifices and understanding, which allowed me to pursue my academic goals. Additionally, I would like to acknowledge my friends and colleagues who have provided me with invaluable support and companionship during this challenging process. Their encouragement, stimulating discussions, and shared experiences have been a constant source of inspiration. I am grateful for their friendship and the countless moments of laughter that helped me maintain balance in my life. Thank you Dr. Geetika, Prashant, Monika, Nidhi, Ravi, Pranjal, Sarthak, Gaurav and others for making my time in Hyderabad memorable.

In conclusion, this thesis represents the culmination of the efforts and contributions of many individuals, and I am truly grateful for their support. Their collective assistance and belief in me have been essential in bringing this project to fruition. I hope that my work can contribute to the existing body of knowledge and inspire future researchers to explore new frontiers. Thank you all for being a part of this remarkable journey.

Finally, I thank you, the reader. I hope you find something valuable here.

Abstract

The issue of ambiguity in natural language poses a significant challenge to computational linguistics and natural language processing. Ambiguity arises when words or phrases can have multiple meanings, depending on the context in which they are used. In natural language processing, addressing the challenge of ambiguity is crucial for building more accurate and effective language models that can better reflect the complexity of human communication. In this thesis, we investigate two specific forms of linguistic ambiguities - polysemy, which is the multiplicity of meanings for a specific word, and tautology, which are seemingly uninformative and ambiguous phrases used in conversations. Both phenomena are widely-known manifestations of linguistic ambiguity - at the lexical and pragmatic level, respectively.

The first part of the thesis focuses on addressing this challenge by proposing a new method for quantifying the degree of polysemy in words, which refers to the number of distinct meanings that a word can have. The proposed approach is a novel, unsupervised framework to compute and estimate polysemy scores for words in multiple languages, infusing syntactic knowledge in the form of dependency structures. The framework adopts a graph-based approach by computing the discrete Ollivier Ricci curvature on a graph of the contextual nearest neighbors. The effectiveness of the framework is demonstrated by significant correlations of the quantification with expert human-annotated language resources like WordNet. The proposed framework is tested on curated datasets controlling for different sense distributions of words in three typologically diverse languages - English, French, and Spanish. The framework leverages contextual language models and syntactic structures to empirically support the widely held theoretical linguistic notion that syntax is intricately linked to ambiguity/polysemy.

The second part of the thesis explores how language models handle colloquial tautologies, a type of redundancy commonly used in conversational speech. Colloquial tautologies pose an additional challenge to language processing, as they involve the repetition of words or phrases that may appear redundant, but convey a specific meaning in a given context. We first present a dataset of colloquial tautologies and evaluate several state-of-the-art language models on this dataset using perplexity scores. We conduct probing experiments while controlling for the noun type, context and form of tautologies. The results reveal that BERT and GPT2 perform better with modal forms and human nouns, which aligns with previous literature and human intuition.

We hope this work bolsters further research on ambiguity in language models. Our contributions have important implications for the development of more accurate and reliable natural language processing systems.

Contents

List of Figures

List of Tables

Chapter *1*

Introduction

Of course language can never be made absolutely neutral and colourless; but unless its ambiguities are understood, accuracy of thought is impossible, and the injury done is proportionate to the logical force and general vigour of character of those who are misled.

James Fitzjames Stephen, *Liberty, Equality, Fraternity, 1873*

Language is one of the most defining characteristics of human beings. It is a complex system of communication that enables us to express our thoughts, feelings, and ideas to others. Language is so deeply ingrained in our lives that we often take it for granted, yet it is a remarkable and fascinating phenomenon that has evolved over thousands of years.

At its most basic level, language is a system of symbols, sounds, and rules that enables communication between individuals [Adrian et al., 2001]. It includes spoken languages, sign languages, and written languages. In many ways, language is what sets us apart from other animals [Balconi, 2010]. While other animals communicate through a range of vocalizations and gestures, human language is far more sophisticated and varied. One of the key features of human language is its **ability to convey abstract concepts and ideas**. Unlike other animals, humans can communicate about things that don't exist in the physical world, such as love, justice, or freedom. Language also enables us to **communicate across time and space**, through written records and digital media. This ability to transmit information and ideas across generations and continents has been a key factor in human progress and evolution [MacWhinney, 2005]. Another important aspect of language is its **adaptability and flexibility**. Languages evolve over time [Christiansen and Kirby, 2003] as they are used by different communities of speakers. New words are created, old words fall out of use, and grammar and pronunciation change. This is evident in the vast number of languages spoken around the world, each with their own unique

characteristics and history. Language also plays a crucial role in **socialization and identity**. Through language, individuals are able to communicate their cultural heritage, beliefs, and values. Language can create a sense of belonging and community, as well as reinforcing differences and divisions between groups [Mercuri, 2012, Rovira, 2008]. This is particularly evident in the context of language and nationalism, where language can be a powerful symbol of national identity and pride.

Despite the many benefits of language, there are also challenges and limitations. One issue is the **diversity of languages** spoken around the world, which makes it difficult to study and compare them all. Another challenge is the fact that **language is constantly evolving** and changing, making it hard to create a stable set of rules that can be applied across time and space. Additionally, there are many debates and controversies within the field itself, such as the best way to approach language acquisition or **the role of culture in shaping language use**. There are also broader social and political issues related to language, such as **linguistic discrimination** [Wee, 2005] or the use of **language as a tool of power** and control. The current thesis is primarily concerned with a specific problem in language, which is **ambiguity**. While there are many issues in linguistics and language more broadly, this thesis focuses on a more intrinsic problem of language. Ambiguity arises when a word or phrase can have multiple meanings or interpretations, leading to confusion or miscommunication. This problem is particularly relevant in natural language processing and machine learning, where ambiguity can create challenges for automated language understanding.

By examining the nature of ambiguity in language and exploring strategies for resolving it, the current thesis aims to contribute to the development of more accurate and effective language technologies. While other issues in linguistics and language are important and worthy of attention, this thesis seeks to make a valuable contribution to the specific area of ambiguity within the field.

1.1 Ambiguity in Language

Ambiguity in language refers to situations where a word or phrase can have multiple meanings or interpretations [Fromkin et al., 2018]. It is a common phenomenon in natural language and can arise due to various factors, such as the context, syntax, or semantics of the language. Ambiguity can have both positive and negative effects on communication. On the one hand, it can lead to creativity and humor in language use. For example, puns and wordplay rely on the multiple meanings of words for their effect ("atheism is a non-prophet institution" is an wordplay on the common phrase "non-profit institution" by replacing the word *profit* to its homophone *prophet*). On the other hand, ambiguity can also lead to confusion and misunderstandings, particularly in situations where clarity is important, such as legal or technical documents. In order to reduce ambiguity in communication, speakers and writers can use various strategies. One strategy is to use more specific language, such as using synonyms or defining terms clearly. Another strategy is to provide additional context or information, such as using examples or illustrations. Finally, speakers and writers can also use language that is appropriate to their audience, avoiding jargon or technical terms that may not be familiar to their listeners or readers.

Typically, three main types of ambiguity in language are considered - lexical, syntactic, and semantic ambiguity [Fromkin et al., 2018]. Each of these types can have different effects on communication and can require different strategies for clarification.

1.1.1 Lexical Ambiguity

Lexical ambiguity arises when a word has multiple meanings. This can occur due to homophones, homographs, or polysemes. Homophones are words that have the same pronunciation but different meanings, such as "write" and "right". Homographs are words that are spelled the same but have different meanings, such as "lead" (the metal) and "lead" (to guide). Polysemes are words that have multiple related meanings, such as "bank" (a financial institution) and "bank" (the edge of a river). In such cases, the intended meaning can be inferred from the context in which the word is used.

1.1.2 Syntactic Ambiguity

Syntactic ambiguity arises when a sentence can be parsed in multiple ways. This can occur due to structural ambiguity or garden path sentences.

Structural ambiguity arises when the sentence structure allows for multiple interpretations, such as in the sentence "I saw the man with the telescope." This sentence can be interpreted in two ways - either the speaker saw a man who was carrying a telescope or the speaker saw a man through a telescope.

Garden path sentences, on the other hand, are sentences that initially lead the reader or listener down a particular syntactic path, only to be followed by an unexpected or incorrect interpretation, such as in the sentence "The horse raced past the barn fell." Initially, when the reader encounters the phrase "the horse raced past the barn," they form a mental image of a horse racing past a barn. However, as they reach the word "fell," they realize that it doesn't fit with the expected structure of the sentence. The phrase "raced past the barn" acts as a prepositional phrase modifying the horse, but the verb "fell" seems disconnected from the rest of the sentence. To resolve the confusion, the reader must backtrack and reinterpret the sentence. They realize that "the horse" is not the subject of the verb "fell" but instead a noun phrase acting as the direct object of the verb "raced." The actual subject of the verb "fell" is omitted, creating a syntactic ambiguity.

1.1.3 Semantic Ambiguity

Semantic ambiguity arises when a word or phrase can be interpreted in multiple ways based on the meaning of the words used. This can occur due to lexical ambiguity, as discussed earlier, or due to structural or contextual ambiguity. Structural ambiguity arises when the sentence structure allows for multiple interpretations based on the meaning of the words used, such as in the sentence "Visiting relatives can be boring.", which can be interpreted as either the relatives being boring or the act of visiting being boring. Contextual ambiguity arises when the meaning of a word or phrase depends on

the context in which it is used, such as in the sentence “He saw her duck” which can be interpreted as either the woman ducking or the man seeing a duck.

Lexical and Semantic ambiguities form the foci of this work, as we will discuss in the subsequent sections.

1.2 Pyramid of Language Analysis

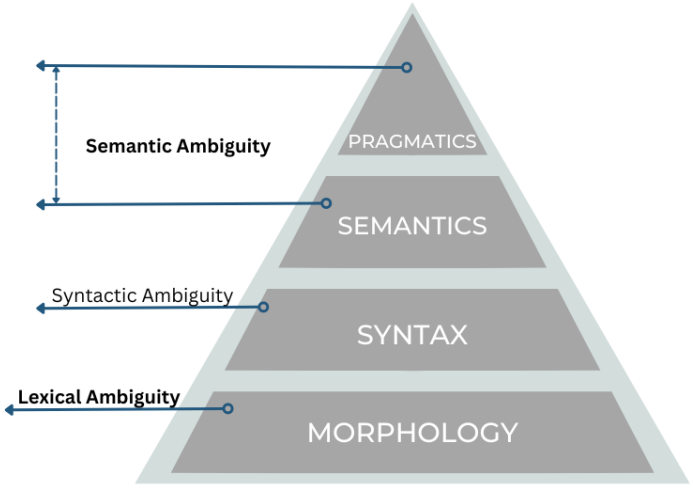

Figure 1.1 Pyramid of Language Analysis. The foci of this thesis - Lexical and Semantic Ambiguities are in bold.

The pyramid of linguistics [Ivanov, 2019] is a theoretical framework that describes the hierarchy of different levels of language analysis. At the base of the pyramid is morphology, which deals with the smallest units of meaning in a language, such as individual words and their parts. Above morphology is syntax, which concerns the rules governing how words are combined to form larger units, such as phrases and sentences. Semantics comes next, which is concerned with the meaning of words and how they combine to form meaningful sentences. Finally, at the top of the pyramid is pragmatics, which is concerned with the context in which language is used and the social and cultural factors that influence its meaning. Together, these levels provide a comprehensive understanding of how language works at various levels of analysis.

Figure 1.1 shows a rough mapping between the levels of language analysis and the types of linguistic ambiguities.[1] In this thesis, we focus on Lexical and Semantic Ambiguities concerning the morphological and semantics levels in the pyramid, respectively.

1.3 Polysemy

Polysemy is a type of lexical ambiguity which refers to the phenomenon where a single word has multiple meanings or senses. It is a common feature of human language and plays an important role in how we communicate and understand each other [Nerlich and Clarke, 2001, Falkum and Vicente, 2015]. Polysemy can be observed in nearly every language and is a product of the ways in which words evolve and change over time [Bréal, 1904].

