Queen Mary
University of London

Word Sense Distance and Similarity Patterns in Regular Polysemy

Insights Gained from Human Annotations of Graded Word Sense Similarity and an Investigation of Contextualised Language Models

by

Janosch Haber

Student-ID: 180993419

A Dissertation Submitted to the
School of Electronic Engineering and Computer Science
in Partial Fulfilment of the Requirements for the Degree of
Doctor of Philosophy (PhD)

November 14, 2022

Supervisor:	*External Examiner:*
Prof. Dr. Massimo Poesio	Prof. Dr. Aline Villavicencio
Co-supervisors:	*Internal Examiner:*
Dr. Julian Hough	Prof. Dr. Matthew Purver
Prof. Dr. Patrick Healey	

Page intentionally left blank

Abstract

This thesis investigates the notion of distance between different interpretations of polysemic words. It presents a novel, large-scale dataset containing a total of close to 18,000 human annotations rating both the nuanced sense similarity in lexically ambiguous word forms as well as the acceptability of combining their different sense interpretations in a single co-predication structure.

The collected data suggests that different polysemic sense extensions can be perceived as significantly dissimilar in meaning, forming patterns of word sense similarity in some types of regular metonymic alternations. These observations question traditional theories postulating a fully under-specified mental representation of polysemic sense. Instead, the collected data supports more recent hypotheses of a structured representation of polysemy in the mental lexicon, suggesting some form of sense grouping, clustering, or hierarchical ordering based on word sense similarity.

The new dataset then also is used to evaluate the performance of a range of contextualised language models in predicting graded word sense similarity. Our findings suggest that without any dedicated fine-tuning, especially BERT Large shows a relatively high correlation with the collected judgements. The model however struggles to consistently reproduce the similarity patterns observed in the human data, or to cluster word senses solely based on their contextualised embeddings.

Finally, this thesis presents a pilot algorithm for automatically detecting words that exhibit a given polysemic sense alternation. Formulated in an unsupervised fashion, this algorithm is intended to bootstrap the collection of an even larger dataset of ambiguous language use that could be used in the fine-tuning or evaluation of computational language models for (graded) word sense disambiguation tasks.

Acknowledgements

My thanks goes to the members of the DALI project and Queen Mary's CogSci group for countless illuminating discussions, especially on how a native speaker of English might interpret a certain word in a given context, how to correctly apply statistical methods to evaluate noisy, continuous annotation data - and how to persuade a contextualised language model to run on a GPU. I especially want to thank Derya Çokal and Andrea Bruera, who had to endure my coming up with ever more examples of lexical ambiguity in an attempt to find my way into the subject - and still provided great feedback and support throughout the past few years.

My gratitude also goes to Pat Healey and Julian Hough for their involvement and input as co-supervisors, and most of all to Massimo, who gave me (what I perceived as) a free rein in developing my research topic, method and materials - even when they slowly started to move away from where we originally had set off - all the while providing immeasurably helpful feedback and ideas drawn from an impressive knowledge of, experience in and passion for (linguistics) research in all of its forms and facets.

This thesis developed in pre-, mid- and post-COVID times, each yielding a very different experience - not just of research, but of life in general. I want to thank my friends and family for being a reliable and balancing constant through all these times; with their calls, online and offline (board)games, and cycling trips (thankfully) forcing me to once in a while stop staring at my screen.

Finally, I want to thank my fiancée Sophie for being so much more than I could have dared to hope for.

The research presented in this thesis was funded by the Disagreements in Language Interpretation (DALI) project, ERC Grant 695662, and in part supported by the Enrichment Scheme of The Alan Turing Institute.

Impact Statement

The primary contribution of this thesis is the presentation of a novel, large-scale dataset of graded word sense similarity. The dataset contains a total of close to 18,000 human annotations rating both the nuanced sense similarity in lexically ambiguous word forms as well as the acceptability of combining their different sense interpretations in a single co-predication structure. We hope that this resource will prove useful for future research in the (psycho-)linguistics and computational linguistics communities, either by providing empirical evidence for the investigation of the processing of lexical ambiguity, or as a means to fine-tune and evaluate computational language models.

This thesis also presents a first detailed analysis of the performance of contextualised language models in predicting human annotations of graded word sense similarity using the new dataset. With black-box models like the contextualised language models tested in this study achieving remarkable down-stream task performances but limiting the interpretability of their representation of semantic content, we hope that this evaluation will provide useful insights allowing for specific optimisations or adjustments in future work.

Lastly, this thesis contains a comprehensive overview of the scientific literature on lexical ambiguity in general and the processing and representation of polysemy in particular. It attempts to clarify any ambiguous or vague use of definitions that have emerged in the past few decades of studies and theories treating this subject, and we hope that this overview will provide a useful starting point for anyone looking to learn more about the field or intends to contribute to it.

Contents

List of Figures

List of Tables

Abbreviations and Acronyms

AMT - Amazon Mechanical Turk

BERT - Bidirectional Encoder Representations from Transformers (Devlin et al., 2019)

CBOW - Continuous Bag-of-Words (Mikolov et al., 2013a,b)

DURel - Diachronic Usage Relatedness (Schlechtweg et al., 2018)

EEG - Electroencephalogram

ELMo - Embeddings from Language Models (Peters et al., 2018)

ERP - Event Related Potential

fMRI - functional Magnetic Resonance Imaging

GL - Generative Lexicon (Copestake and Briscoe, 1995; Pustejovsky, 1995)

GloVe - Global Vectors (Pennington et al., 2014)

GLUE - General Language Understanding Evaluation benchmark (Wang et al., 2018)

GPT - Generative Pre-trained Transformer (Radford et al., 2018)

HIT - Human Intelligence Task (AMT Questionnaire)

IAA - Inter-annotator Agreement

JSD - Jensen-Shannon divergence

KDE - Kernel Density Estimate

kNN - k Nearest Neighbour clustering

LM - (Computational) Language Model

LSTM - Long Short-Term Memory Hochreiter and Schmidhuber (1997)

MEG - Magnetoencephalography

MLM - Masked Language Modelling (Devlin et al., 2019)

MSE - Mean Squared Error

NLP - Natural Language Processing

NN - Neural Network

NNLM - (Feed-forward) Neural Net Language Model

NMI - Normalised Mutual Information

NSP - Next Sentence Prediction (Devlin et al., 2019)

OLS - Ordinary Least Squares Regression

RAW-C - Relatedness of Ambiguous Words Trott and Bergen (2021)

RNN - Recurrent Neural Net

RNNLM - Recurrent Neural Net Language Model

RT - Relevance Theory, Reaction Time, Reading Time

SEL - Sense Enumeration Lexicon

SVM - Support Vector Machine

t-SNE - t-Distributed Stochastic Neighbour Embedding (van der Maaten and Hinton, 2008)

Usim - Word Usage Similarity (Erk et al., 2009)

WiC - Words in Context Pilehvar and Camacho-Collados (2019)

WSD - Word Sense Disambiguation

WSI - Word Sense Induction

WSsim - Word Sense Similarity (Erk et al., 2009)

Chapter 1

Introduction

The past few years have seen the emergence of deep contextualised language models, large neural networks with up to multiple billions of parameters designed to encode the meaning of a specific sentence - and each word within it. One of the main drives behind this development has been the problem of correctly representing and interpreting *homonyms.* Homonyms are words that can take on completely different meanings in different contexts, like for example *match*:

(1) a. The *match* fell on the carpet and left a burn mark.
    b. The *match* ended without a winner even after going into overtime.

While a large array of scientific work in computational linguistics and psycholinguistics thus has been focused on investigating homonymy and suggesting models for its processing in the human brain and in computational approaches, both fields so far have paid less attention to a closely related, but arguably much bigger phenomenon: *polysemy.*

Polysemous words also can take on different interpretations in different contexts - but what distinguishes them from homonyms is that their interpretations are closely related, and often invoke different aspects of or perspectives on the same concept. Take for example the different interpretations of *school* in example (2):

(2) a. The *school* has a dull brown facade. (building)
    b. The *school* has prohibited light-up sneakers. (administration)
    c. The *school* won last year's play-offs. (sports team)
    d. The *school* is well respected among researchers. (institution)

Most if not all content words are considered to be polysemous to some degree, and accumulating evidence suggests that the phenomenon is far less homogeneous than often assumed: eye-tracking as well as Electro- and Magnetoencephalography

(EEG and MEG) studies have indicated not just differences in the processing of homonyms and polysemes, but also between different types and interpretations of polysemous sense extensions.

With the work presented in this thesis we hope to contribute to the investigation of the representation of polysemic word sense in both the human language processor and computational language models. To do so, we present a novel, carefully designed, human-annotated dataset of word sense similarity, and use it to evaluate different models of the mental lexicon as well as a range of contextualised language models to analyse their capability in predicting the collected judgements. The full dataset is publicly available, for example on `https://github.com/dali-ambiguity/Patterns-of-Lexical-Ambiguity`.

1.1 The Challenge of Representing Lexical Ambiguity

One of the central approaches to explaining how humans process language is to stipulate a mental lexicon that connects words with their interpretation. This approach is immediately challenged by the existence of ambiguity - and especially phenomena of lexical ambiguity like homonymy and polysemy - which require a context-specific disambiguation of a word before its interpretation can be assigned.

To allow for selecting different interpretations in different contexts, words with multiple meanings usually are proposed to be represented by different entries in the mental lexicon - much like they are in physical dictionaries. The representation of polysemic senses on the other hand is much more debated, with approaches ranging from treating them in the same way as homonyms (so-called Sense Enumeration Lexicons), to complex, semantically under-specified representations of meaning (the One Representation hypothesis). While the former usually requires each distinct entry to have its own clear interpretation, as well as necessary discrimination and selection criteria, the latter often assumes that the different senses of polysemic words are in fact so similar that no specific one is selected in their interpretation. Under-specification approaches therefore require polysemic word senses to be so similar that all of them invoke the same, under-specified entry in the mental lexicon.