One of the key features of polysemy is that it can be difficult to determine which meaning of a word is intended in a given context. For example, the word "bank" can refer to a financial institution, the side of a river, or a place where airplanes park, among multiple other meanings. The intended meaning of "bank" can often be determined by context, such as the surrounding words or the speaker's tone of voice. However, in some cases, the intended meaning may be ambiguous, which can lead to confusion or misunderstandings. Polysemy can also be influenced by culture and history. For example, the word "revolution" has different meanings in different contexts. In a political context, it typically refers to a violent overthrow of a government, while in a scientific context, it can refer to a complete rotation of an object around its axis. The historical events that led to the French Revolution have also influenced the way the word is used in contemporary language, with "revolution" often carrying connotations of radical change and upheaval. Another important aspect of polysemy is that it allows for linguistic creativity and innovation [Murphy, 1997]. Words can take on new meanings or senses over time, and these changes can be driven by cultural shifts, technological advancements, or other societal changes. For example, the word "tweet" was originally used to describe the sound made by a bird, but it has now taken on a new meaning in the context of social media. Polysemy can also pose challenges for language learners, particularly those who are trying to learn a new language as an adult [Crossley et al., 2010]. Because the same word can have multiple meanings, learners may struggle to understand the intended meaning of a word in a given context. However, as learners become more familiar with a language, they begin to develop a sense of how words are used in different contexts and how they can be used to convey multiple meanings. By recognizing the complexities of polysemy, we can better appreciate the richness and diversity of human language.

The first contribution of the thesis deals with quantification of polysemy. An interesting application of a quantifiable polysemy score is that it can help natural language systems to generate less ambiguous outputs.

[1] While there are more granular levels of ambiguity defined in literature, like discourse and pragmatic ambiguity, in the scope of this thesis, we only consider the three levels of ambiguity.

1.4 Tautology

A colloquial tautology is a type of expression in which a word or phrase is unnecessarily repeated or restated, resulting in redundant or unnecessary information [Gibbs and McCarrell, 1990]. These types of phrases are commonly used in everyday speech and are often considered to be a feature of colloquial language rather than formal writing. One of the most common examples of a colloquial tautology is the phrase "ATM machine," where "ATM" stands for "automated teller machine." In this case, the word "machine" is repeated, which is unnecessary because the "M" in "ATM" already stands for "machine." Other examples of colloquial tautologies include "free gift," "hot water heater," and "added bonus." While colloquial tautologies can sometimes be used for emphasis or rhetorical effect, they are generally considered to be a sign of imprecise or sloppy language use. In formal writing, it is generally recommended to avoid tautological expressions and to use precise and concise language to convey meaning [Wierzbicka, 1987]. In addition to colloquial tautologies, there are also other types of tautologies, such as logical tautologies and mathematical tautologies. Logical tautologies are statements that are always true, regardless of the truth value of their constituent parts. For example, the statement "A or not A" is always true, regardless of the value of A. Mathematical tautologies, on the other hand, are mathematical statements that are true by definition or by virtue of their logical structure.

Colloquial tautologies like "boys will be boys" can create ambiguity by relying on vague or imprecise language that can be interpreted in different ways depending on the context. In this particular phrase, the tautology "boys will be boys" implies that boys will behave in a certain way regardless of the situation or environment they are in. However, this phrase can be interpreted in different ways depending on the context, which can create ambiguity. For example, if this phrase is used to excuse or dismiss inappropriate behavior by boys or men, it can perpetuate harmful gender stereotypes and suggest that such behavior is acceptable or even expected. In this context, the phrase can be seen as ambiguous or even contradictory, as it seems to suggest that certain behavior is innate to boys while simultaneously acknowledging that such behavior may be problematic.

The second contribution of the thesis concerns itself with probing Large Language Models (LLMs), specifically BERT [Devlin et al., 2019] and GPT2 [Radford et al., 2019] - to investigate their competence in handling colloquial tautologies. We add to the growing body of literature on analysing the capabilities of language models across various linguistic phenomena.

1.5 Thesis Contribution

As discussed in the previous sections, this thesis deals with ambiguity in languages and aims to solve the challenges of current NLP systems dealing with linguistic ambiguities. To this end, we make the following core contributions:

1. In the first work, we propose a novel, graph-based syntax-aware framework to measure polysemy scores of words in multiple languages (English, Spanish and French). We outperform state-of-

the-art and random baselines and observe an increment of 0.3 points in correlation with human rankings of polysemy measures. [2]

2. We validate the long-held notion that syntax is intricately linked with semantics, thus influencing the polysemy of words in a language.

3. In the second work, we first create a dataset of tautologies while controlling for noun types, syntactic form and the context for a tautology. We test two state-of-the-art models: GPT2 and BERT on their pragmatic competence on tautologies using perplexity measures. We provide evidence of a unique type of pragmatic understanding of LLMs by highlighting that GPT2 outperforms BERT with lower perplexities of tautologies. Additionally, we highlight that, akin to human understanding, LLMs also prefer tautologies with human nouns and modal forms as compared to concrete or abstract nouns.

1.6 Thesis Organization

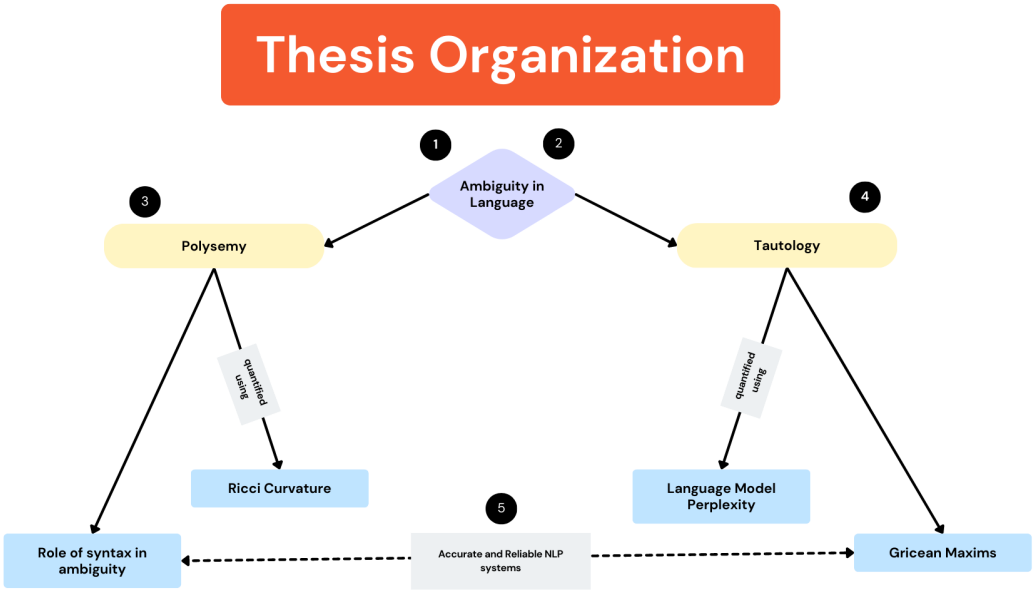

Figure 1.2 Thesis Organization.

Figure 1.2 illustrates the logical outline of the thesis and shows the associated concepts for each contribution.

[2]The work is a full paper accepted at **EMNLP 2022**.
Goel, A., Sharma, C., and Kumaraguru, P. An Unsupervised, Geometric and Syntax-aware Quantification of Polysemy. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP).

The thesis is organised into 6 chapters. Chapter 2 discusses relevant and related work dealing with ambiguities in NLP with a focus on corpora, models and frameworks. In Chapter 3, we discuss the polysemy quantification framework which is syntax and graph-aware. Furthermore, we conduct ablation studies to establish the influence of syntax on polysemy scores. In Chapter 4, we discuss the colloquial tautology dataset and the motivation for creating a controlled dataset. We then conduct perplexity-based measurements on LLMs to investigate the pragmatic competence of current LLMs in handling tautologies with discussions on its relation to the Gricean Maxims. Finally, we conclude the thesis in Chapter 7 and discuss implications and limitations of the current work along with future directions.

Chapter *2*

Background

Ambiguity seems to be an essential, indispensable element for the transfer of information from one place to another by words.

Lewis Thomas, *The Lives of a Cell: Notes of a Biology Watcher* [Thomas, 1978]

This chapter will cover the necessary background and motivation from previous literature on ambiguity in languages. Later, we highlight the relevant background specifically for Polysemy and Tautology. Figure 2.1 represents the structure of this chapter.

2.1 Ambiguity in Language

Table 2.1 provides a summary of key takeaways from relevant papers on ambiguity in language which provide the relevance and contextualization of the contributions of this thesis.

Ambiguity of language has been addressed as early as in the writings of Aristotle but relatively recent linguistic research in the form of Zipf's Principle of Least Effort [Bain, 1950] heralded a new understanding of human cognition and language systems positing the tradeoff between efficiency and brevity in communication systems [Piantadosi et al., 2012].

[Eddington and Tokowicz, 2015] suggests that the majority of words in the English language have multiple meanings, and these meanings are processed and represented differently in the human mind highlighting the importance of understanding how meaning similarity influences ambiguous word processing. [Kess and Hoppe, 1978] highlights that ambiguity is not a problematic source of difficulty for individuals, and the study of the resolution of ambiguity may be a useful tool in the comprehension of sentence processing in general. This suggests that ambiguity is a natural and integral part of language, and that understanding how it is resolved can provide insight into how language is processed and un-

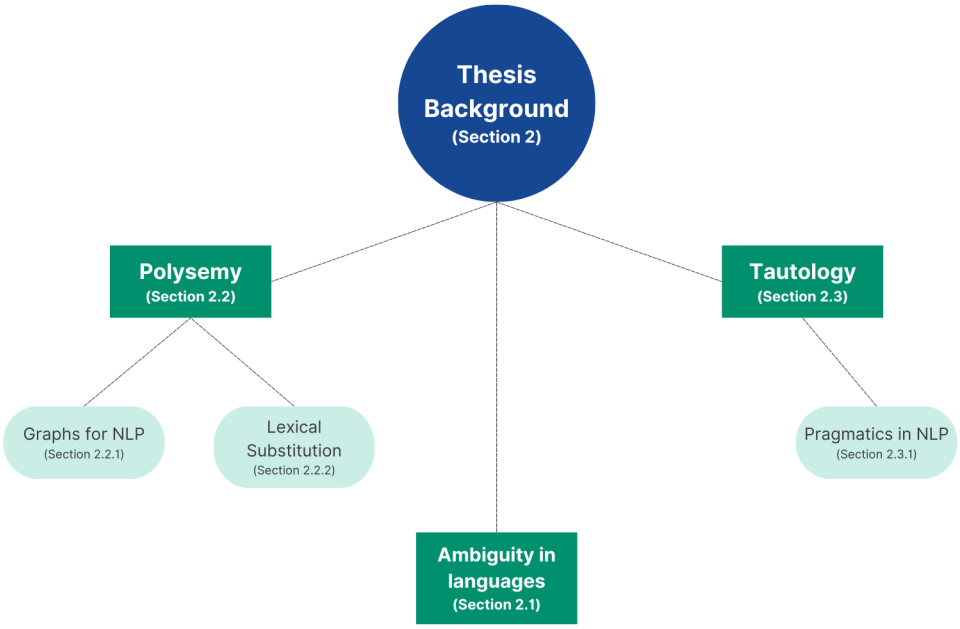

Figure 2.1 Background of this thesis.

derstood. The importance of disambiguation and the potential role of word sense in improving search accuracy of information retrieval systems is highlighted by [Krovetz and Croft, 1992] which establishes that considerable amount of ambiguity exists in databases and word senses provide a significant separation between relevant and non-relevant documents. [Winkler, 2015] emphasizes the importance of exploring how production and perception of ambiguity interact with each other and a reference system when ambiguity is generated and resolved. This highlights the need to understand the context in which ambiguity occurs and how it is resolved. [Mohammad, 2018] suggests that the factors causing ambiguity are shared among languages and ambiguity is transferable. This finding supports the idea that understanding ambiguity can provide insight into language processing across different languages. [Andreu and Corominas-Murtra, 2011] argues that the emergence of ambiguous codes is an unavoidable byproduct of efficient communication proposing that ambiguity may be a natural consequence of language evolution and adaptation. [MacDonald et al., 1994] suggests that lexical and syntactic ambiguities are resolved by the same processing mechanisms underscoring the importance of understanding the role of lexical and syntactic information in disambiguation. Finally, [Tuggy, 1993] suggests that ambiguity and vagueness occupy opposite ends of a continuum, with polysemy in the middle. This highlights the need to understand the different types of ambiguity and their relationship to other linguistic phenomena.