In this thesis we present an empirical investigation of human judgements of word sense similarity that allows us to scrutinise recently proposed hybrid models of the mental processing of polysemes. Based on conflicting observations in previous research, some of these new hybrid models for example suggest a hierarchical or distance-based arrangement of polysemic word senses within a single under-specified entry. This assumes that the similarity between different polysemic sense interpretations is not a binary measure, but rather a continuous one that determines their grouping or clustering in the mental lexicon - which in turn causes the processing differences observed in previous studies.

The data collected in this study suggests that some polysemic alternations are perceived to be less similar in their interpretation than others, and that these differences form relatively consistent patterns across different word forms allowing for the same type of alternation. We will argue that these observations cannot be readily explained by either, sense enumeration or fully under-specified models of the mental lexicon, and instead indicate that - at least for regular polysemes - sense similarity indeed might be a central factor in structuring their mental representation.

The second focus of this thesis will be an evaluation of contextualised language models based on the human-annotated data. With models like ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) and GPT (Radford et al., 2018), computational approaches only recently gained the ability to represent a given word within the context of a particular use. While this has shown to drastically improve performance of a wide range of downstream natural language processing (NLP) tasks, analyses of what information exactly is encoded in their contextualised embedding vectors have proven difficult and inconclusive. Given that now we have available a dataset of human annotations judging the nuanced differences in polysemic sense interpretations, we propose to utilise this dataset to evaluate how well the different contextualised language models can predict these judgements. We will show that especially BERT Large's contextualised word representations seem to capture polysemic sense alternations relatively well, with its predicted word sense similarity scores correlating with the human judgements to the same degree as the human annotations correlate with each other.

Finally, we will investigate the use of BERT Large's contextualised embeddings as a heuristic to bootstrap the automatic expansion of the collected dataset. To this end we present a pilot algorithm aimed at detecting additional word forms that allow for the same alternations as those previously tested in the study. We will leverage the variance of contextualised word embeddings derived from an unlabelled corpus sample to provide an unsupervised assessment of a target word's polysemic potential, and to determine whether or not a given word allows for a specific sense extension. This information then could be used to select further targets for annotation and increase the size of the dataset to a scale more suitable for the fine-tuning and evaluation of deep neural networks.

1.2 Research Questions

In this thesis, we will address the following research questions:

Q1 Does empirical evidence on word sense distance support the traditional hypotheses of word sense enumeration or fully under-specified mental representation of polysemes in the mental lexicon?

Q1a Does empirical evidence on word sense similarity indicate differences in the interpretation of different polysemic senses?

Q1b Does word sense distance form discernible patterns in the interpretation of regular polysemic senses?

Q2 Do computational approximations of word sense similarity correlate with empirical indications of word sense distance?

Q2a Can contextualised word embeddings be used to classify and identify (new) polysemic targets?

The first set of research questions are specifically aimed at contributing to the ongoing debate on the mental representation of polysemic word senses. While a perceived identity in meaning between two different polysemic senses is difficult to explain with a sense enumeration approach (what distinguishes these two interpretations to warrant distinct entries in the lexicon), significant differences in the similarity of two polysemic senses prove difficult to reconcile with a fully under-specified approach (if two senses are interpreted differently, how can they both be derived from the same entry).

Instead of a purely theoretical contribution, this thesis provides empirical data -and an investigation thereof - that is suitable to investigate support or challenges for the different models of the mental lexicon. Since we cannot present an exhaustive analysis of all polysemes, we focus on a subset of the most regular polysemic alternations identified in previous literature. Results for these regular types should yield the clearest results possible while allowing for generalising findings to more irregular subtyes. They also allow for an investigation of polysemy patterns which, by removing the risk of analysing one-off outliers, will allow for even stronger arguments in the discussion of the possible mental representation of polysemic word sense.

The second set of questions investigates the performance of - and, by extension, their potential as a research tool - a recent range of contextualised language models.

If these models produce word sense similarity approximations that correlate well with the human judgements collected in the first part of this study, the contextualised language models could present an inexpensive means to expanding corpora and research on the interpretation and representation of polysemic sense.

1.3 Thesis Outline

Derived from our research questions, we set ourselves a range of research goals:

G1 Provide a comprehensive overview of the (psycho-)linguistics literature on lexical ambiguity, and specifically the phenomenon of polysemy and its representation in the language processor.

G2 Establish a strict definition of a specific type of polysemic sense extension to be investigated, and determine a set of seminal and experimental target word forms for this type.

G3 Develop materials and methodology suitable to collect human-annotated data reflecting the nuanced differences between the polysemic sense extensions of the selected targets.

G4 Collect sufficient annotated data to allow for the validation of the developed materials and methodology.

G5 Collect a large-scale dataset of fine-grained word sense similarity judgements.

G6 Using the collected data, investigate the notion of word sense similarity and a potential distance-based grouping or clustering of word senses in the mental lexicon.

G6 Using the collected data, investigate how well different contextualised language correlate with human annotations of word sense similarity.

G7 Investigate the potential of contextualised language models in collecting additional data on lexical ambiguity.

The following chapters will address these research goals in the context of answering our main research questions. As a result, this thesis is structured as follows: the literature review of Chapter 2 will define polysemy as a form of lexical ambiguity occupying a unique middle ground between multiplicity of meaning and identity of sense, and further sub-divide the phenomenon into different classes and types of sense extensions. In Section 2.4 we will investigate seminal and contemporary work focusing on the mental processing of polysemy, discussing theoretical models of word sense representation together with empirical evidence that either supports or challenges different models. Here we also will introduce the notion of word sense distance, and indicate how investigating word sense representation through word sense similarity can contribute to the ongoing debate between fully structured and under-specified approaches to the mental lexicon.

After exploring these (psycho-)linguistic accounts, in Chapter 3 we will briefly present the computational linguistics discipline of distributional semantics. Distributional semantics aims to approximate word senses by inferring the relationships between words from large amounts of corpus data. Decades of research in this field led to the recent introduction of contextualised language models (Section 3.3), a range of neural networks able to generate computational representations of individual words relative to their respective context. We will show how these models have been applied to improve performance on NLP tasks, and indicate specifically their use in word sense disambiguation (WSD) tasks (Section 3.3.6).

In Chapter 4 we present a data collection pilot intended to validate a novel approach to establishing a crowd-sourced data set of human annotated judgements of word sense similarity. We will report in detail the materials (Section 4.3) and methodology (Section 4.4) used in the pilot, and present a preliminary analysis of the data collected for a small number of polysemic targets to evaluate the chosen approach (Section 4.5). We will then also already provide a preliminary investigation of how well contextualised language models can predict the human annotations, finding that the small sample set at that point precludes conclusive insights.

After having validated the data collection approach, Chapter 5 presents the collection of a large-scale dataset with close to 18,000 judgements for 28 polysemic targets representing ten different types of regular, metonymic polysemy. With materials and methodology largely unchanged, Section 5.3 presents a thorough analysis of the collected data, indicating that some types of alternations exhibit relatively consistent patterns of word sense similarity visible in different kinds of human annotations. Section 5.3.3 then revisits the investigation of contextualised language models based on the full set of annotated data, now finding clear links between predicted similarity scores and human judgements, with especially BERT Large's predictions correlating well with the collected word sense similarity judgements. All computational models however struggle to consistently re-create the specific similarity patterns observed in the human annotations.

Motivated by the good correlation between BERT predictions and human annotations, Chapter 6 presents a pilot algorithm for automatically extending the collected dataset by using an unsupervised heuristic to detect additional target words that allow for the same set of alternations as the polysemic targets included in the annotation experiments. While still showing some teething problems, the presented pilot provides a promising proof of concept, and with a few modifications might prove to be a useful tool in creating a dataset of word sense similarity large enough to allow for a new way of fine-tuning of contextualised language models for word sense discrimination tasks.

Chapter 7 finally will summarise our findings, discuss them in the context of previous work, and pinpoint potential avenues for continuing the investigation of both the mental and computational representation of polysemy.

Chapter 2

Lexical Ambiguity and the Mental Lexicon

Lexical ambiguity is the phenomenon of a word exhibiting multiplicity of meaning, i.e. allowing for different interpretations in different contexts. Lexical ambiguity is ubiquitous in everyday language, and poses interesting questions and challenges to both (psycho-)linguists and computational linguists research: why do we use ambiguous expressions? Why do they not interfere with communicative goals more often? How do we process ambiguous words; how are they stored in our brains? And how should computational language models represent and deal with instances of lexical ambiguity?

A central issue in investigating lexical ambiguity - interestingly or ironically - is that traces of ambiguity and vagueness also found their way into the literature on this topic, affecting nomenclature and definitions. To give an example, traditionally lexical ambiguity is further differentiated into two closely related phenomena, homonymy and polysemy. Definitions for these two phenomena however are highly debated and can differ among specific fields of research - while some scholars do not make a distinction at all - and in some cases the term polysemy has been used as a synonym for lexical ambiguity itself. In addition, a large part of seminal work on the subject is based on introspective arguments and paradigmatic or anecdotal evidence, which will inevitable fall short of capturing and representing the full complexity of the phenomena involved, and, in consequence, is likely to over-simplify especially the requirements for the processing of lexically ambiguous words.

One of the main goals of this thesis is to provide a sound, empirical investigation of a well-defined sub-set of polysemic alternations in order to i) showcase the heterogeneity and diversity of effects, but also any systematic patterns present in that sample, ii) through the collected data provide representative evidence to inform the development of mental models of language processing, and iii) establish a reliable dataset for the evaluation of computational approaches to (graded) word sense disambiguation.

As investigating phenomena of ambiguity under ambiguous definitions is doomed to generate ambiguous results at best, in this chapter we will review seminal and contemporary approaches to classifying lexical ambiguity in general, and distinguishing polysemy and homonymy in particular. We will begin by introducing and developing central terminology in Section 2.1, followed by a detailed investigation of different types and classifications of polysemy in Section 2.2. In Section 2.3 we will address seminal accounts of vagueness and under-specification that seem to question the sheer possibility of clear and succinct definitions, but ultimately can be - and have been - incorporated into theoretical accounts of lexical ambiguity. In Section 2.4 we introduce the most prominent models of the mental lexicon proposed to explain the processing of lexical ambiguity in language users, followed by linguistic and behavioural evidence collected in support or in opposition of these different models (Section 2.5). Finally, in Section 2.6, we will briefly summarise the main findings of the literature review.

2.1 Multiplicity of Meaning and Multiplicity of Sense

Modern investigations of multiplicity of meaning in the widest sense date back to at least Breal (1897), who noticed that many expressions in everyday interactions were ambiguous, but surprisingly rarely led to miscommunication. From among the different phenomena of ambiguity observed in natural language (see Poesio, 2020, for a recent overview), this thesis focuses on lexical ambiguity, the phenomenon of a single word exhibiting multiplicity of meaning, i.e. allowing for different interpretations in different contexts. Traditionally phenomena of lexical ambiguity are further subdivided into homonymy and polysemy, separating *multiplicity of meaning* from *multiplicity of sense*. In an attempt to better distinguish these two phenomena, Weinreich (1964) for example note that 'homonymy is observed in lexical items that *accidentally* carry two distinct and unrelated meanings' (also see Klepousniotou et al., 2012). This notion of two things 'accidentally' being assigned the same name also is reflected in the choice of its description, with the term *homonym* being derived from Ancient Greek ὁμο- (homo-, *same.*) and ὄνυμα (ónuma, *name*). Seminal, paradigmatic examples of homonyms include nouns like *match*, *bat* and *mole*, which can invoke at least two different, arguably unrelated interpretations:

(3) a. The *match* fell on the carpet and left a burn mark.

b. The *match* ended without a winner even after going into overtime.

(4) a. The *bat* was found hibernating in an attic.
b. The *bat* was expertly crafted from a single piece of wood.

(5) a. The *mole* dug a number of tunnels through the front yard.
b. The *mole* on her shoulder stopped bothering her after a while.

Contrasting the unrelatedness of homonymic interpretations, polysemy[1] traditionally has come to signify words that can invoke different distinct but related interpretations (Lyons, 1977; Swinney, 1979; Simpson, 1994; Pinkal, 1995; Cruse, 1995; Ravin and Leacock, 2000), like *school* in Example (6).

(6) a. The *school* has a dull brown facade. (building)
b. The *school* has prohibited light-up sneakers. (administration)
c. The *school* won last year's play-offs. (sports team)
d. The *school* is well respected among researchers. (institution)

Both of these phenomena are set in opposition to *monosemy*, where words are assumed to be associated with just one, fixed interpretation. Initially, distinguishing homonymy from polysemy was often considered of little theoretical interest, and literature was more focused on investigating the distinction between monosemes and lexically ambiguous word forms instead (see e.g. Kempson, 1977; Cruse, 1986). Vicente (2015) later argued that 'part of this neglect is due to the fact that philosophical and a good part of linguistics semantics have been focused on sentential, truth-conditional, meaning, instead of on lexical meaning for a long time. But another part has to do with [...] the idea that, barring homonymy, each word-type has a unique simple denotation.' Lexical ambiguity however is a ubiquitous phenomenon, with Durkin and Manning (1989) for example estimating that 40% of frequent English words are polysemous, while scholars like Zipf (1945); Rodd et al. (2002) and Travis (2008) even argue that basically every content word can be used polysemically - a notion we will explore in Section 2.3. Estimates for homonymy are a little more conservative, with for example Dautriche (2015) suggesting that about 4% of English words can have multiple, unrelated meanings.

Distinguishing homonymy from polysemy based on the notion of 'relatedness of meaning' however also has been met with strong criticism (e.g. Lyons, 1977; Kilgarriff, 1997): relatedness itself is at best a vague proposition open to contextual biases, subjective judgement or 'folk etymology,' while determining homonymy based on historically 'formally distinct items in some earlier stage of the language' (Klepousniotou, 2002) suffers from unclear historical derivations, and begs the question how

[1] from Ancient Greek πολύς (polús, *many, much*) and σῆμᾰ (sêma, *mark, sign, token*)

far back one should got in tracing the history of words (Lyons, 1977). The resulting vague boundary between polysemy and homonymy is at least partially responsible for the sometimes conflicting observations about the processing of homonymic and polysemic words in previous literature - as we will discuss in Section 2.5. When presenting previous work, we will therefore - whenever possible - aim to clarify how authors classify specific samples.

2.1.1 A Note on Terminology

Most linguistics literature will label a given word to be either a monoseme, polyseme or homonym. As however many (if not all) content words - including homonyms - will have polysemic sense alternations, this can lead to confusing terminology. A range of seminal publications for example introduce either the phenomenon of polysemy or the concept of homonymy with the word *bank* (see e.g. Klein and Murphy, 2002; Nerlich and Clarke, 2003; Poesio, 2020) - as it has both, polysemic and homonymic alternations:

(7) a. The *bank* was involved in a scandal. (financial - institution)
b. A robber crashed a car into the *bank*. (financial - building)
c. The *bank* was littered with tons of plastic waste. (landscape - feature)

As a result, labelling specific words as homonyms or polysemes strictly speaking is not very meaningful, and instead their different interpretations should be considered polysemic or homonymic in their relation to one another. With that in mind, in the remainder of this thesis we will use the terms *homonymic* or *polysemous* to refer to specific alternations of a given ambiguous word. To further clarify this distinction, we will refer to the different polysemic extensions of a word as different word *senses*, and the different interpretations of homonymic alternations as *meanings*. Note that this naming convention will be in conflict with some previous literature using the these terms to refer to either or neither of the phenomena in particular.

2.2 Polysemy: Types, Classes and Other Subdivisions

Besides unclarity in their definition, a second central aspect impeding with a clear distinction between homonymy and polysemy is the observation that the latter phenomenon appears to be quite heterogeneous in its appearance, and that the 'relatedness' of polysemic sense interpretations can range between near identity on one end of the spectrum to resembling homonymy on the other. To address this issue, a number of publications have been attempting the definition of sub-types, classes, and other distinctions of specific polysemic alternations, each presented in conjunction with hypotheses as to how the relatedness in sense interpretations evident in this sub-type affects the proposed models of the mental processing of polysemes.

One of the most well-known and commonly accepted distinctions is that polysemic alternations are considered to either be *idiosyncratic* (sometimes labelled *accidental*), or *regular.* Following the definition of Apresjan (1974), a lexeme A with senses a_1 and a_2 is an example of regular polysemy if there exists at least a second lexeme B for which its senses b_1 and b_2 are 'semantically distinguished in exactly the same way as a_1 and a_2' - although later publications like Falkum (2015); Vicente and Falkum (2017) and Ortega-Andrés and Vicente (2019) note that to actually exhibit regularity in its polysemic sense alternation, it should have more than one corresponding alternative. According to Vicente and Falkum (2017), 'regular polysemy is typically associated with senses generated by metonymic extensions, and irregular polysemy with senses that tare derived metaphorically' (also see Apresjan, 1974; Bowdle and Gentner, 2005). These descriptions link the observed relation between two different sense extensions to two more well-known figures of speech: metaphoric (from Greek μεταφορά (metaphorá), *transference*) sense alternations are observed in cases where an interpretation that is more inherent to one concept is transferred to another, (unrelated) concept - but still evokes a related meaning. An example for this kind of metaphoric extension is noun *mouth* in Example (8)

(8) a. She has a number of freckles on her nose and close to her **mouth**.
    b. The river never is more than 20 feet across, except close to its **mouth**.

Metonymic extensions (from Greek μετωνυμία, metōnymía, *a change of name*) on the other hand usually indicate sense extensions referring to different aspects or facets of the same entity. The different uses of *school* in Example (6) showcase this kind of polysemic sense extension, with different contexts invoking different aspects of the concept *school.*

A wide range of previous research has been focused on naming and specifying some of the most frequent alternations observed in regular polysemy, including for example *animal/meat* alternations (see Example (9), cf. Copestake and Briscoe, 1995; Frisson and Frazier, 2005; Falkum, 2015), *container/content* alternations (Example (10), e.g. Schumacher, 2013) or *physical/information* alternations (Example (11), cf. Pustejovsky, 1993; Antunes and Chaves, 2003; Frisson, 2015):

(9) a. The **chicken** pecked for some seeds in the shadow of the barn.
    b. The **chicken** was seasoned deliciously and served with potato wedges.

(10) a. Nervously waiting for his date, he peeled the label off his **beer**.

b. He didn't remember much as he had had way too much **beer**.

c. She carefully placed the priceless **bottle** in a padded box.

d. She only had about half a **bottle**, but she was feeling tipsy already.

(11) a. They found the **book** wedged under a window to create some airflow.

b. After two semesters they were able to cite most of the **book**.

These alternations cannot only be found in different English words,[2] but exhibit matching counterparts in many other languages, with Srinivasan and Rabagliati (2015) presenting evidence of 27 distinct cases of English polysemy also being present in 14 different languages, 'suggesting that polysemy arises from conceptual constraints rather than arbitrary, language-specific conventions' (Murphy, 2021).

Pustejovsky (1995) later introduced an even more fine-grained distinction between different types of regular polysemy: while some expressions are *merely regular*, they considered others to be *inherently* polysemous. For a term to exhibit inherent polysemy, the different senses need to be 'somehow inherent to the entity that the term denotes.' Ortega-Andrés and Vicente (2019) for example propose that the noun *book* could be said to have inherently polysemic interpretations, as both the *physical* (He put the *book* back in the shelf) as well as the *information* reading (He read the *book* in under two hours) are inherent to what a book is. They however also argue that this characterisation of inherent polysemy is rather vague, as there is no clear definition as to when certain sense interpretations are inherent or not.