Within the NLP community, [Liu et al., 2023] propose a linguist-annotated benchmark dataset called *AMBIENT* to evaluate state-of-the-art pretrained LLMs on ambiguity and disambiguation. [Garí Soler and Apidianaki, 2021] investigate the capabilities of BERT on ambiguity and present a dataset for polysemous words. [Ortega-Martín et al., 2023] present the first linguistic ambiguity analysis of chatGPT

Paper	Takeaways
How meaning similarity influences ambiguous word processing: the current state of the literature [Eddington and Tokowicz, 2015]	Majority of words in the English language have multiple meanings and these meanings are processed and represented differently in the human mind.
On psycholinguistic experiments in ambiguity [Kess and Hoppe, 1978]	Most experimental paradigms agree that ambiguity is not a problematic source of difficulty for individuals and the study of the resolution of ambiguity may be a useful tool in the comprehension of sentence processing in general.
Lexical ambiguity and information retrieval [Krovetz and Croft, 1992]	Considerable amount of ambiguity exists in databases. Word senses provide a significant separation between relevant and non-relevant documents.
Ambiguity: Language and Communication [Winkler, 2015]	The production and perception of ambiguity can only be understood by exploring how these factors interact with each other and a reference system when ambiguity is generated and resolved.
The Nature of Ambiguity across Languages [Mohammad, 2018]	The factors causing ambiguity are shared among languages and ambiguity is transferable.
On ambiguity. Its locus in the architecture of Language and its origin in efficient communication [Andreu and Corominas-Murtra, 2011]	The emergence of ambiguous codes is an unavoidable byproduct of efficient communication.
The lexical nature of syntactic ambiguity resolution [MacDonald et al., 1994]	Lexical and syntactic ambiguities are resolved by the same processing mechanisms.
Ambiguity, polysemy and vagueness [Tuggy, 1993]	Ambiguity and vagueness occupy opposite ends of a continuum with polysemy in the middle.

Table 2.1 Summary of Literature on Ambiguity in Language (arranged in increasing order of relevance to the thesis; top to bottom)

and highlight the capabilities of chatGPT in detecting ambiguities like homonymy and polysemy. However, syntactic ambiguity is shown to be a challenging setting for chatGPT to detect.

2.2 Polysemy

Polysemy is a natural outcome of lexical semantic change [Bréal, 1904] by virtue of words gaining new meanings over time. Even though polysemous words present processing costs, their use in everyday discourse has significant pragmatic and discourse functions [Nerlich and Clarke, 2001]. Polysemy is notoriously difficult to treat and characterize both theoretically and empirically [Falkum and Vicente, 2015]. Evidence from cognitive linguistics suggests that polysemous words connect to the same abstract lexical representations in our minds but are distinct within that representational set [Pylkkänen et al., 2006]. Recent evidence highlights the role of context and pragmatics for a unified understanding and disambiguation of polysemy [Falkum, 2015]. [Kelih, 2008] characterize polysemy in six unrelated languages and connect ambiguity to factors like sample size and parts of speech. [Glynn, 2009] underscore the importance of considering a network based approach to understand linguistic ambiguities. [Kelih and Altmann, 2015, Durkin and Manning, 1989] suggest that polysemous words inhabit a continuous space of semantic meaning and the variation in the multiple senses of polyemous words is captured by a continuum.

Recent works in computational linguistics for ambiguity mostly deal with word sense disambiguation [Pasini et al., 2021, Wiedemann et al., 2019], word-in-context tasks [Pilehvar and Camacho-Collados, 2019] and analyzing polysemy in language models like BERT [Garí Soler and Apidianaki, 2021]. While some previous works [Erk and McCarthy, 2009, Friedrich et al., 2012] acknowledge polysemy even in particular instances, relatively less attention has been paid towards quantifying polysemy using current NLP tools. [Pimentel et al., 2020] measure ambiguity in language from an information-theoretic lens but their approach requires a large number of sentences to give a good upper bound on ambiguity estimates. [Xypolopoulos et al., 2021] leveraged contextual language models like BERT to estimate polysemy but they rely on dimensionality reduction and sensitive hyperparameters.

Works like [Reif et al., 2019, Haber and Poesio, 2021] have explored the geometry of BERT embeddings and their relation to polysemy levels thus highlighting the importance of neural embeddings in the quantification of polysemy levels of lexicons.

2.2.1 Lexical Substitution

Lexical substitution is the task of finding relevant contextual replacements of a word given its context. To generate good quality contextual replacements, previous works have relied heavily on distributional semantic models like word2vec [Mikolov et al., 2013] and specialized language models like context2vec [Melamud et al., 2016]. In all models, the generated substitutes are ranked based on some relation with the target word to be replaced. Recent advances in language models like the Transformer-based BERT [Devlin et al., 2019] and XLNet [Yang et al., 2020] rely on the bidirectional context and the special [MASK] token based training to generate contextual substitutes. [Zhou et al., 2019] showed that BERT performs poorly on lexical substitution and proposed a dropout based approach which is even more computationally expensive due to the large number of forward passes required. Supervised

approaches [Lacerra et al., 2021] often rely on manually curated databases and sense inventories like WordNet, Wikipedia or BabelNet. [Arefyev et al., 2020] is a recent neural lexical substitution method which injects information about the target word in the form of probability distribution of possible word subtitutes based on word frequencies.

2.2.2 Graphs and NLP

Traditional works in linguistics have used language networks and graphs for analyzing morphological complexity [Inglese and Brigada Villa, 2021], ambiguity [Čech et al., 2017] and phonetics [Yamshchikov et al., 2020]. Graph-based frameworks like the Chinese Whispers algorithm [Biemann, 2006] have worked at the intersection of graph theory and linguistics with applications in word sense disambiguation. [Mitra et al., 2014] propose to use dependency-based contextual neighbors for the task of novel word sense induction. Our contribution combines ideas from [Mitra et al., 2014] with contextual word embeddings to measure polysemy. Recent advances in Graph Neural Networks (GNNs) has opened new avenues to apply network based approaches to language problems. While language networks have been analysed before, GNNs provide an alternative to traditional methods with more natural inductive biases for syntactic models to work with. The combination of graphs and language models has proved to be effective in incorporating semantics and syntax in language problems [Marcheggiani and Titov, 2020, Ahmad et al., 2021, Xu et al., 2021].

2.3 Tautology

Scarce attention has been paid towards understanding tautological constructions in English [Vilinbakhova and Escandell-Vidal, 2020], especially within computational linguistics. Just based on form, tautologies seem redundant and uninformative; however, they are frequently used in speech to evoke and imply a shared assumption about the world. This has important implications in understanding discourse and pragmatics, especially so for computational models of pragmatics. Gricean conversational implicature - the additional meaning or implied information that is inferred by the listener or reader in a conversation, beyond the literal meaning of the words used, based on principles of cooperation, relevance, and shared knowledge - is most often evoked to justify the use of tautologies [Ward and Hirschberg, 1991]. [Gibbs and McCarrell, 1990] provide foundational work on understanding nominal tautologies by conducting acceptability studies. They establish that syntax and lexical construction of tautologies influence the acceptability ratings of tautologies in English speakers. [Vilinbakhova and Escandell-Vidal, 2020] establish the contribution and synchrony of different dimensions of knowledge (like encyclopaedic, normative, etc) in interpreting tautological constructions in English. Tautological expressions have been interpreted and characterized by syntax and lexical form [Gibbs and McCarrell, 1990], context at the pragmatic level [Farghal, 1992], case markers in specific languages like Japanese [Kwon, 2009] and proper names [Vilinbakhova and Escandell-Vidal, 2021].

Tautologies in languages other than English have also been analysed with [Kwon, 2014] using Korean nominal tautologies to study cognitive aspects of pragmatics. [Sonnenhauser, 2017] establish the confluence of semantics, syntax and pragmatics in understanding the German tautology "Wer kann, der kann" ("he who can, can") and [Kwon, 2009] studying Japanese tautologies with respect to case markers.

2.3.1 Pragmatics in NLP

Recent advances in neural approaches to language models have lead to improvements in performance on benchmarks built for NLP tasks ranging from Named Entity Recognition [Wang et al., 2021] to Sentiment Analysis [Raffel et al., 2020]. Pragmatic reasoning, the highest frontier on the NLP pyramid (Figure 1.1), is a major milestone for current NLP models. Neural approaches to understand the subtleties and nuances of discourse - beyond the lexical and semantic meaning of word forms - have been gaining recent traction within the NLP community [Kabbara, 2019]. [Ettinger, 2020] explore LLMs and their efficacy on semantic and pragmatic tasks. Priming experiments on BERT reveal that LLMs are sensitive to contextual constraints which has important implications for pragmatic development of pretrained LLMs [Misra et al., 2020]. More recently, [Pandia et al., 2021] test the pragmatic competence of LLMs through discourse connectives.

With the recent democratisation of LLMs and their increased access via tools like GPT [Brown et al., 2020] and chatGPT [OpenAI, 2023], it is imperative to test the pragmatic competence of such models.

Chapter *3*

Syntax-aware Polysemy Quantification[1]

. . . polysemy, which is the greedy habit some words have of taking more than one meaning for themselves.

Eric McKean, *The joy of lexicography, 2007*[2]

3.1 Motivation

Polysemy is a phenomenon prevalent in everyday language use where the same lexical unit (or word form) is associated with multiple distinct yet *related* meanings (or senses). Determining which words are polysemous can help in filtering data for linguist studies, creation of sense corpora and the anthropological study of language. Consider the following sentences:

1a His **aunt** is his legal guardian.

2a The dog would always **bark** at mailmen.

2b The tree's **bark** was rusty brown.

3a The **mouth** of the wine was dry.

3b I have three **mouths** to feed.

3c You can see the **mouth** of the river from here.

Polysemy is distinct from monosemy (a word form with only one meaning; 1a) and homonymy (multiple *unrelated* senses of the same word form; 2a-b). The polysemous senses of a word often have

[1] **The work is a full paper accepted at EMNLP 2022.**
***Goel, A.*, Sharma, C., and Kumaraguru, P. An Unsupervised, Geometric and Syntax-aware Quantification of Polysemy. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP).**

[2] https://www.ted.com/talks/erin_mckean_the_joy_of_lexicography

metonymic (3a-b) or metaphorical (3c) relations among them [Vicente and Falkum, 2017]. Polysemy is a central feature of natural languages and proliferates almost every word to varying degrees in the lexicon of a language. Attempts [Piantadosi et al., 2012] at explaining the presence of ambiguity[3] in language suggest that polysemy is a desirable property for language systems since it allows efficient communication by allowing simpler units to be reused. Ambiguity and polysemy have sparked debate among linguists and philosophers for decades but relatively little attention has been paid to analyze and measure polysemy in language by computational linguists. While a human listener is easily able to disambiguate the specific sense of the word being used in context, it is notoriously difficult for NLP systems to separate the distinct senses of a word being used [Yenicelik et al., 2020].

Recently, there has been widespread attention on including syntactic knowledge in various computational linguistic systems and studies - ranging from syntax aware language models [Zhou et al., 2020] to syntax informed sentiment analysis [Hou et al., 2021]. Recent works have identified [Čech et al., 2017] an intricate link between the syntactic properties of a lexical unit and its ambiguity (or lack thereof) since the meaning of a word is influenced by its syntactic as well as semantic context. The fact that most open class word forms are associated with multiple related senses hints at the possible role that syntax plays in influencing polysemy. Syntactic structures can constrain the possible contexts a word form may be used in, thus there is an implicit relation between the semantics of a lexical unit and its associated polysemy. Motivated by these recent linguistic findings, we operationalize the polysemy of a word form as being influenced by both - its semantic variability and its importance in the syntactic network.

The level of polysemy a word possesses is highly subjective and varies widely across annotators [Artstein and Poesio, 2008]. To aid annotators in creating, validating and qualitatively analysing sense inventories, having an estimate of the ambiguity a word possesses could be very helpful. This measure then acts as a proxy to how many (or how few) senses a word in a certain language possesses. A quantification of polysemy is also helpful in Information Retrieval systems as they can be used to rank more relevant results [Krovetz, 1997]. Polysemic knowledge can also help improve cross-lingual alignment of embedding spaces and cross-lingual transfer [Garí Soler and Apidianaki, 2021].