Dölling (2020) offers an alternative distinction, contrasting *metonymic* and *inherent* polysemy. Following their definition, metonymic polysemy describes 'cases where one of the related senses is primary and the others are metonymically derived from it,' while inherent (or logical) polysemy 'involves senses where there are no substantial reasons for assuming that one or another of them is prior' (based on earlier observations by e.g. Nunberg, 1995; Copestake and Briscoe, 1995). From their collection of systematic polysemes, Dölling identify nouns such as *rabbit, apple, oak, beer* and *bottle* as exhibiting metonymic extensions where 'even though each of the senses are in equal measure usual, one of them is primary,' and all interpretations exhibit a 'normal, conventionalised use.' As a rule of thumb, alternations that can be described through patterns like *animal-for-meat*, *fruit-for-pulp* or *container-for-content* are likely to be cases of metonymic polysemy. Inherent polysemic sense extension on the other hand were identified for nouns like *book, speech, bank, newspaper* and *lunch*, where 'neither interpretation can be viewed as more basic.' Potential targets here are words where function and physical realisation both are integral - a

[2] or German, in the case of Schumacher (2013)

book without its physical realisation or content for example would arguably not be a book.

Recent literature more and more demands reliable empirical data to replace the use of individual paradigmatic examples in the investigation of polysemic alternations (see e.g. Klepousniotou et al., 2008; Erk and McCarthy, 2009; Schumacher, 2013; Ortega-Andrés and Vicente, 2019). Löhr (2021) for example call for a more stringent definition of the term 'polysemy' itself to substitute seminal definitions derived from anecdotal evidence, lamenting that the under-specified concept of 'relatedness of meaning' does not suffice for this task. Without presenting any data themselves, they later propose a reformulated definition of polysemy, following Recanati (2017) in reserving the term for 'related senses that are conventionalised.' This however still seems to fail to address the previously highlighted issue of the loose definition of relatedness - and merely adds the similarly loosely defined concept of conventionalisation to the mix. To better address the need for more empirical evidence, one of the main goals of this thesis will be the collection of a representative dataset for a well-defined sub-set of polysemic alternations that allows for a data-driven evaluation of different phenomena of lexical ambiguity. We will be focusing on regular polysemic alternations, allowing us to investigate multiple target words per type of alternation, as well as metonymic extensions (which usually are considered regular) to showcase the diversity in effects even for alternations proposed to be very closely related.

2.3 Vagueness and Coercion

Both homonymy and polysemy are considered facets of ambiguity rather than vagueness, i.e. the potential interpretations of polysemic or homonymic expressions form a discrete set rather than a continuous transition. Vagueness however also plays a central role in the theoretical conceptualisation of these phenomena and their definitions: when are two senses of a word different enough to be considered distinct interpretations? Where is the cutoff-point for the relatedness of interpretations? And how can we account for the infinite sets of (discourse and deictic) contexts impacting the meaning and interpretation of a word?

The latter question is reflected by a phenomenon which Cruse (2000) called *ways of seeing*, and often also is referred to as *context coercion* (also see Anderson and Ortony, 1975, for an ealry discussion). Context coercion can best be understood with verbs like *run* that exhibit an extraordinary amount ambiguous productivity (Brugman, 1988; Gilliver, 2013). Consider sentences like

(12) a. John is **running** at least 5k every morning.
b. The bank robber is **running** from the police.
c. The dog is **running** in the park.
d. The water is **running** down the steps.
e. My nose just won't stop **running**.
f. The coffee machine is **running** all morning.

While all of these uses of *run* elicit arguably closely related interpretations, the different contexts of the sentences in Example (12) ever so slightly change the meaning of the word: running for exercise is a different kind of running than when chased by the police (being *on the run*); a dog runs differently than a human does; water and other liquids run differently again - with a running nose arguably eliciting at least a different connotation than a babbling stream; and while water runs through it, a coffee machine again invokes a different conceptualisation of *run*.[3]

Much of the discussion on context coercion is exemplified by the debate famously held by Jackendoff (1989) and Fodor (1998), who over a period of time discussed specifically the verb *to keep*, with Jackendoff arguing that *keep* must sure be polysemous with its uses in phrases like *keep the change*, *keep your car in the garage*, *keep the crowd happy*, while Fodor argues that *keep* in fact only has a single meaning, and 'the apparent differences in meaning are simply an artefact of the different contexts in which the verb appears' (also see Falkum and Vicente, 2015).

Seeing context coercion as the main drive of polysemic sense extensions, some scholars take an entirely pragmatic approach in explaining phenomena of lexical ambiguity and specifically polysemy (which we will briefly discuss in Section 2.4.3), and others postulate that all content words are in fact polysemous (see e.g. Zipf, 1945; Travis, 1997; Rodd et al., 2004). While the scope of the 'multiplicity of sense' classification of polysemy allows for the inclusion of context coercion, a definition like this blurs the line between ambiguity and vagueness, as no longer all senses could be clearly distinguished.

Tying together phenomena of vagueness and ambiguity in a single formalisation, Pinkal (1985) proposed the concepts of h-type and p-type ambiguity to better classify lexical ambiguity (also see Poesio, 2020). Following Pinkal's approach, an expression is h-type ambiguous if and only if its 'indefinite base level is inadmissible.' As a consequence, h-type ambiguous words have to be immediately disambiguated because they do not allow for an under-specified representation of their base level. P-type ambiguous words on the other hand do allow for an under-specified representation,

[3] These are but a selection of the 645 meanings Gilliver compiled when revising the Oxford English Dictionary.

and therefore do not require an immediate disambiguation. These formalisations of h-type and p-type ambiguity can directly be applied to homonymy and polysemy, suggesting that homonyms do not have an admissible under-specified base level, while polysemes do - allowing for vagueness in the interpretation of polysemes but not so for homonyms.

While acknowledging that context coercion will play a central role in the individual, nuanced interpretation of an encountered ambiguous expression, in this thesis we will focus on the discrete set of polysemic and homonymic readings a given ambiguous target can elicit, and investigate how these interpretations relate to one another. We will therefore use the term *polysemy* in a stricter sense, referring only to those pre-defined, inherent interpretations of an ambiguous expression, and attempt to invoke those interpretations as clearly as possible when investigating phenomena of lexical ambiguity in Chapters 4 and 5.

2.3.1 Under-specification and Good-enough Representations

Another central concept in literature distinguishing homonymy from polysemy is the *ambiguity advantage*: the idea that not having to fully specify an interpretation can reduce processing demands while still allowing language users to achieve their intended communicative goals (see e.g. Piantadosi et al., 2012; Winkler, 2015). Swets et al. (2008) for example suggest that ambiguity advantage is a consequence of *strategic under-specification*, meaning that comprehenders will (automatically) spend less time resolving ambiguous expressions when this is not required by the current task or overall processing goal. When translated into an experimental setting, any ambiguity advantage for example should disappear in a relative clause (RC) attachment task when participants are explicitly asked to resolve the attachment. The authors tested this by having three groups of participants reading the same RC samples. The first group was asked questions explicitly focused on RC attachment, the second group was asked superficial questions unrelated to the relative clause, and the third group was occasionally asked about RC attachment. Swets et al. found evidence of an ambiguity advantage only in the second group, but not so in either of the groups that were asked specific RC questions (also see Logačev and Vasishth, 2016).

In a similar approach, Ferreira et al. (2002); Ferreira and Patson (2007); Karimi and Ferreira (2016) investigate the notion of *good-enough representations*. Based on the observation that readers sometimes only seem to establish a shallow reading of a sentence - and in some cases completely misunderstand it - the authors suggest that the language processor might only establish a rudimentary representation of semantic content that is ‘good-enough’ to proceed with an interpretation. Indications of good-enough processing primarily came from observations such as the Moses illusion and garden path sentences, but were then also discovered in the interpretation of co-reference. The Moses illusion is based on a simple question: ‘How many of each type of animal did Moses take on the ark?’ In a study by Erickson and Mattson (1981) most participants answered ‘two’ - failing to notice that not Moses but Noah is said to have built the ark. In a similar way, asking where to bury the survivors of a plane crashed on the border between two nations usually stumps participants rather than having them notice that survivors should not be buried at all (Barton and Sanford, 1993). Garden path sentences include examples like ‘While Mary bathed the baby played in the crib.’ This sentence usually is misunderstood on first encounter as *the baby* appears to be the object of *bathed*, but is in fact the subject of *played* (also see Christianson et al., 2001). Asking participants whether Mary bathed the baby, Ferreira et al. (2001) found that participants still replied ‘yes’ after having a chance to re-read and correctly interpret the sentence, indicating that the initial, wrong interpretation still lingered. This effect was also observed for much simpler passive sentences such as ‘the dog was bitten by the man,’ which was rated plausible by 25% of the participants in Ferreira (2003).

Ferreira and colleagues proposed that the explanation for these observations lies in the way the language processor works: instead of producing perfect representation of a speaker’s intention or a sentence’s interpretation, the language processor ‘responsibility is to create representations that are suitable for the task that the listener [or reader] wants to perform with the help of the linguistic input’ (Ferreira and Patson, 2007). Usually, this task is to produce an appropriate follow-up to continue a dialogue, or to proceed with the next sentence of a text - none of which require for example a full evaluation of truth conditions. Recasens et al. (2011) later also found that a good-enough interpretation might account for perceived near-identity in co-reference, spanning the bridge to a possible good-enough, under-specified representation of polysemic sense in lexically ambiguous targets: if polysemic sense extensions indeed are merely facets of the same concept, most processing tasks might not require a full disambiguation, and polysemic sense can be left under-specified. Homonymic interpretations on the other hand - referring to completely unrelated concepts - will require a full disambiguation to allow for even a shallow interpretation of a sentence. Christianson (2016) summarised the ‘good-enough’ perspective on the language processor as follows:

1. The language processor is bounded, i.e. operating under a set of restrictions derived through task demands, time, processing load and cognitive limitations

2. Any language input could contain inconsistencies - syntactic, statistical, semantic, pragmatic, contextual or otherwise - that the language processor might not attempt to reconcile
3. The usual goal of the language processor is not to build a veridical representation of the input, but rather to facilitate communication (Ferreira and Patson, 2007)
4. The language processor appears to favour a 'fast and frugal' (Ferreira, 2003) when approaching any of the aforementioned points - potentially leading to good enough, under-specified or shallow representations not by accident, but by design
5. The misinterpretations resulting from such good-enough, under-specified or shallow representations are systematic and predictable, and they offer insights into the architecture of the language processor

2.4 Models of the Mental Lexicon

The (human) language processor is a crucial aspect in most - if not all - theories on lexical ambiguity: if a single word can indeed have multiple senses, and sometimes even multiple meanings, how are these connections stored in our mental representation of those words? Or - more figuratively speaking - what is the makeup of our mental lexicon?