While recent contextual embedding models like BERT, XLM and RoBERTa have been shown to possess the ability to distinguish between different senses of a word [Garí Soler and Apidianaki, 2021], less attention has been paid towards quantifying the level of polysemy that a word represents - a measure which is continuous and can be compared across lexica. Attempts at quantifying polysemy either rely on large amounts of data [Pimentel et al., 2020] and/or on carefully tuned hyperparameters and embedding distortion due to dimensionality reduction of the contextual space of language models [Xypolopoulos et al., 2021].

We operationalize polysemy of a word form as a quantity influenced by its contextual semantic neighbors and its syntactic role in a syntactic network. In particular, we construct a contextual nearest-neighbor graph of lexical units using a pretrained language model like BERT [Devlin et al., 2019]. We leverage the discrete Ricci curvature [Ni et al., 2015] measure defined on graph edges as an indicator

[3] In the context of this paper, we use ambiguity and polysemy of a word form interchangeably.

of ambiguity of a word form. The Ricci curvature can be used to determine edge roles like bridge, cliques, etc. in a graph as illustrated in Figure 1. Additionally, we construct a syntactic network for the (ambiguous) word form based on the dependency trees of the randomly sampled contexts in which the word has occurred. This network acts as another linguistic signal guiding the polysemy measure. We rely on the ability of pretrained language models to distinguish between word senses [Garí Soler and Apidianaki, 2021] and the power of graph entropy methods to identify syntactic importance of word forms in the syntactic network.

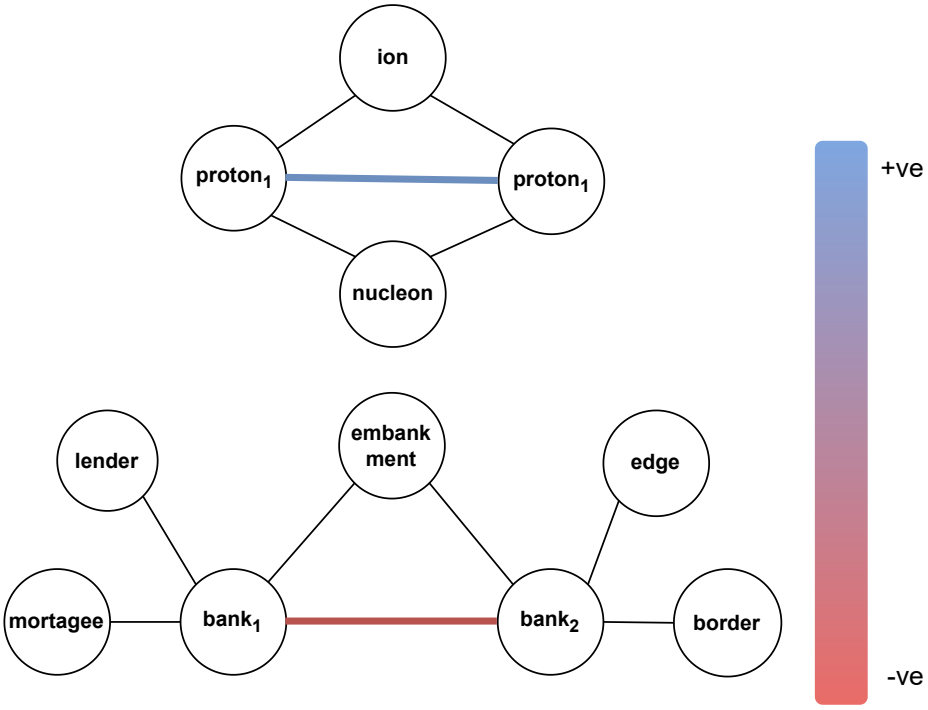

Figure 3.1 An illustrative example of Ricci curvature. The red edge (more negative) acts as a bridge connecting two distinct neighborhoods (distinct senses of the word *bank*) while the blue edge (more positive) is an edge within the same neighborhood (sense cluster of the monosemous word *proton*).

3.2 Preliminaries

3.2.1 Notations

Given a set of vertices $\mathcal{V}$ and set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, an undirected graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. For each node $v \in \mathcal{V}$, $\mathcal{N}(v)$ denotes the set of its 1-hop neighbors and $k_v = |\mathcal{N}(v)|$ denotes its degree.

3.2.2 Ricci Curvature

Traditionally, curvature is the geometric characteristic that measures how flat or curved an object is. The discrete Ollivier Ricci curvature [Ni et al., 2015] is the coarse graph generalization of curvature measures usually defined on smooth surfaces or manifolds. For $u, v \in \mathcal{V}$, m_u and m_v are probability measures of total value (mass in geometric terms) 1 each centered at u and v respectively. The Wasserstein (Earth Mover) distance $W(m_u, m_v)$ finds the optimal transportation plan ξ between probability distributions m_u and m_v.

$$W(m_u, m_v) = \inf_{\xi} \int \int d(u, v) d\xi(u, v) \tag{3.1}$$

It gives a metric to measure the minimum amount of work required to transform one probability distribution into another. The Ricci curvature thus becomes

$$\kappa_{uv} = 1 - \frac{W(m_u, m_v)}{d(u, v)} \tag{3.2}$$

where $d(u, v)$ is the number of edges in the shortest path between u and v.

Based on [Lin et al., 2011], we define the probability measure on node $v \in \mathcal{V}$ with $\alpha \in [0, 1]$ as:

$$m_v^{\alpha}(v_i) = \begin{cases} \alpha & \text{if } v_i = v \\ (1\text{-}\alpha)/k_v & \text{if } v_i \in \mathcal{N}(v) \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

We use $\alpha = 0.5$ based on previous literature since it assigns equal weights to the node and its neighbors. The edge Ricci curvature acts as an indicator of the importance and structural role of an edge in a graph. It is representative of the intrinsic geometry and local topology of the edges in a graph. This property of the discrete Ricci curvature has been used to analyze the geometry of the Internet topology [Ni et al., 2015], Graph Neural Networks [Luo et al., 2021] and community detection [Sia et al., 2019].

3.3 Proposed Approach

In this section, we introduce our proposed approach to quantify polysemy as illustrated in Figure 3.2.

3.3.1 Semantic Module

Given a word w and its list of sentences (contexts where it occurs), $S = \{s_1, s_2, \cdots, s_k\}$, we consider each instance of the word w in its corresponding sentence s_i as a separate lemma w_i. For example, if we have two contexts for the word *bank*: 1) I went to the *bank1* to deposit money, and 2) Flowers grow along the river *bank2*, we consider *bank1* and *bank2* as two separate lemmas during the graph construction.

We add each instance w_i of the word as a node to a graph $\mathcal{G}$ and pass each sentence $s_i \in S$ through a lexical substitution system [Arefyev et al., 2020] to retrieve the top contextual neighbors C_k of the

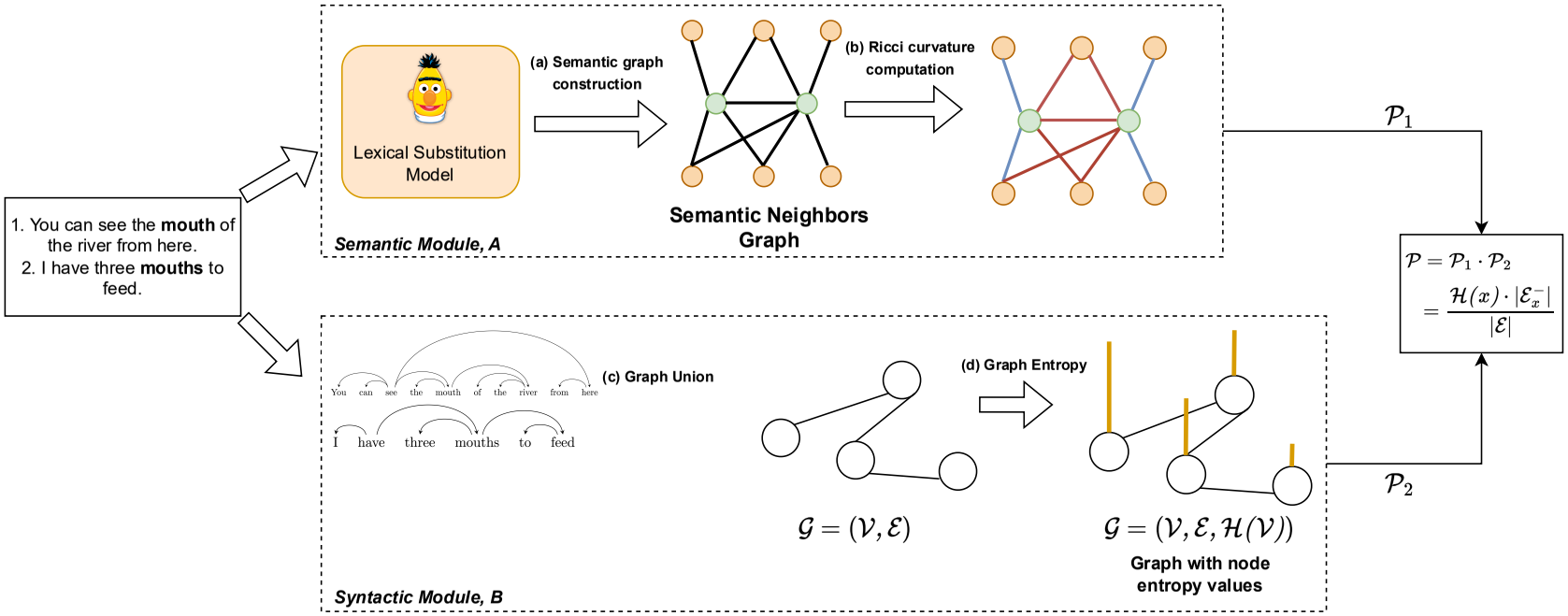

Figure 3.2 Proposed Approach for Polysemy Quantification. The set of contexts of the polysemous word is passed through: the Semantic Module (A) and the Syntactic Module (B). In the semantic module, (a) a contextual semantic graph is constructed with the help of a lexical substitution model and (b) Ricci curvature is computed on the graph edges. In the syntactic module, (c) dependency trees of the input sentences are constructed and combined in a global syntactic network using the Graph Union operator, and (d) the graph entropy is computed on the syntactic network. Both semantic (A) and syntactic (B) modules are then combined to derive the final measure of polysemy.

word w_i, adding an edge between the nearest neighbor word $c_k^i \in C_k$ and the lemma w_i. The lexical substitution model gives the most appropriate contextual replacement of a word in the input sentence, thus we can derive the semantic neighbors of a word given its context which renders the construction of the graph $\mathcal{G}$ possible.

We now efficiently compute the Ricci curvature on each edge of the graph $\mathcal{G}$ based on the linear programming method introduced by [Ni et al., 2015] and Equations 3.2 and 3.3:

$$\begin{aligned} \min \quad & \textstyle\sum_{y \in V} \sum_{x \in V} d(x,y) \rho_{xy} m_u^{\alpha}(x), \\ \text{s.t.} \quad & 0 \leq \rho_{xy} \leq 1 \quad \forall x, y \in V, \\ & \textstyle\sum_{y \in V} \rho_{xy} = 1 \quad \forall x \in V, \\ & \textstyle\sum_{x \in V} \rho_{xy} m_u^{\alpha}(x) = m_v^{\alpha}(y) \quad \forall y \in V, \end{aligned} \tag{3.4}$$

where ρ is the transportation plan matrix.

For the graph $\mathcal{G}$, we now have the edge feature matrix $E \in \mathbb{R}^{\mathcal{E} \times 1}$. Based on the intuition that negative edges act as bridge across clusters, we hypothesize that negatively curved edges connect distinct senses of the same word w. We derive the negative edges normalized by total edges in the graph as:

$$\mathcal{P}_1 = \frac{|E^-|}{|E|} \tag{3.5}$$

where $|E^-|$ is the number of negative edges in the graph. This formulation describes the variation of the curved edges in the graph. While we describe here a ratio-based definition of $\mathcal{P}_1$, it can also be operationalised as the variation of edge weights in the graph, with similar results.

3.3.2 Syntactic Module

For the given word w and its list of contexts, we derive the syntactic dependency trees of each sentence $s_i \in S$. Note that, here we do not make any distinctions between the instances of the word w unlike in the case of the Semantic Module. The obtained dependency trees are converted to their corresponding adjacency matrix A with $A_{ij} = 1$ if there is a dependency relation between tokens i and j. Each adjacency matrix corresponding to each sentence s_i can be converted to an unweighted, undirected graph D_i.

We then construct a single, global syntactic graph $\mathcal{D} = \{D_1 \bigcup D_2 \cdots \bigcup D_k\}$ where $\bigcup$ is the graph union operator, i.e., for two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, $G_1 \bigcup G_2 = (V_1 \bigcup V_2, E_1 \bigcup E_2)$. The global syntactic graph $\mathcal{D}$ contains tokens as nodes and edges as syntactic dependencies between the tokens. It thus represents the syntactic relations between input word w and the tokens of contexts. Based on ideas proposed by previous work [Čech et al., 2017] that syntactic relations influence polysemy of words across languages, we utilize the relations encoded in this graph as a signal to our polysemy measure. Inspired by recent advances in graph signal processing [Wijesinghe et al., 2021, Nouranizadeh et al., 2021, Luo et al., 2021], we compute the node entropy of the word w to quantify the importance of a node as a function of its structure.

The adjacency matrix A of the global graph $\mathcal{D} = (V, E)$ contains first order links of the graph. We define $A^2 = A^T A$ to study second order links. D represents the degree vector of the graph. We define D_r as the normalized degree vector which contains information about first and second order links.

$$D_r = D^T A_r^2 \tag{3.6}$$

Here A_r^2 is the normalized second order adjacency matrix defined as,

$$A_r^2[i,j] = \frac{A^2[i,j]}{\sum_j A^2[i,j]} \tag{3.7}$$

Here, $A^2[i,j]$ is the i-th row and $j-th$ column of the second order adjacency matrix.

Following principles of information theory, the entropy of a node, x is thus defined as

$$\begin{aligned} \mathcal{P}_2 = H(x) &= -P_x \log P_x \\ &= -\frac{D_r[x]}{\sum_x D_r[x]} \log \frac{D_r[x]}{\sum_x D_r[x]} \end{aligned} \tag{3.8}$$

3.3.3 Polysemy Quantification

To derive the quantification for polysemy, we combine Equations 3.5 and 3.8 as:

$$\begin{aligned} \mathcal{P} &= \mathcal{P}_1 \cdot \mathcal{P}_2 \\ \mathcal{P} &= \frac{H(x) \cdot |E^-|}{|E|} \end{aligned} \tag{3.9}$$

We thus derive the final measure of polysemy as desrcibed in Equation 3.9. This operationalization of polysemy in a graph-based measure incorporates syntactic signals as well as semantic structural variation.

3.4 Experiments

In this section, we first describe the data used in the current study (§3.4.1) followed by a description of the flow of the proposed approach (§3.4.2). Next we describe the evaluation metrics used (§3.4.3) and the implementation details of the current study (§3.4.4). Finally, we discuss the results of the proposed approach (§3.4.5) and perform an ablation study of investigating the individual contribution of semantics and syntax towards polysemy (§3.4.6).

3.4.1 Data

We utilize the data introduced by [Garí Soler and Apidianaki, 2021]. Sentences were sampled from SemCor 3.0 [Miller et al., 1993] dataset controlling for sense distributions in polysemous words that occur at least ten times in the corpus. For each polysemous word, we have 2 sets of sentences:

- **Random senses (poly-rand)**: Randomly sampling 10 sentences which captures the natural distribution of the senses of a word.
- **Balanced senses (poly-bal)**: 10 sentences of the word containing distinct senses. This is a controlled setting where the variation in the senses of the word is maximized.

A sample of 15 words considered in the English experiments are:

```
'exceed', 'blow', 'accord', 'identify', 'build', 'national',
'popular', 'flow', 'emotional', 'check', 'maximum', 'ability',
'west', 'instant', 'arise'
```

The original English dataset is composed of 836 polysemous words, and their corresponding 8,195 unique sentences. For French and Spanish, the sentences are taken from the Eurosense corpus [Delli Bovi et al., 2017] which contains texts from Europarl automatically annotated with BabelNet word senses [Navigli and Ponzetto, 2012]. In the multilingual corpus, we have 418 polysemous words.

We use the Frequency and Random baselines as described by [Xypolopoulos et al., 2021]. In the frequency baseline, words are ranked in decreasing order of their frequency in the Wikipedia dump. The random baseline assigns scores by sampling from a Log Normal distribution.

3.4.2 Setup

We pass each sentence in the sentence pool (poly-bal or poly-rand) through the semantic module (§3.3.1) to get a contextual nearest neighbor graph and compute $\mathcal{P}_1$ (Equation 3.5) via the Ricci curvature. Parallel to this, the sentences are also passed through the syntactic module (§3.3.2) to build a global syntactic network to compute $\mathcal{P}_2$ (Equation 3.8). Finally, based on Equation 3.9, we compute the polysemy score for the input word.

3.4.3 Evaluation

Following previous literature in polysemy quantification [Xypolopoulos et al., 2021], we utilised Spearman correlation as our evaluation metric.[4] We also perform significance tests of the correlation across all languages tested.

3.4.4 Implementation Details

We use the Stanford Stanza library [Qi et al., 2020] to build the dependency trees of sentences. [5] We use the author's implementation of LexSubGen [Arefyev et al., 2020] as the lexical substitution module

[4] We use Spearman correlation as evaluation metric due its use in previous studies [Xypolopoulos et al., 2021] which can enable fair comparison. Additionally, Spearman correlation is robust to outliers and non-linear associations in the data.

[5] https://stanfordnlp.github.io/stanza/

Method	poly-bal	poly-rand
Random	0.11	0.15
Frequency	0.18	0.20
[Garí Soler and Apidianaki, 2021]	0.29	0.32
D2L8	0.30	0.27
Ours	**0.62**	**0.60**

Table 3.1 Spearman correlation of WordNet senses and polysemy scores on English data. Our approach improves the correlation by 0.3 points over D2L8. Numbers in bold are statistically significant ($p < 0.05$)

in our framework.[6] To compute the Ricci curvature on graphs, we used the implementation based on [Ni et al., 2015].[7] All other code is written in PyTorch and uses Huggingface Transformers library [Wolf et al., 2020].

We use language-specific models for each of the language tested in our study. For English we use the state-of-the-art Lexical Substitution system described by [Arefyev et al., 2020]. For languages other than English, we rely on the Masked Language Model prediction of the model which has been shown to be effective for lexical substitution by [Qiang et al., 2021]. We use *bert-base-uncased* [Devlin et al., 2019] for English, *flaubert-base-uncased* [Le et al., 2020] for French and *bert-base-spanish-wwm-uncased* [Caete et al., 2020]. We compare our results with the model based on dimensionality reduction and multiresolution grids on the reported hyperparameters proposed by [Xypolopoulos et al., 2021].

3.4.5 Results

In this section, we discuss the results of the proposed quantification measure. We assume the number of senses of a word in the WordNet is a good representative of the ambiguity it possesses [Pimentel et al., 2020] and calculate its correlation with our proposed metric. Prior work like [Xypolopoulos et al., 2021] have used WordNet as ground truth and empirically demonstrated that WordNet, WordNet-reduced and domain-specific WordNet all produce highly similar polysemy rankings despite the different sense granularities they have. Hence we report our results on the classic WordNet data. Henceforth, we refer to the approach proposed by [Xypolopoulos et al., 2021] as D2L8.

In Table 3.1, we observe that our measure shows higher significant correlations with the WordNet rankings on English data. For poly-rand setting, where the natural sense distribution of a word is captured, we observe an increment of 0.3 points in the correlation as compared to the D2L8 baseline which is based on the notion of multiresolution grids where volume is approximated hierarchical discretization

[6] https://github.com/Samsung/LexSubGen
[7] https://github.com/saibalmars/GraphRicciCurvature

of the embedding space [Nikolentzos et al., 2017]. The poly-bal data is a controlled setting where the number of contexts is balanced. Although the baseline was described to work on randomly sampled sentences in English, we apply it to the controlled setting where it achieves a much better correlation of 0.3 and comparable to ours.

	French		Spanish	
	D2L8	Ours	D2L8	Ours
poly-bal	0.48	0.45	0.48	0.62
poly-rand	0.19	0.43	0.14	0.20

Table 3.2 Spearman correlation of the proposed polysemy quantification with WordNet number of senses across different languages.

We apply our measure in a cross-lingual setting to measure polysemy across 2 diverse languages other than English - French and Spanish. We also extend the baseline D2L8 to our cross-lingual setting.[8] Table 3.2 reports the Spearman correlations of the number of senses of a word in the Multilingual WordNet [Bond and Paik, 2012] of the language with our proposed quantification. We observe significant correlations across all languages and all settings (poly-bal and poly-rand). The poly-bal data setting shows consistently strong correlations as compared to poly-rand setting which is quite intuitive due to the carefully controlled sense distribution in poly-bal sentences. We note here that since we only take 10 sentences in each context pool (poly-bal and poly-rand), it is a highly constrainted setting as compared to previous works [Xypolopoulos et al., 2021, Pimentel et al., 2020] which randomly sampled greater than 10,000 sentences for each word. Our motivation behind taking this constrained approach is to enable our method to perform even for low-resource languages.

3.4.6 Ablation study

We perform an ablation study in order to investigate the individual contribution of Semantic and Syntactic Module. In Table 3.3, we report the Spearman correlations of polysemy measure taken from each module with the English WordNet rankings.

We observe that both semantic and syntactic module are positively correlated with the number of senses a word possesses. This result validates previous findings linking syntax and polysemy [Čech et al., 2017]. These results suggest that studies in ambiguity should investigate syntax along with semantics of an utterance.

[8] https://github.com/ksipos/polysemy-assessment

	Syntax Module	Semantic Module
poly-bal	0.28	0.33
poly-rand	0.48	0.46

Table 3.3 Spearman correlation of individual measures from syntax and semantic modules with English WordNet ground truth rankings.

3.4.7 Error Analysis

Since we rely on a lexical substitution module [Arefyev et al., 2020], the errors in this model might propagate into the final score. For example, in some cases, the substitution model fails to generate enough number of word substitutions given the context, thus resulting in a sparse graph where Ricci curvature might not be a good metric to compute polysemy. In some cases, the model also generates variations of the same semantic word, *home* and *homes*, which can further reduce the important signals required for the model to compute a good polysemy score.

3.5 Discussion

Since our method aims to quantify the tendency of a word to have more meanings, words assigned higher values are assumed to be more polysemous. While this operationalisation does not explicitly allow for the discovery of new polysemy relations, we observe, for example, that accord (6 ground truth senses) is assigned a higher polysemy score relative to a word like maximum (4 ground truth sense). This case is interesting since WordNet provides very similar ratings for both while our method accentuates the difference between the two, intuitively, giving accord much higher score.

3.5.1 Conclusion

In this study, based on previous linguistic evidence, we posit that including syntactic information in the form of dependency structural knowledge can help in the quantification of lexical ambiguity or polysemy of a wordform. To investigate this, we propose a simple operationalization of polysemy based on the Ricci curvature of the contextual nearest neighbors graph of a word and the entropy of its combined syntactic network.

We show that our proposed measure shows high correlations with number of word senses in WordNet across multiple languages. Our approach is fully unsupervised, simple and grounded in previously established linguistic theories. We hope that similar graph-based approaches can help in creation and validation of sense inventories across languages.

3.5.2 Limitations

Our work acts as a proxy for the ambiguity of a word form and the scores are continuous but it does not quantify the discrete counts of the senses of a word. We rely on the availability of good quality language-specific language models which can be used as the lexical substitution model in the Semantic Module. Any errors in the language model may propagate into our score.

We tested our framework on sentences sampled from the SemCor 3.0 dataset which is a good resource for sense analysis in NLP but is naturally limited to sentences in formal English. A lack of diversely sourced corpora for a study in polysemy may limit the generalizability of a quantification measure to other domains.

3.5.3 Future Work

We leave the utility of polysemy quantification to improve extrinsic tasks Word Sense Disambiguation or Word In Context for future work. We hope that works in polysemy quantification also lead to interesting linguistic analyses about the nature of ambiguity in natural languages and the relationship between morpho-syntactic information like Part-Of-Speech Tags, dependency relations and thematic information with polysemy scores of word units.