Linguistics literature has produced a range of proposals attempting to answer these questions, commonly split into three groups: sense enumeration approaches, one representation models, and pragmatic approaches. We will here discuss these different proposals in more detail, and present them with some preliminary support and principled objections. In Section 2.4.4 we will then introduce a number of contemporary hybrid models that will be the subject of investigation in the remainder of this thesis, and continue by presenting behavioural data generated to evaluate and scrutinise all of these models of the mental lexicon in Section 2.5.

2.4.1 The Sense Enumeration Lexicon

One of the earliest models of the mental lexicon was offered by Katz and Fodor (1963); Katz (1972), who in the grammar of their natural language semantics model included a dictionary in which all senses of a word were to be listed, that, taken together, constitute a word's meaning. This type of mental representation later has come to be known as a sense enumeration approach, or Sense Enumeration Lexicon (SEL). SEL approaches usually do not make a principled distinction between homonymic and polysemic interpretations, with each of them simply being listed as another possible meaning of a given word.[4]

As Falkum and Vicente (2015) noted, sense enumeration models are '*prima facie* the simplest way to deal with polysemy on theoretical grounds,' explaining all variability in the semantic contribution of an expression through its 'different senses stored as distinct representations.' SEL approaches however have not received much support from the academic community. Given the previously mentioned observations that polysemy is a pervasive phenomenon and that some words can have up to hundreds of possible meanings and sense interpretations, assuming individual entries for all of them would require an immense storage complexity and cause a combinatorial explosion when processing sentences containing multiple ambiguous words. Similarly, a number of philosophical concerns have been raised concerning definitional theories in general, with scholars like Kilgarriff (1997) lamenting the difficulty in 'deciding when two senses are different enough to warrant a new entry, and how to represent the information that is common to multiple different senses' and Hanks (2000) questioning whether different senses actually can be represented as disjoint classes defined by necessary and sufficient conditions (also see Wittgenstein, 1953; Tuggy, 1993; Laurence and Margolis, 1999).

More recently - and more specifically - Vicente and Falkum (2017) noted that semantic markers proposed to distinguish senses in SEL approaches cannot account for many of the observed polysemic alternations, and Dölling (2020) remarked that sense enumeration accounts 'miss the generalisation that can be made with regard to the underlying patterns of multiple meaning' and, as a consequence, 'blur the distinction between homonymy, non-systematic polysemy and systematic polysemy, and ultimately denies the existence of the latter.'[5]

2.4.2 One Representation Models

Nowadays, most scholars subscribe to a so-called *one representation model* of the mental lexicon. In one representation models, the 'senses of a polysemous expression either belong or depend on a single representation' (Falkum and Vicente, 2015). This line of thinking dates back to works like Nunberg (1979), who argued that there was

[4] Defenders of the model may distinguish between polysemy and homonymy based on whether the different senses or meanings belong to a single lexical entry - but ultimately both are stored as distinct representations (Falkum and Vicente, 2015).

[5] Dölling's observation links back to a concept sometimes called the *polysemy fallacy* as introduced by Sandra (1998), complaining that SEL approaches 'fail to distinguish between those aspects of meaning that are part of the word meaning proper, and those that result from its interaction with the context' (Falkum and Vicente, 2015).

no need to represent all interpretations of an ambiguous word in our mental lexicon; what needed to be stored was a core representation (conceding that it was entirely unclear what would be included in this core representation and how it would be derived; also see Caramazza and Grober, 1976; Miller and Johnson-Laird, 1976).

One representation models often are also called under-specification accounts, since - in contrast to SEL models - they do not require the full specification of all sense interpretations, but instead postulate a single, under-specified entry accessed for all interpretations of a polyseme. There are however different proposals concerning the question how much semantic information is stored in this representation, ranging from thin semantics models containing merely a set of constraints for what interpretations a word can take on, to rich semantics approaches that sometimes can postulate an over-specified core representation that makes all necessary information for all possible interpretations available at once.

Thin Semantics

In thin semantics models, the mental representation of a word is 'impoverished' compared to the meaning it can take on within a specific context (Falkum and Vicente, 2015), i.e. upon encountering a (polysemic) expression, an under-specified mental concept of its meaning is activated and subsequently enriched with relevant contextual information to form a specific interpretation. Thin semantics models often propose that the mental representation of a word is merely lexical, containing only information necessary to 'constrain the range of concepts that words can express' (Ortega-Andrés (2021), also see Travis, 2008; Falkum, 2011; Carston, 2013), or even that the under-specified representation is so thin that it carries no semantic content at all. Pietroski (2005) for example proposed that the mental representation of a word is simply a set of 'instructions for how to access and assemble concepts' (Ortega-Andrés, 2021), linking at or pointing to a number of concepts involved in its realisation.

When taking a thin semantics stance, mental representations of polysemous words are often brought back to Nunberg's core meaning approach, where 'the semantic representation of polysemous terms consists in a set of features or a common core that is shared by all senses' of that expression (Falkum and Vicente, 2015). This can best be explained by Jackendoff (1989)'s example of the verb *keep*, for which they postulate a mental meaning representation that simply states

(13) CAUSE [STATE OF X THAT ENDURES OVER TIME]

a core definition common to all interpretations, where X can take on different semantic values including possession, location or memory.

Rich Semantics

Rich semantics take the opposite approach to defining the mental representation of a polysemous word by postulating that all semantic information necessary to specify its different interpretations is available in the lexicon entry. One of the most prominent and influential rich semantics models is the so-called Generative Lexicon (GL) originally proposed by Pustejovsky (1993, 1995). The generative lexicon proposes that the lexical representation of meaning consists of four structures: an argument structure, an event structure, a lexical inheritance structure, and a qualia structure. The latter is the hallmark of Pustejovsky's model, designed to contain information on the roles that a word can fulfil in its different functions. This information includes aspects of 'about how the object came into being (its agentive role), what kind of object it is (formal role), what it is for (telic role) and what it is constituted of (constitutive role)' (also see Falkum and Vicente, 2015). As its name implies, in the generative lexicon word meaning is *generated* by accessing specific information from this over-specified lexical entry when encountering a target word in a specific context.

For unambiguous words, the information contained in the qualia structure decides whether a word is permissible in a given context, i.e. whether it fulfils the context's selectional restrictions. If on the other hand a word is polysemous, it can fulfil different selectional restrictions. According to Pustejovsky, this means that at least regular, or logical polysemous words must have complex qualia structures that allow for the selection of different roles in different contexts. In order specify these complex qualia structures, the generative lexicon postulates a special type, the so-called dot object (also see Asher and Pustejovsky, 2006; Asher, 2011). Dot objects represent polysemous expressions that combine at least two different senses into a single, under-specified type. Usually, these senses are inherent to the realisation of a concept, and could be described as facets or aspects of the complex type (cf. Cruse, 2004; Frisson, 2009; Paradis, 2004). As an example, the noun *book* would be represented as dot object **physical object•information** combining its realisation as a *physical object* and its *information* or *content* sense.

Based on this work, Arapinis and Vieu (2015) present a linguistic investigation of the notion of inherent polysemy, arguing that the kind of phenomena observed for targets like *book* or *country* might actually be grounded on specific ontological relations involving the entities referred to. Given the the observation that systematic polysemy is 'very productive and pervasive, mobilising general patterns of conceptual relatedness that structure our perception of the domain of reference' (see e.g. Apresjan, 1974), Arapinis and Vieu constitute that 'merely listing the senses inevitably fails to account for the conceptual or ontological mechanisms that trigger such multiple meaning phenomena.' Instead, they argue that items like *books* are complex 'materialised informational contents' that correspond neither to the conjunction of their disjoint aspects, nor to their disjunction. Proposing that the linguistic motivation for dot-types can have a direct ontological counterpart, they suggest to extend the notion of constitution beyond material coincidence to 'furnish the ontological counterpart of the semantic relation between each single type and the complex type or dot-type formed on them' (also see Arapinis, 2013).

As an example, Arapinis and Vieu suggest that an extended notion of coincidence can explain that *inflammation* denotes a complex **process•physical object** type through co-location of the anatomical structure and the process, or that *construction* may fail co-predication tests (cf. Ježek and Melloni, 2011) while being represented as a complex **process•result** object because 'the resulting object only comes into being after the process is over.' To introduce their extended constitution, the authors propose an alternative operator, the *general mereological sum operator* (+), which creates sums that will be 'filtered both according to the category of the entities summed and according to the presence of a coincidence relation between them, itself consequence of dependence relations' - addressing an earlier objection of Asher (2011) describing the mereological conception of dot-objects as 'fatally flawed.'

2.4.3 Literalist and Pragmatic Approaches

A third principled approach to modelling the mental processing of (ambiguous) words suggests that for each word, we store a single, 'concrete and semantically determined representation' (Falkum and Vicente, 2015), its 'literal meaning.' Once this literal meaning has been activated, a context-specific interpretation is derived either through a set of lexical rules, or through pragmatic modulation.

Among the literature explaining mostly regular polysemic alternations through a set of lexical rules applied to an initial literal interpretation (see e.g. Gillon, 1992, 1999; Kilgarriff, 1992; Ostler and Atkins, 1991; Asher and Lascarides, 2003), one of the most well-known proposals is Pelletier (1975)'s, and subsequently Copestake and Briscoe (1995)'s work on the 'universal grinder,' a model explaining *count/mass* alternations like the famous 'there was *rabbit* all over the highway' through a set derivation rules. But while gaining some attention in seminal formal and early computational semantics literature, rule-based literalist approaches have not received much support in recent years. One of the main reasons for this is that all rule-based approaches suffer from the limitation that they can only be applied to a small subset of the observed phenomena, and that even then they can be over-productive in some cases, requiring not only a formulation of derivation rules, but also a set of idiosyncratic exceptions to them. Falkum (2015) for example lists three theoretical arguments undermining fully rule-based approaches to polysemic sense extension as offered by Copestake and Briscoe (1995); Pustejovsky (1995). Firstly, it seems unclear how in a sentence like

(14) Peter enjoyed the nice weather.

the (assumed) intended reading of 'Peter enjoyed *being outside in* the nice weather' could be generated when 'there seems to be no telic information in the lexical representation' of *weather* that could be used as input in the compositional process deriving this interpretation. Secondly, the author argues that it is difficult to see how rule-based accounts can avoid making wrong predictions about many compositional interpretations, as for example in the VP *begin a car*, which according to the telic function of *car* should be interpreted as *begin driving a car.* And thirdly, they fail to see how when modelling (metonymic) polysemy entirely in terms of a lexicon-internal process, a rule-based approach can 'account for the interpretative flexibility that is arguably involved in its construction.' This concern is illustrated by sample sentences like

(15) a. Will a hamster bite if it smells *rabbit* on my hands? (*rabbit odour*)
b. [Biology teacher]: *Rabbit* is smaller than hare. (*rabbit faeces*)
c. [Hunter]: This time of year I prefer using *rabbit.* (*electronic rabbit calls*)
d. Last winter, we discovered *rabbit* and fox in our garden. (*rabbit tracks*)

where 'their one-off character makes it seem unlikely that any of them can be generated by a lexical rule.' Instead, Falkum favours a radical pragmatic account. Radical pragmatic approaches were common in early AI models, where all meanings would be generated via general commonsense reasoning - see e.g. Hobbs et al. (1993) - and are still favoured by many cognitive linguists, presenting an alternative to postulating rule-based derivations of contextualised interpretations from a literal meaning. As we briefly mentioned before, some scholars support the notion that basically every content word can be used polysemically. As this would entail an impossibly large number of senses or derivation rules that would be needed to be stored in our mental lexicon, pragmatic approaches suggest that we only store a single, fully conceptual representation of a word, and derive any contextualised readings pragmatically in an ad hoc fashion (see e.g. Recanati, 1998; Carston, 2002).

According to Traugott (2017), ‘a fundamental claim in cognitive linguistics is that words do not have fixed meanings. They evoke meanings and are cues to potential meaning, instructions to create meanings as words are used in context’ (also see e.g. Brugman, 1988; Kilgarriff, 1997; Paradis, 2011). As a consequence, radical pragmatic accounts ‘see the role of the linguistic system as being that of providing a minimal input or clue - a *sketch* or *blueprint* of the speaker’s meaning - which the pragmatic inferential system uses as evidence to yield hypotheses about occasion-specific, speaker-intended meanings’ (Falkum, 2015).

2.4.4 Hybrid Models

Falkum (2015) however also argue that while ‘overall, a radical pragmatic account provides the most promising basis for a unified account of the role of polysemy in several domains, [...] depending on their degree of conventionalisation, some senses may be stored in our mental lexicons, [and] some may be contextually derived.’ Returning to the *count/mass* alternation in *rabbit*, the authors therefore suggest that the input to the pragmatic processing of polysemes like this is composed of a rich, pragmatic representation of context and encyclopaedic information, and a highly under-specified conceptualisation of the target itself, which are combined to construct a narrower, *ad hoc* concept (e.g. *rabbit meat*). Some of these constructions like the *animal/meat* alternation of words like *rabbit*, *chicken* and *lamb* may become ‘progressively more routinised,’ developing ‘pragmatic routines’ (cf. Vega Moreno, 2007) that ‘increase the accessibility of certain interpretations and thereby contribute to a reduction of hearers’ processing efforts.’ These regularities then are proposed to give rise to the ‘sense of regularity’ observed in metonymic polysemes.

This account introduces a last addition to the range of mental models on the processing of ambiguous expressions, which we will preliminary label *hybrid models*. Hybrid models usually are based on one of the traditional mental models of language processing, but borrow some aspects of others. Klepousniotou et al. (2008) for example note that while their experiments in principle support a rich under-specification model, they also find that ‘high-overlap polysemous words differ from moderate- and low-overlap ambiguous words in comparison [and] there are several potential ways in which they may differ in representation,’ suggesting that a more structured representation of polysemic word sense might replace a fully under-specified core entry. Similarly, Asher (2011) presented a different version of hybrid model, suggesting that pragmatics are involved whenever a non-default interpretation is involved. This fall-back is intended to augment their originally over-specified one representation approach with pragmatic aspects for context coercion, but, while now allowing for these

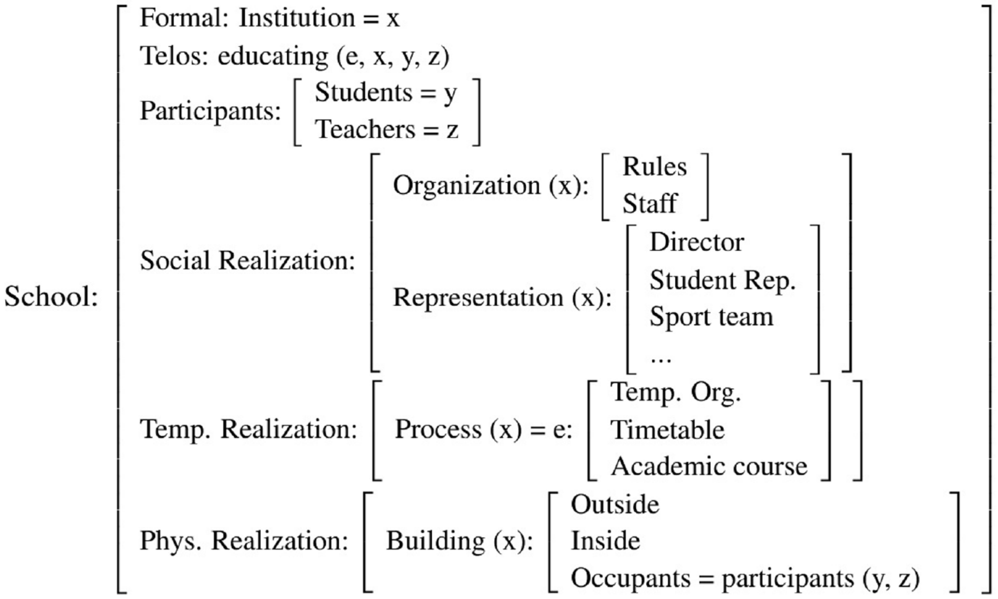


Figure 2.1: Schema of the knowledge structure of the polysemic realisations of word *school* according to the activation package model proposed by Ortega-Andrés and Vicente (2019). Figure replicated from ibid.

specific cases, does raise the question of when and how the fall-back is activated.

Finally, Ortega-Andrés and Vicente (2019); Ortega-Andrés (2021) recently proposed a hierarchical ordering within the under-specified representation of polysemic sense to allow a traditional Pustejovskyan model to account for processing differences among polysemic senses. Based on a rich under-specification account, Ortega-Andrés and Vicente extend a target's knowledge structure with multiple realisers that each specify a certain range of interpretations of the overall concept. Figure 2.1 shows a schematic of the hierarchical structure for polyseme *school* in Ortega-Andrés and Vicente's model. According to their hypothesis, the different interpretations that can be invoked by a given realisation (e.g. *rules* and *staff*) form so-called activation packages, groupings of interpretations that are so closely related to one another that an under-specified interpretation invoked by their realiser includes all of them simultaneously. This means that interpretations included in an activation package should allow for cost-free sense shifting, while moving to an interpretation evoked by a different realiser will lead to processing difficulties. Besides these explicit activation packages, a hierarchical representation like this however also implies an underlying notion of sense similarity which determines the representation - and consequently suggests at least different levels of similarity in the interpretation of polysemic senses.

2.5 Differential Processing: Linguistic Tests and Behavioural Evidence

One of the main drives behind Ortega-Andrés and Vicente's model was to explain processing differences that recently had been shown not just between polysemic and homonymic alternations, but also between different polysemic sense extensions. In this section, we will present some of the central literature investigating the differential processing in different types of lexical ambiguity in general, and specifically between different types of polysemic alternations. The experiments and data presented here often were produced alongside the theoretical models of language processing presented in the previous section, and each through their specific focus and setup was aimed at providing at least partial evidence in support of or in opposition to a specific model of the mental lexicon. We will here aim to evaluate both the established data and claims derived from it.

2.5.1 Co-predication Tests

While most of the initial motivation for distinguishing homonymy from polysemy or classifying different types of polysemic alternations comes from paradigmatic or anecdotal examples, literature also has produced a number of linguistic tests for this subject, including most notably the so-called co-predication test. Originally devised to determine identity of interpretation in ambiguous words (Zwicky and Sadock, 1975), Norrick (1981) was among the first to propose that co-ordination tests like in Example (16) could be used to test for complex polysemy, i.e. the activation of regular or possibly inherent polysemic sense extensions:

(16) The book was interesting$^{\text{INFO}}$ and weighed a ton$^{\text{PHYS}}$

Given that the co-ordinated structure here is acceptable even though it evokes two different senses of the target *book*, Norrick would consider this test to come down in support of a complex sense representation for the target word.