Chapter *4*

Investigating Pretrained Language Models and Colloquial Tautologies

The reader I seek is a tautology, for he/she is simply the person who wants to read what I have written.

Will Self, *The Guardian, 2007*[1]

4.1 Motivation

The phrase *Boys will be boys* is an example of a **colloquial tautology**, a type of statement that is frequently used in everyday language, literature, and advertising. Although such phrases may seem nonsensical at first glance, they are easily understandable to human speakers and do not appear to be redundant. In this study, we are interested in examining colloquial tautologies like *Boys will be boys* or *A promise is a promise* and exploring the factors that influence their interpretation by Pretrained Language Models (PLMs) such as BERT [Devlin et al., 2019] and GPT [Radford et al., 2019]. We also aim to uncover why some tautological phrases are acceptable to PLMs while others are less acceptable.

While linguists and philosophers have previously examined the meaning and interpretation of tautological statements, there has been little research on how PLMs process and understand them. Our work seeks to fill this gap by investigating the pragmatic, semantic, and syntactic factors that impact the interpretation of colloquial tautologies by PLMs. Specifically, we use perplexity scores as a proxy for acceptability of colloquial tautologies by PLMs. By understanding how these models process and interpret such statements, we hope to gain insights into the capabilities and limitations of these increasingly important tools in natural language processing.

[1] https://www.theguardian.com/books/2007/may/09/willself

4.2 Theoretical Approaches to Tautologies

We ground our experiments and contrast our results using two theoretical models of interpreting nominal tautologies, viz., pragmatic and semantic views.

4.2.1 The Gricean Maxims

[Grice, 1975] proposed a pragmatic model on how listeners and speakers communicate and cooperate in conversations. The Gricean view proposes that most of the information conveyed by a speaker is implied rather than explicitly asserted. Hence, *Boys will be boys* does not explicitly state any new information but implies something about the nature of boys which is expected to be realized by the listener. Grice proposed four **maxims of conversation.**

Gricean Maxims are a set of principles that describe how people use language to communicate with each other effectively. They are considered to be a fundamental concept in the field of pragmatics. There are four Gricean Maxims: the Maxim of Quantity, the Maxim of Quality, the Maxim of Relation, and the Maxim of Manner. Each of these maxims outlines a principle that people follow when communicating with each other.

4.2.1.1 The Maxim of Quantity

It refers to the idea that speakers should give as much information as necessary, but no more. Speakers should provide enough information to convey their message clearly, but should not provide unnecessary or irrelevant information that could confuse or distract the listener.

4.2.1.2 The Maxim of Quality

It refers to the idea that speakers should be truthful and should provide accurate information. Speakers should avoid making false or unsupported claims, and should provide evidence or support for their statements if necessary.

4.2.1.3 The Maxim of Relation

It refers to the idea that speakers should stay on topic and should provide information that is relevant to the conversation. Speakers should avoid introducing unrelated or tangential information that could distract or confuse the listener.

4.2.1.4 The Maxim of Manner

It refers to the idea that speakers should be clear and concise in their communication. Speakers should avoid using ambiguous or overly complex language, and should strive to make their message as easy to understand as possible.

While the Gricean Maxims are not considered to be absolute rules that speakers must follow, they are seen as useful guidelines for effective communication. By following these maxims, speakers can communicate more clearly and avoid misunderstandings or confusion. However, it is important to note that the Gricean Maxims can also be violated deliberately or unintentionally. Speakers may choose to withhold information or provide false information, or they may use ambiguous language for rhetorical effect. Additionally, cultural and linguistic differences can also affect how the Gricean Maxims are interpreted and applied in different contexts. Even though *Boys will be boys* conflicts with the Maxim of Quantity and the Maxim of Manner and despite this apparent lack of cooperation, a listener will consider the speaker to be cooperative in the conversation.

Remark. *The Pragmatic View or the Gricean view suggests that the interpretation of nominal tautologies is context-dependent. Same tautology can take on different meanings depending on the conversational context.*

4.2.2 The Semantic View

Critics of the Pragmatic View argue that the interpretation of tautologies is not solely based on their pragmatic implications, but rather also on the syntactic patterns and nominal classifications of the phrases [Wierzbicka, 1987]. For example, tautologies of the form *N is N* (e.g., "War is war") are typically abstract-singular and convey a negative mood. In contrast, tautologies of the form *N will be N* generally convey negative aspects of the topic with an indulgent undertone. The Semantic View to nominal tautologies supports the idea that the syntactic form of a phrase contributes to its overall meaning. In other words, the way that the words are arranged in a sentence can impact the interpretation of the tautology. For example, the order of the words in the phrase *"Boys will be boys"* contributes to the indulgent undertone towards the negative aspects of the behavior being referred to. Furthermore, the classification of the nominal phrase can also influence the interpretation of the tautology. A nominal phrase that refers to an abstract concept, such as "war," may have a different connotation than a phrase that refers to a concrete object, such as "rock." This is because abstract concepts may have more emotional or ideological associations than concrete objects.

Overall, these factors suggest that the interpretation of tautologies is not solely based on their pragmatic implications, but rather on a complex interplay between syntactic form, nominal classification, and pragmatic context. Understanding these factors is crucial for accurately interpreting and generating natural language using tools such as Pretrained Language Models (PLMs).

Remark. *The Semantic View to nominal tautologies suggests that syntactic form of phrases contribute semantic information to the interpretation of tautologies.*

4.3 Language Model Scores

Sequence log probability scores are a measure of how likely a sequence of words is according to a transformer-based language model. For autoregressive models like GPT, this score is calculated by multiplying the conditional probability of each token in the sequence given all the previous tokens. The sequence log probability score is the sum of the logarithm of these conditional probabilities. This score can be used to evaluate the acceptability of texts given a language model [Misra, 2022].

Consider a sentence S of length $|S|$ with tokens $w_0, w_1, \ldots, w_{|S|}$ and a given Language Model with pretrained parameters θ.

4.3.1 Autoregressive Language Models

The formula for sequence log probability for autoregressive models like GPT is:

$$\log p(S) = \sum_{i=1}^{|S|} \log P(w_i | w_{1:i-1}; \theta) \tag{4.1}$$

, where $\log P$ is the conditional probability of the current token w_i given the previous context $w_1 : w_{i-1}$.

4.3.2 Masked Language Models

For Masked Language Models (MLMs), since true log probabilities are not tractable, we use Pseudo-Log Probability (PLL) scores [Salazar et al., 2019].

$$PLL(S) = \sum_{i=1}^{|S|} \log P_{MLM}(w_i | S_{\backslash w_i}; \theta) \tag{4.2}$$

, where $\log P_{MLM}$ is the Pseudo-Log Probability of a masked token w_i given the remaining context $S_{\backslash w_i}$.

4.4 Data

To create a dataset for understanding tautologies, we followed the methodology of [Gibbs and McCarrell, 1990] by selecting 12 nouns each from three noun types: human role, concrete objects, and abstract concepts. The data was constructed using recent neural approaches [Yoo et al., 2021] on text generation and augmentation for building high-quality datasets. Specifically, we use few-shot prompting to synthetically generate contexts using GPT-3.[2] This resulted in a total of 36 nouns. Nominal tautologies were constructed using these nouns based on three syntactical forms: singular, plural, and

[2]The prompts are written by using positive and negative contexts for a noun from the dataset by [Gibbs and McCarrell, 1990] as in-context samples. The model is then prompted to generate contexts for a new noun (which is the subject of our study).

modal.[3] As a result, we generated 108 systematically generated nominal tautologies. To further extend the dataset, we generated two short contexts for each of the 36 nouns. One context was positive, while the other was negative. The length of each context was roughly the same. This resulted in a total of 216 sentences (36 nouns x 3 syntactic forms x 2 contexts). Table 4.1 shows sample sentences from the dataset used in this study.

The contexts are used in conjunction with the tautologies. We append a tautology (it could be of either form out of singular, plural or modal) to the negative/positive context and probe the pretrained language models. This experiment is again motivated by [Gibbs and McCarrell, 1990] and allows us to investigate whether pretrained models have learnt negative associations during pretraining.

4.5 Results

4.5.1 Acceptability of Colloquial Tautologies without context

The two tables (Table 4.2 and 4.3) show the log likelihood scores of colloquial tautologies without context measured by two different pretrained language models: BERT and GPT2. The scores are presented in three different noun types (human, concrete, and abstract) and three different syntactic forms (singular, plural, and modal). By comparing the two tables, we can infer that the acceptability scores of the tautologies measured by GPT2 are generally higher than those measured by BERT, regardless of noun type and syntactic form except for the case of plural tautologies where BERT outperforms GPT2 scores consistently. This suggests that GPT2 performs better in handling the acceptability of colloquial tautologies than BERT in general while BERT is better than GPT2 for plural tautologies across all noun types. Moreover, we can observe that the noun type and syntactic form have an impact on the acceptability scores of the tautologies. For instance, plural syntactic form generally receives higher scores than singular form with concrete plural having a high spike in scores, and human nouns receive higher scores than concrete and abstract nouns, which may indicate that the acceptability of tautologies is influenced by both syntactic and semantic factors. Surprisingly, LLMs seem to prefer plural tautological constructions, contrary to previous literature on humans' preference for modal forms [Gibbs and McCarrell, 1990].

4.5.2 Acceptability of Colloquial Tautologies with context

The two tables (Table 4.4 and 4.5) present log-likelihood scores for Colloquial Tautologies with different noun types and syntactic forms, in various contexts. The likelihood scores are averaged over the sequence length to account for the larger length of contexts in this setting. The first table presents the scores from the BERT model, while the second table presents the scores from the GPT2 model. The tables suggest that the scores for the different noun types vary depending on the syntactic form and

[3] We borrow the terminology to describe morphological forms as syntactic forms inspired by [Gibbs and McCarrell, 1990]

	Noun Type		
Syntactic Form	**Human**	**Concrete**	**Abstract**
Singular	A politician is a politician.	A diamond is a diamond.	A war is a war.
Plural	Politicians are politicians.	Diamonds are diamonds.	Wars are wars.
Modal	Politicians will be politicians.	Diamonds will be diamonds.	Wars will be wars.
Context			
Positive	Politicians have the power to make a real difference in people's lives. They can create laws and policies that improve our communities and protect our rights.	Diamonds are a symbol of love and commitment, and the perfect way to celebrate important milestones. They are also incredibly valuable.	Wars can bring about social and political reforms and technological advances. Wars have led to the overthrow of oppressive regimes and the establishment of more democratic societies.
Negative	Politicians are notorious for being dishonest and corrupt. They will say anything to get elected and then break their promises once they're in office.	The mining of diamonds can have devastating effects on the environment and can be tied to human rights abuses. The high cost of diamonds creates a culture of materialism and consumerism.	Wars lead to widespread suffering, displacement, and trauma, and can result in the loss of millions of lives. The aftermath of war often leads to lasting divisions and resentment between nations.

Table 4.1 Constructed dataset sample for each category of noun types, syntactic form and context.

context. In general, the scores are higher for negative contexts and for plural and modal syntactic forms. This suggests that models encode negative factual connotations for tautological constructions, similar to human behaviour [Gibbs and McCarrell, 1990]. However, Table 4.4 shows higher scores for positive contexts within concrete and abstract nouns. This suggests that BERT's behaviour for non-human nouns is preferentially encoded to positive stereotypes. The human noun type tends to have the highest scores,

	Noun Type		
Syntactic Form	Human	Concrete	Abstract
Singular	2.4	2.7	2.4
Plural	9.5	8.4	7.5
Modal	5.9	6.0	5.5

Table 4.2 BERT Log Likelihood score of Colloquial Tautologies without context. Human nouns in the plural syntactic form have the highest acceptability.

	Noun Type		
Syntactic Form	Human	Concrete	Abstract
Singular	6.0	5.8	5.9
Plural	7.5	7.8	6.1
Modal	7.9	7.2	6.4

Table 4.3 GPT2 Log Likelihood score of Colloquial Tautologies without context. Abstract nouns with plural syntactic form have the highest acceptability.

	Noun Type					
	Human		**Concrete**		**Abstract**	
Context	**+ve**	**-ve**	**+ve**	**-ve**	**+ve**	**-ve**
Syntactic Form						
Singular	1.26	1.79	1.0	0.91	0.81	0.79
Plural	1.29	1.87	1.01	1.0	0.85	0.8
Modal	1.39	1.97	1.2	1.1	0.95	0.82

Table 4.4 BERT Log Likelihood score of Colloquial Tautologies with context. Human nouns in the modal form with negative context have high acceptability.

followed by concrete and then abstract noun types. Additionally, there are differences between the BERT and GPT2 models. The GPT2 model tends to produce higher scores overall, particularly for negative contexts and abstract noun types. In general, we observe that negative contexts are accepted higher than positive ones, in agreement with the pragmatic approach to tautologies.

	Noun Type					
	Human		Concrete		Abstract	
Context	**+ve**	**-ve**	**+ve**	**-ve**	**+ve**	**-ve**
Syntactic Form						
Singular	2.95	3.14	3.1	3.3	2.95	3.0
Plural	3.0	3.2	3.15	3.3	2.98	3.06
Modal	3.1	3.2	3.2	3.4	3.02	3.09

Table 4.5 GPT2 Log Likelihood score of Colloquial Tautologies with context. Negative contexts for human and concrete nouns have higher acceptability.

Overall, these inferences suggest that the choice of model, syntactic form, and noun type can have a significant impact on the performance of a PLM when dealing with Colloquial Tautologies.

4.6 Discussion

The results suggest that there are two distinct views on how to interpret tautologies: the semantic view and the pragmatic view. The first experiment supports the semantic view, which holds that the meaning of a statement is determined by the logical relationships between its constituent terms. In contrast, the second experiment supports the pragmatic view, which holds that meaning is determined by the context in which the statement is used and the intentions of the speaker.

However, the statement also suggests that the two views are not mutually exclusive and that there may be interactions between them. The implication is that a syncretic approach, which takes into account both semantic and pragmatic factors, may be necessary to fully understand and interpret tautologies.

The implications of this are significant for future research, as it suggests that a more nuanced and complex understanding of language is required. A syncretic approach may require researchers to take into account a wider range of factors, including context, intention, and cultural background. It may also require the development of new methods for analyzing language, such as natural language processing algorithms that can recognize and account for pragmatic factors. Our results suggest that there is no simple or straightforward way to interpret tautologies using PLMs, and that a more nuanced and integrated approach is needed to fully understand the complexities of language and meaning.

4.6.1 Limitations

- **Lack of transparency:** PLMs are complex models that are difficult to interpret. While log likelihood scores can provide a numerical representation of sentence acceptability, it is not always

clear how these scores are generated or what factors contribute to them. As a result, it can be challenging to understand why a particular sentence receives a high or low score.

- **Overreliance on the training data:** The performance of PLMs is heavily influenced by the training data they are exposed to. Therefore, using log likelihood scores from a PLM trained on a particular dataset may not accurately reflect the acceptability of sentences that are not represented in that training data. This is especially true for rare or uncommon constructions, which may not be well-represented in the training data and therefore not accurately captured by the PLM.

- **Domain-specificity:** PLMs are typically trained on large, general-purpose datasets, which may not reflect the language usage in specific domains. Therefore, using log likelihood scores from a general-purpose PLM to measure sentence acceptability in a domain-specific context may not be appropriate.

- **Sensitivity to noise:** Log likelihood scores are sensitive to noise, including typographical errors, misspellings, and other errors in the input text. Therefore, small variations in the input text can significantly impact the log likelihood score and thus the estimated sentence acceptability.

- **Lack of correlation with human judgments:** While PLMs can generate log likelihood scores quickly and efficiently, these scores do not always correlate well with human judgments of sentence acceptability. In some cases, PLMs may give high scores to sentences that are considered unacceptable by human judges, and vice versa.

4.6.2 Future Works

- **Investigating the impact of context length and complexity:** The current work used two short contexts of similar length for each noun in the dataset. Future work could explore the effect of context length and complexity on the acceptability of tautologies by LLMs. This could involve using longer or more complex contexts and analyzing the resulting LLM predictions.

- **Comparing different LLMs:** The current paper used only two LLMs (GPT-2, BERT) to analyze the acceptability of tautologies. Future work could investigate the acceptability of tautologies by other LLMs, such as RoBERTa, and compare the results to those obtained with GPT-2. This could provide insights into the relative strengths and weaknesses of different LLM architectures for this task.

- **Exploring different types of tautologies:** The current work focused on nominal tautologies (i.e., tautologies that involve nouns), but tautologies can also be constructed using other parts of speech, such as adjectives or verbs. Future work could investigate the acceptability of tautologies of different types by LLMs and compare the results to those obtained with nominal tautologies.

- **Examining the effect of fine-tuning:** The current paper used a pre-trained LLM and did not fine-tune it on the tautology dataset. Future work could investigate the effect of fine-tuning on the LLM's ability to predict the acceptability of tautologies. This could involve training the LLM on the tautology dataset and comparing its performance to that of the pre-trained LLM.
- **Testing human judgments:** The current paper used LLM predictions as a proxy for human acceptability judgments. Future work could directly test human judgments of tautologies to validate the LLM's predictions. This could involve conducting human acceptability experiments using the same tautology dataset used in the current paper.

Chapter *5*

Conclusion & Future Works

Man must not attempt to dispel the ambiguity of his being but, on the contrary, accept the task of realizing it.

Simone de Beauvoir, The Ethics of Ambiguity [de Beauvoir, 1947]

This thesis focused on computational experiments involving linguistic ambiguities. The research first introduced a novel method to quantify polysemy that incorporates syntax and geometry. This was followed by an examination of how pretrained language models handle colloquial tautologies. Through a literature review, gaps in the existing research were identified and addressed in this study. Chapter 2 reviewed existing literature on ambiguity in NLP, graph-based NLP, and pragmatics. Shortcomings of these approaches were highlighted. Chapter 3 addressed the lack of syntactic information in previous polysemy quantification measures by proposing a syntax-aware and graph-theoretic framework. By incorporating dependency structures, the proposed method effectively measured polysemy across three languages. The correlation with the WordNet benchmark demonstrated the efficacy of grounding NLP models with syntactic knowledge. The use of graphs in NLP was also shown to be effective. Chapter 4 investigated the ability of pretrained language models to handle colloquial tautologies. The study used a systematic dataset to control for noun type, context, and syntactic form of the tautologies. The results revealed that BERT and GPT2 perform better with modal forms and human nouns, which aligned with previous literature and human intuition.

In conclusion, this thesis highlighted the shortcomings of existing research on linguistic ambiguities and proposed various solutions to overcome them. The incorporation of syntax and geometry into polysemy quantification is a novel contribution that demonstrates the effectiveness of syntactically motivated methods. The study on colloquial tautologies sheds light on the capability of pretrained language models in handling tautological constructions and the factors influencing them. This emphasizes the need for further investigation into this area.

5.1 Future Works

The thesis forms the bedrock for future work in analysing, interpreting and measuring linguistic ambiguities using computational tools. Some broad future directions to build upon this thesis are discussed below:

1. **Syntax-guided Controlled Generation:** An interesting application of polysemy quantification is the use of syntactic signals to guide natural language generation in order to produce less (or more) ambiguous texts.

2. **Pragmatic Alignment of LLMs:** Theories like the Gricean Maxims can be used to align the behaviour of LLMs with real-world communication pragmatics to build robust and natural dialogue agents.

Related Publications

- **An Unsupervised, Geometric and Syntax-aware Quantification of Polysemy**, *Anmol Goel*, Charu Sharma, Ponnurangam Kumaraguru. *In Conference on Empirical Methods in Natural Language Processing (EMNLP) 2022*.

Other Publications

- **SyMCoM - Syntactic Measure of Code Mixing A Study Of English-Hindi Code-Mixing**, Prashant Kodali , *Anmol Goel*, Monojit Choudhury, Manish Shrivastava, Ponnurangam Kumaraguru. *In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022*
- **HLDC: Hindi Legal Documents Corpus**, Arnav Kapoor, Mudit Dhawan, *Anmol Goel*, Arjun T H, Akshala Bhatnagar, Vibhu Agrawal, Amul Agrawal, Arnab Bhattacharya, Ponnurangam Kumaraguru, Ashutosh Modi. *In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022*

Bibliography

[Adrian et al., 2001] Adrian, A., Richard, A. D., Ann, K., and Robert, M. (2001). Linguistics: An introduction to language and communication. *United States: Massachusetts Institute of Technology*.

[Ahmad et al., 2021] Ahmad, W., Li, H., Chang, K.-W., and Mehdad, Y. (2021). Syntax-augmented multilingual BERT for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4538–4554.

[Andreu and Corominas-Murtra, 2011] Andreu, J. F. and Corominas-Murtra, B. (2011). On ambiguity. its locus in the architecture of language and its origin in efficient communication. *ArXiv*, abs/1107.0193.

[Arefyev et al., 2020] Arefyev, N., Sheludko, B., Podolskiy, A., and Panchenko, A. (2020). Always keep your target in mind: Studying semantics and improving performance of neural lexical substitution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1242–1255.

[Artstein and Poesio, 2008] Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

[Bain, 1950] Bain, R. (1950). Human behavior and the principle of least effort: An introduction to human ecology. by george kingsley zipf. cambridge, mass.: Addison-wesley press, inc., 1949. 573 pp.

[Balconi, 2010] Balconi, M. (2010). Biological basis of linguistic and communicative systems: From neurolinguistics to neuropragmatics.

[Biemann, 2006] Biemann, C. (2006). Chinese whispers-an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the first workshop on graph based methods for natural language processing*, pages 73–80.

[Bond and Paik, 2012] Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71.

[Bréal, 1904] Bréal, M. (1904). *Essai de sémantique (science des significations).* Hachette.

[Brown et al., 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[Caete et al., 2020] Caete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Prez, J. (2020). Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

[Čech et al., 2017] Čech, R., Mačutek, J., Žabokrtskỳ, Z., and Horák, A. (2017). Polysemy and synonymy in syntactic dependency networks. *Digital Scholarship in the Humanities*, 32(1):36–49.

[Christiansen and Kirby, 2003] Christiansen, M. H. and Kirby, S. (2003). Language evolution: Consensus and controversies. *Trends in cognitive sciences*, 7(7):300–307.

[Crossley et al., 2010] Crossley, S., Salsbury, T., and McNamara, D. (2010). The development of polysemy and frequency use in english second language speakers. *Language Learning*, 60(3):573–605.

[de Beauvoir, 1947] de Beauvoir, S. (1947). *The Ethics of Ambiguity*.

[Delli Bovi et al., 2017] Delli Bovi, C., Camacho-Collados, J., Raganato, A., and Navigli, R. (2017). EuroSense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 594–600.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

[Durkin and Manning, 1989] Durkin, K. and Manning, J. (1989). Polysemy and the subjective lexicon: Semantic relatedness and the salience of intraword senses. *Journal of Psycholinguistic Research*, 18:577–612.

[Eddington and Tokowicz, 2015] Eddington, C. M. and Tokowicz, N. (2015). How meaning similarity influences ambiguous word processing: The current state of the literature. *Psychonomic bulletin & review*, 22:13–37.

[Erk and McCarthy, 2009] Erk, K. and McCarthy, D. (2009). Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 440–449, Singapore. Association for Computational Linguistics.

[Ettinger, 2020] Ettinger, A. (2020). What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

[Falkum, 2015] Falkum, I. L. (2015). The how and why of polysemy: A pragmatic account. *Lingua*, 157:83–99.

[Falkum and Vicente, 2015] Falkum, I. L. and Vicente, A. (2015). Polysemy: Current perspectives and approaches.

[Farghal, 1992] Farghal, M. (1992). Colloquial jordanian arabic tautologies. *Journal of Pragmatics*, 17(3):223–240.

[Friedrich et al., 2012] Friedrich, A., Engonopoulos, N., Thater, S., and Pinkal, M. (2012). A comparison of knowledge-based algorithms for graded word sense assignment. In *Proceedings of COLING 2012: Posters*, pages 329–338, Mumbai, India. The COLING 2012 Organizing Committee.

[Fromkin et al., 2018] Fromkin, V., Rodman, R., and Hyams, N. (2018). *An Introduction to Language (w/MLA9E Updates)*. Cengage Learning.

[Garí Soler and Apidianaki, 2021] Garí Soler, A. and Apidianaki, M. (2021). Lets play mono-poly: Bert can reveal words polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

[Gibbs and McCarrell, 1990] Gibbs, R. W. and McCarrell, N. S. (1990). Why boys will be boys and girls will be girls: Understanding colloquial tautologies. *Journal of Psycholinguistic Research*, 19:125–145.