Co-predication now usually is defined as 'a grammatical construction in which two predicates jointly apply to the same argument' (Asher, 2011; Gotham, 2014) and used to test for (types of) polysemy in nominals (starting with e.g. Cruse, 1986). Murphy (2021) recently proposed to view co-predication tests more generally as a test for conflict, where 'semantic theories see co-predication as a conflict in type selection, whereas pragmatic and philosophical theories see it as a conflict in referential relations,' and reviewed various methods of constructing co-predication structures.

One of the most widely approaches however still is by *conjunction reduction* (Zwicky and Sadock, 1975), where two different interpretations of an ambiguous expression are combined into a single sentence by reducing a second, independent sentence into a conjunctive clause of another:[6]

(17) a. The *city* has 500,000 inhabitants.
b. The *city* outlawed smoking in bars last year.
c. The *city* has 500,000 inhabitants **and** outlawed smoking in bars last year.

Co-predication however is not limited to two senses only; if a word has multiple polysemic extensions, one could in principle generate a co-predication structure containing any or all of them as well (see Example (18) adapted from Ortega-Andrés and Vicente, 2019):

(18) Brazil is a large$^{\text{PHYS}}$ Portuguese-speaking$^{\text{CULT-LANG}}$ republic$^{\text{INST}}$
that scores very low in inequality rankings$^{\text{NATION}}$ but often leads
the FIFA ranking$^{\text{CULT-SPORT}}$

While authors such as Asher (2011) distinguish between logical and accidental polysemy by postulating that logical polysemy passes co-predication tests and accidental polysemy does not, and Ortega-Andrés and Vicente (2019) suggest the use of co-predication tests to tell apart inherent from other types of regular polysemy, linguistic tests in general are heavily context dependent, and can be made to yield inconsistent results by carefully manipulating these contexts (Geeraerts, 1993; Antunes and Chaves, 2003; Schumacher, 2013; Falkum, 2015; Murphy, 2021). Consider the following examples:[7]

(19) a. ? Judy's *dissertation* is thought provoking though yellowed with age.
b. Judy's *dissertation* is still thought provoking though yellowed with age.

(20) a. # They took the *door* off its hinges and walked through it.
b. The *door* was smashed in so often that it had to be bricked up.

(21) a. ? This *book* revolutionised the western world and is full of coffee stains.
b. That *book* is wrong about nearly everything it says about biology and full of coffee stains.

(22) a. # Mary fed and enjoyed the lamb.

[6] Example from Asher (2011)

[7] Examples from Norrick (1981), Cruse (1995), and Antunes and Chaves (2003), respectively. We will use question marks (?) to indicate questionable acceptability or sentence felicity, and hashes (#) to indicate arguably unacceptable or infelicitous structures in our examples.

b. Mary had fed the lamb herself and so she couldn't possibly enjoy it very much at dinner.

In each pairing, a slight modification of the predications involved in the co-predication structures - sometimes by simply adding a more descriptive context - can make an infelicitous or at least questionable sentence more acceptable - and vice versa. Dölling (2020) therefore note that 'it is apparent that co-predication may not only depend on the kind of pattern connecting word meanings but also on the discourse context and the rhetorical connections between the two predications.'

Additionally, evidence has been accumulating that acceptability judgements are not as objective as often assumed in literature. Collecting crowd-sourced acceptability annotations for textbook examples exhibiting grammatical inconsistencies, Lau et al. (2014) for example found that participants often rated acceptability differently than assumed in the original materials. When given a graded rating scale, grammaticality judgements also more resembled the results of a control study rating a shown character to be 'fat' or 'thin' rather than a second control rating them 'male' or 'female,' indicating that for many annotators grammatical acceptability seem to lie on a spectrum rather than representing a binary signal.[8] In order to mitigate at least the effects of subjective judgements, some studies have adopted a more empirical approach to investigating co-predication acceptability, either using corpus statistics or by collecting and aggregating annotations from a larger number of (layperson) raters. Following the first approach, Ježek and Vieu (2014) suggest that 'the *variability* of co-predication contexts is the key to distinguishing complex type nouns (i.e. inherent polysemy or polysemy proper) from nouns subject to coercion.' As an example, they argue that the *event* sense of *sandwich* in a sentence like (23)a is a result of coercion and not an inherent sense alternation, as the phrase 'during the sandwich' has far fewer (in this case zero) corpus occurrences than a related expression like 'during lunch.'

(23) a. Sam grabbed and finished the sandwich in one minute.
b. during lunch (780 corpus hits for the Italian equivalent)
c. during the sandwich (0 hits for the Italian equivalent)

To test their approach, the authors extracted all occurrences of [V [Det N Adj] patterns matching 28 selected nouns that could exhibit a *physical/information* alternation like *book* or *newspaper* from an Italian text corpus. The estimated recall of this procedure was reported at about 6%, and precision varied between 0% and 80% depending on the target noun. Among the collected sentences, Ježek and Vieu

[8] When assuming that gender is a binary construct that is.

(2014) found between 3% and 0.07% matches to be co-predications (i.e. having different type restrictions in the V and Adj elements).[9] From the list of target words, *lettera* (letter), *giornale* (newspaper) and *documento* (document) showed the highest ratio of co-predication vs. single predication matches (ratios 2.4%, 2.1% and 1.6%, respectively), and *pezzo* (piece), *prodotto* (product) and *fenomeno* (phenomenon) exhibited the lowest ratios (0.32%, 0.26% and 0.11%, respectively). As a result, the authors concluded that the *physical/information* alternation of the top lemmas is significantly more prevalent in their variability, and that these lemmas therefore are more likely to be representatives of polysemy proper than the latter ones.

While this corpus-based approach is a first step in capturing phenomena of polysemy empirically rather than based on a small set of seminal case studies, it is surprising that the authors focus exclusively on co-predication patterns to distinguish polysemy proper from coercion - as, like they seem to argue with the corpus results in Example (23), any single predication will already have type restrictions and the variability and other distributional information of these type restrictions could be used to investigate effects of regularity and prototypicality. Investigating selectional restrictions directly also would help to overcome the extreme sparsity of their data, and would be likely to give a more representative view of the true variability of the target nouns. As a side node, Ježek and Vieu (2014) also claim that *vino* (Italian for *wine*) does not appear to have an inherent *container* interpretation as 'it cannot be coerced into a container type by **any** predicate that would felicitously apply to *bottiglia* (*bottle*), as shown by Example (24)a. It is questionable how such a statement can be supported by a single example - especially given their corpus approach to other types of polysemy - and indeed there are plenty of predications that allow for a *container* type interpretation of *bottle*; one of them presented by the authors themselves just a few sentences prior.[10]

(24) a. # *Ho rotto il vino rosso.* (I broke the red wine.)
b. He opened the red wine.
c. He put the red wine back on the shelf.
d. He dropped the red wine on the floor.

More recently, Murphy (2019, 2021) collected annotator judgements on co-predication acceptability in a range of experiments aimed at investigating effects of sense ordering, complexity and coherence. Among their results, they observed significant effects

[9] The authors do not explain how the selectional restrictions were determined in this experiment.

[10] Ironically with the remark that 'a single occurrence of a relevant co-predication context is not enough to identify a complex type'. Note that here we only provide examples for English, it could indeed be the case that in Italian only *open* can elicit a *container* reading for *vino*.

of both sense order and sentence type on acceptability ratings, with for example stimuli invoking a concrete interpretation first and co-predicating an abstract second interpretation rated to be more acceptable than if these interpretations were presented in the inverse order. Based on their findings, the author suggests a theory of Incremental Semantic Complexity, stating that the language processor overall favours the presentation of input in ascending order of semantic complexity - which becomes explicit in co-predication. The author however also finds that co-predication acceptability depends on a wide range of other factors besides complexity, with samples not normed for frequency or controlled for coherence often failing to achieve significance. Co-predication acceptability thus should not be interpreted as a surefire sign of identity of sense or inherent polysemy, but as a complex signal illuminating some of the underlying mechanics of the language processor.

2.5.2 Eye-tracking and Event-related Potentials

Besides collecting data on the prevalence and acceptability of co-predication, a second central line of research contributing crucial inputs to the evaluation of mental language processing models have been behavioural studies. While for example acceptability judgements provide an off-line indicator of the processing of a certain phrase, behavioural studies focus on on-line effects as measured through e.g. reading times, eye-movements or brain activations, and often focus on investigating *when* and *how* an ambiguous item is interpreted. Together, the answers to these questions can reveal interesting insights into any differential processing between homonyms and polysemes, and, as a result, evidence for or against specific language processing models: if experiments were to find no notable differences in the participant's processing of homonymic and polysemic samples, the distinction should be considered purely theoretical and their mental processing can be assumed to proceed identically - which would support a sense enumeration approach to the mental lexicon. If on the other hand behavioural differences can be found in the processing of homonyms and polysemes - or specific types of polysemes - and some polysemic alternations were to facilitate processing compared to homonyms, behavioural insights could be used to support under-specified one-representation models.

Studies in Support of One Representation Models

Historically, (psycho-)linguistics literature produced a number of different - and oftentimes conflicting - semantic processing principles, including models of immediate and delayed semantic interpretation, completely-specified (maximal) and minimal commitments, and a so-called 'default assignment strategy' where a particular option is selected based on frequency or pragmatic plausibility (cf. Frazier and Rayner, 1990). As partial evidence had been presented in favour and against any of these models, Frazier and Rayner (1990) suggested that each of these processing principles could be involved in different aspects or elements of the language comprehension process, and proposed to re-formulate the central goal of language comprehension research to 'determine which class of decisions falls under which strategy, and why.'