[Glynn, 2009] Glynn, D. (2009). Polysemy, syntax, and variation. a usage-based method for cognitive semantics. inv. evans & s. pourcel (eds.), new directions in cognitive linguistics (pp. 77–106).

[Grice, 1975] Grice, H. P. (1975). Logic and conversation. In *Speech acts*, pages 41–58. Brill.

[Haber and Poesio, 2021] Haber, J. and Poesio, M. (2021). Patterns of polysemy and homonymy in contextualised language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2663–2676, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Hou et al., 2021] Hou, X., Qi, P., Wang, G., Ying, R., Huang, J., He, X., and Zhou, B. (2021). Graph ensemble learning over multiple dependency trees for aspect-level sentiment classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2884–2894.

[Inglese and Brigada Villa, 2021] Inglese, G. and Brigada Villa, L. (2021). Inferring morphological complexity from syntactic dependency networks: A test. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 10–22.

[Ivanov, 2019] Ivanov, G.-B. (2019). *NLP for Hackers*. Manning.

[Kabbara, 2019] Kabbara, J. (2019). Computational investigations of pragmatic effects in natural language. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 71–76, Minneapolis, Minnesota. Association for Computational Linguistics.

[Kelih, 2008] Kelih, E. (2008). Modelling polysemy in different languages: A continuous approach. *Glottometrics*, 16(2008):46–56.

[Kelih and Altmann, 2015] Kelih, E. and Altmann, G. (2015). A continuous model for polysemy. *Glottometrics*, 31:31–37.

[Kess and Hoppe, 1978] Kess, J. F. and Hoppe, R. A. (1978). On psycholinguistic experiments in ambiguity. *Lingua*, 45(2):125–140.

[Krovetz, 1997] Krovetz, R. (1997). Homonymy and polysemy in information retrieval. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 72–79.

[Krovetz and Croft, 1992] Krovetz, R. and Croft, W. B. (1992). Lexical ambiguity and information retrieval. *ACM Trans. Inf. Syst.*, 10(2):115141.

[Kwon, 2009] Kwon, I. (2009). A tautology is a tautology: specificity and categorization in nominal tautological constructions. In *Annual Meeting of the Berkeley Linguistics Society*, volume 35, pages 211–222.

[Kwon, 2014] Kwon, I. (2014). Categorization and its embodiment: Korean tautological constructions in mental spaces theory. *Language Sciences*, 45:44–55.

[Lacerra et al., 2021] Lacerra, C., Tripodi, R., and Navigli, R. (2021). GeneSis: A Generative Approach to Substitutes in Context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10810–10823.

[Le et al., 2020] Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT: Unsupervised language model pretraining for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490.

[Lin et al., 2011] Lin, Y., Lu, L., and Yau, S.-T. (2011). Ricci curvature of graphs. *Tohoku Mathematical Journal, Second Series*, 63(4):605–627.

[Liu et al., 2023] Liu, A., Wu, Z., Michael, J., Suhr, A., West, P., Koller, A., Swayamdipta, S., Smith, N. A., and Choi, Y. (2023). We're afraid language models aren't modeling ambiguity. *arXiv preprint arXiv:2304.14399*.

[Luo et al., 2021] Luo, G., Li, J., Su, J., Peng, H., Yang, C., Sun, L., Yu, P. S., and He, L. (2021). Graph entropy guided node embedding dimension selection for graph neural networks. *arXiv preprint arXiv:2105.03178*.

[MacDonald et al., 1994] MacDonald, M., Pearlmutter, N., and Seidenberg, M. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological review*, 101:676–703.

[MacWhinney, 2005] MacWhinney, B. (2005). Language evolution and human development. *Origins of the social mind: Evolutionary psychology and child development*, pages 383–410.

[Marcheggiani and Titov, 2020] Marcheggiani, D. and Titov, I. (2020). Graph convolutions over constituent trees for syntax-aware semantic role labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3915–3928.

[Melamud et al., 2016] Melamud, O., Goldberger, J., and Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61.

[Mercuri, 2012] Mercuri, S. P. (2012). Understanding the interconnectedness between language choices, cultural identity construction and school practices in the life of a latina educator. *Gist: Education and learning Research journal*, (6):12–43.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

[Miller et al., 1993] Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

[Misra, 2022] Misra, K. (2022). minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

[Misra et al., 2020] Misra, K., Ettinger, A., and Rayz, J. T. (2020). Exploring bert's sensitivity to lexical cues using tests from semantic priming. *arXiv preprint arXiv:2010.03010*.

[Mitra et al., 2014] Mitra, S., Mitra, R., Riedl, M., Biemann, C., Mukherjee, A., and Goyal, P. (2014). That's sick dude!: Automatic identification of word sense change across different timescales. *arXiv preprint arXiv:1405.4392*.

[Mohammad, 2018] Mohammad, H. M. (2018). The nature of ambiguity across languages. *International Journal for Innovation Education and Research*, 6(12):149157.

[Murphy, 1997] Murphy, G. L. (1997). Polysemy and the creation of novel word meanings.

[Navigli and Ponzetto, 2012] Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

[Nerlich and Clarke, 2001] Nerlich, B. and Clarke, D. D. (2001). Ambiguities we live by: Towards a pragmatics of polysemy. *Journal of Pragmatics*, 33(1):1–20.

[Ni et al., 2015] Ni, C.-C., Lin, Y.-Y., Gao, J., Gu, X. D., and Saucan, E. (2015). Ricci curvature of the internet topology. In *2015 IEEE conference on computer communications (INFOCOM)*, pages 2758–2766. IEEE.

[Nikolentzos et al., 2017] Nikolentzos, G., Meladianos, P., and Vazirgiannis, M. (2017). Matching node embeddings for graph similarity. In *Thirty-first AAAI conference on artificial intelligence*.

[Nouranizadeh et al., 2021] Nouranizadeh, A., Matinkia, M., Rahmati, M., and Safabakhsh, R. (2021). Maximum entropy weighted independent set pooling for graph neural networks. *arXiv preprint arXiv:2107.01410*.

[OpenAI, 2023] OpenAI (2023). Gpt-4 technical report. *ArXiv*, abs/2303.08774.

[Ortega-Martín et al., 2023] Ortega-Martín, M., García-Sierra, Ó., Ardoiz, A., Álvarez, J., Armenteros, J. C., and Alonso, A. (2023). Linguistic ambiguity analysis in chatgpt. *arXiv preprint arXiv:2302.06426*.

[Pandia et al., 2021] Pandia, L., Cong, Y., and Ettinger, A. (2021). Pragmatic competence of pre-trained language models through the lens of discourse connectives. *arXiv preprint arXiv:2109.12951*.

[Pasini et al., 2021] Pasini, T., Raganato, A., Navigli, R., et al. (2021). Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press.

[Piantadosi et al., 2012] Piantadosi, S. T., Tily, H., and Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

[Pilehvar and Camacho-Collados, 2019] Pilehvar, M. T. and Camacho-Collados, J. (2019). WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273.

[Pimentel et al., 2020] Pimentel, T., Hall Maudslay, R., Blasi, D., and Cotterell, R. (2020). Speakers fill lexical semantic gaps with context. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4004–4015.

[Pylkkänen et al., 2006] Pylkkänen, L., Llinás, R., and Murphy, G. L. (2006). The representation of polysemy: Meg evidence. *Journal of cognitive neuroscience*, 18(1):97–109.

[Qi et al., 2020] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

[Qiang et al., 2021] Qiang, J., Li, Y., Zhu, Y., Yuan, Y., Shi, Y., and Wu, X. (2021). Lsbert: Lexical simplification based on bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3064–3076.

[Radford et al., 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

[Raffel et al., 2020] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

[Reif et al., 2019] Reif, E., Yuan, A., Wattenberg, M., Viegas, F. B., Coenen, A., Pearce, A., and Kim, B. (2019). Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32.

[Rovira, 2008] Rovira, L. C. (2008). The relationship between language and identity. the use of the home language as a human right of the immigrant. *REMHU-Revista Interdisciplinar da Mobilidade Humana*, 16(31):63–81.

[Salazar et al., 2019] Salazar, J., Liang, D., Nguyen, T. Q., and Kirchhoff, K. (2019). Masked language model scoring. *arXiv preprint arXiv:1910.14659*.

[Sia et al., 2019] Sia, J., Jonckheere, E., and Bogdan, P. (2019). Ollivier-ricci curvature-based method to community detection in complex networks. *Scientific reports*, 9(1):1–12.

[Sonnenhauser, 2017] Sonnenhauser, B. (2017). Tautologies at the interfaces: Wer kann, der kann. *Journal of Pragmatics*, 117:16–28.

[Thomas, 1978] Thomas, L. (1978). *The lives of a cell: Notes of a biology watcher*. Penguin.

[Tuggy, 1993] Tuggy, D. (1993). Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, 4:273–290.

[Vicente and Falkum, 2017] Vicente, A. and Falkum, I. L. (2017). Polysemy. In *Oxford Research Encyclopedia of Linguistics*.

[Vilinbakhova and Escandell-Vidal, 2020] Vilinbakhova, E. and Escandell-Vidal, V. (2020). Interpreting nominal tautologies: Dimensions of knowledge and genericity. *Journal of Pragmatics*, 160:97–113.

[Vilinbakhova and Escandell-Vidal, 2021] Vilinbakhova, E. and Escandell-Vidal, V. (2021). Tautologies with proper names in discourse: Rhetorical relations and interpretation. *Language & Communication*, 76:79–99.

[Wang et al., 2021] Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F., and Tu, K. (2021). Automated concatenation of embeddings for structured prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2643–2660, Online. Association for Computational Linguistics.

[Ward and Hirschberg, 1991] Ward, G. L. and Hirschberg, J. (1991). A pragmatic analysis of tautological utternances. *Journal of pragmatics*, 15(6):507–520.

[Wee, 2005] Wee, L. (2005). Intra-language discrimination and linguistic human rights: The case of singlish. *Applied Linguistics*, 26(1):48–69.

[Wiedemann et al., 2019] Wiedemann, G., Remus, S., Chawla, A., and Biemann, C. (2019). Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*.

[Wierzbicka, 1987] Wierzbicka, A. (1987). Boys will be boys:'radical semantics' vs.'radical pragmatics'. *Language*, pages 95–114.

[Wijesinghe et al., 2021] Wijesinghe, A., Wang, Q., and Gould, S. (2021). A regularized wasserstein framework for graph kernels. *arXiv preprint arXiv:2110.02554*.

[Winkler, 2015] Winkler, S., editor (2015). *Ambiguity*. De Gruyter, Berlin, Mnchen, Boston.

[Wolf et al., 2020] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

[Xu et al., 2021] Xu, Z., Guo, D., Tang, D., Su, Q., Shou, L., Gong, M., Zhong, W., Quan, X., Jiang, D., and Duan, N. (2021). Syntax-enhanced pre-trained model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5412–5422.

[Xypolopoulos et al., 2021] Xypolopoulos, C., Tixier, A., and Vazirgiannis, M. (2021). Unsupervised word polysemy quantification with multiresolution grids of contextual embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3391–3401.

[Yamshchikov et al., 2020] Yamshchikov, I. P., Saha, C. M. N., Samenko, I., and Jost, J. (2020). It means more if it sounds good: Yet another hypothesis concerning the evolution of polysemous words. *arXiv preprint arXiv:2003.05758*.

[Yang et al., 2020] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2020). Xlnet: Generalized autoregressive pretraining for language understanding.

[Yenicelik et al., 2020] Yenicelik, D., Schmidt, F., and Kilcher, Y. (2020). How does BERT capture semantics? a closer look at polysemous words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.

[Yoo et al., 2021] Yoo, K. M., Park, D., Kang, J., Lee, S.-W., and Park, W. (2021). Gpt3mix: Leveraging large-scale language models for text augmentation. *arXiv preprint arXiv:2104.08826*.

[Zhou et al., 2020] Zhou, J., Zhang, Z., Zhao, H., and Zhang, S. (2020). LIMIT-BERT : Linguistics informed multi-task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461.

[Zhou et al., 2019] Zhou, W., Ge, T., Xu, K., Wei, F., and Zhou, M. (2019). BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373.