The Immediate Partial Interpretation Hypothesis. Frazier and Rayner presented an eye-tracking study designed to investigate the validity of two general hypotheses labelled *Immediate Complete Interpretation* and *Immediate Partial Interpretation* hypotheses. According to the Immediate Complete Interpretation hypothesis, the processor 'maximises its immediate semantic commitments by interpreting each phrase fully as the phrase is encountered.' While a number of previous studies seemed to agree on their observation that semantic interpretation occurs rapidly (e.g. Crain and Steedman, 1985; Marslen-Wilson and Tyler, 1980; Just and Carpenter, 1980), Frazier and Rayner note that this does not imply that interpretations are necessarily complete. As a result, the Immediate Partial Interpretation hypothesis proposes that 'the processor may delay semantic commitments if this does not result in either i) a failure to assign any semantic value whatsoever to a word or major phrase, or ii) the need to maintain multiple incompatible values for a word, phrase or relation.' The authors suggest that this *partial specification* for example could occur when commitments have to be made for interpretations involving 'partially compatible specifications, as when two options overlap, sharing some but not all features'.

To test the applicability of these two hypotheses, the authors presented an eye-tracking experiment involving late disambiguation of expressions with two different meanings (i.e. homonyms), assumed to have two lexical representations, and words with two different senses (i.e. polysemes), arguing that the Immediate Partial Interpretation hypothesis implies that the processor will be forced to make a 'semantic selection' for meaning ambiguities, but not for sense ambiguities. Specifically, the different interpretations of a homonym are considered incompatible with one another and therefore cannot be maintained without selection, while different interpretations of polysemes sharing a sufficient amount of features do allow for a minimal commitment or partial interpretation 'since it will not result in either multiple analyses or failure to assign an analysis.'

Based on the data generated by 20 participants, Frazier and Rayner found that late disambiguation only led to increased reading times for samples containing homonymic targets, with the polysemic samples exhibiting reading times similar to the unambiguous controls - supporting the Immediate Partial Interpretation model. However, the authors also found that when a polysemic target was preceded by a disambiguating context, reading times were longer when it instantiated the dis-preferred reading as opposed to the preferred one, which suggests that 'readers commit themselves to a particular sense of a word when the intended sense is implied by the content of the prior context,' like they do for homonyms, too. In conclusion, Frazier and Rayner suggest that their results indicate that 'the shared or overlapping features of the various senses of the [polysemic] target are assigned as the initial semantic value or interpretation of the target phrase in the late disambiguation condition. In this condition, nothing in the prior context invites or requires additional specification of the interpretation of the phrase,' allowing for minimal semantic commitments when encountering the ambiguous expression without generating incompatible entailments.

Priming and Sense Dominance. Supporting approaches postulating a generative lexicon (e.g. Copestake and Briscoe, 1995; Pustejovsky, 1995) responsible for the ad-hoc generation of contextualised sense interpretations in the interpretation of lexically ambiguous items, Klepousniotou (2002) similarly suggests that processing times should differ between homonyms, where distinct senses need to be selected, and polysemes, where a 'basic semantic value' suffices to continue processing. Specifically, Klepousniotou expected that polysemes would be processed faster than homonyms, and that within the polysemic samples metonymic alternations should show larger priming effects than metaphoric ones due to the additional lexicalisation involved in metaphoric alternations. In their experiments, the authors presented participants with sentences priming either a more frequent (primary) or less frequent (secondary) reading of ambiguous targets divided into four types: homonyms, metaphoric polysemes (such as *eye* for human body organ or opening in a needle), metonymic polysemes (constituted solely of *count/mass* alternations like *chicken* referring to an animal or foodstuff) and a Name condition containing *author/work* alternations (such as *Dali* referring to the artist or one of his paintings). The priming sentence was followed either by a non-word, a target word (W) or a control item, where controls were words matched for overall corpus frequency (CFW) or the dominance of the primary reading (CAW). Participants were asked to judge whether a shown word is a real word of English by pressing a designated yes/no key on a keyboard, and reaction times (RT) were measured from the onset of the target.

Figure 2.2 shows the mean reaction times measured in their experiments. As

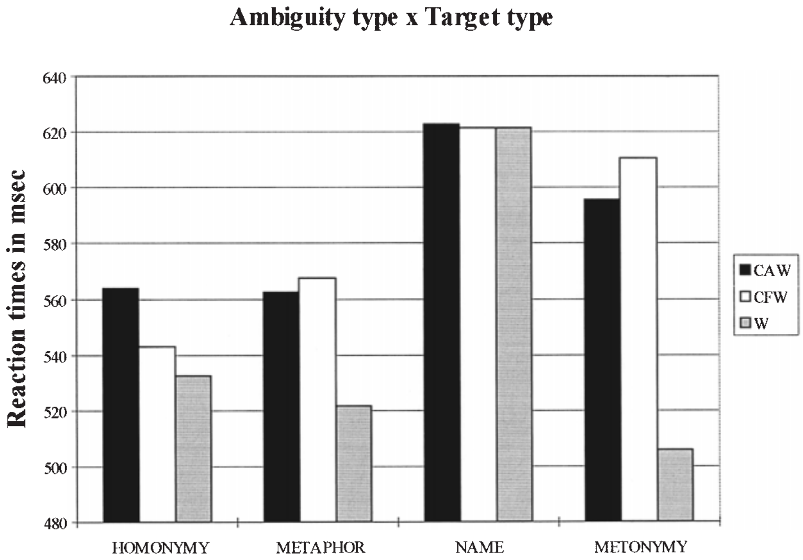


Figure 2.2: Mean reaction times (RT) for control ambiguity word (CAW), control frequency word (CFW), and ambiguous word (W) target conditions for each type of ambiguity. Figure from Klepousniotou (2002).

expected, reaction times were significantly faster for metaphoric and metonymic polysemes than for homonyms (grey bars).[11] Priming effects were strongest for the metonymic targets, with both target reaction times significantly lower than in the metaphor and homonymy conditions, and control reaction times much longer than for the other ambiguity types. Metaphoric polysemes not only led to significantly lower processing times than homonymic targets, but were also processed significantly faster than their respective controls. Homonymic targets were processed significantly faster than their ambiguity-matched controls, but no significant priming effect was found in comparison to the respective frequency controls. The author took this data to confirm their hypothesis that the processing of polysemic targets is faster than that of homonyms, and interprets their results as additional evidence against traditional SEL models. Instead, finding that metonymic polysemes provided an even larger under-specification advantage than their metaphoric counterparts, Klepousniotou suggests that their experiments support generative approaches where polysemic interpretations are constructed from a single, rich lexicon entry based on the contextual requirements.

They later however abandon this stance in favour of a thin semantics approach

[11] Noting that the data obtained from the Name condition revealed that the recognition of proper names appears to follow different processes altogether leading to longer processing times than in any of the other condition and no priming effects whatsoever, this condition was excluded from the analysis for this investigation.

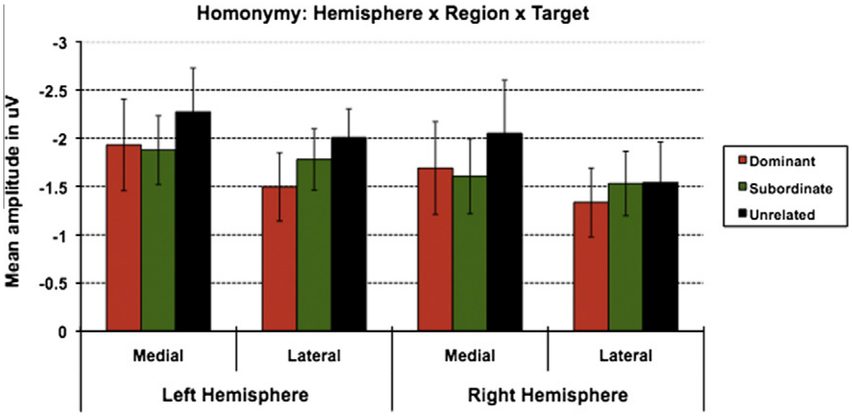


Figure 2.3: Hemisphere×Region×Target interaction for the dominant, subordinate and unrelated target conditions in Homonymy. Hemisphere (left/right) and Region (medial/lateral) are plotted on the x-axis; mean amplitude (and standard errors) (in microvolts) is plotted on the y-axis. Figure from Klepousniotou et al. (2012)

when presenting another range of experiments in Klepousniotou et al. (2012). Revisiting previous work using Electroencephalogram (EEG) data, the authors here used what they called *unbalanced homonymous* (*pen*), *balanced homonymous* (*panel*), *metaphorically polysemous* (*lip*) and *metonymically polysemous* words (*rabbit*) in a single-word priming delayed lexical decision task. Finding that the theoretical distinction between homonymy and polysemy was reflected in the N400 component, both balanced and unbalanced homonymous words showed priming effects with reduced N400 signals predominantly for dominant readings, while all polysemous primes lead to reduced N400 amplitudes for both readings. Assuming the N400 component to reflect lexical activation and semantic processing, the authors thus concluded that while homonyms are processed by directly selecting their dominant reading, polysemes (both metaphoric and metonymic) facilitate the selection of any of their alternative interpretations. In a similar fashion, Rodd et al. (2002) and Beretta et al. (2005) showed that words with more than one meaning (i.e. homonyms) were accessed more slowly than words with a single meaning (i.e. they elicited later M350 peak latencies and slower reaction times), and that words with many senses (i.e. productive polysemes) were accessed faster than words with fewer senses.

Based on these results, Klepousniotou et al. ultimately argue that when assuming a core meaning representation for polysemes, polysemous primes should lead to a faster processing of disambiguating targets invoking any of its alternative readings, while homonyms (with meaning represented in different entries) should only facilitate the processing of a target invoking the dominant reading.

Besides obtaining the previously summarised ERP data in line with their overall hypothesis, a significant Hemisphere×Region×Target interaction for homonyms also suggested that 'for dominant meanings the full set of the semantic representation (distributed across both hemispheres) is activated leading to more robust priming effects. In contrast, for subordinate meanings only a subset of the semantic representation (distributed predominantly over the left hemisphere) is activated leading to weaker priming effects' (see Figure 2.3). A similar effect was found for metaphorically polysemous words, too, but here only over the left hemisphere. As a result, processing advantages for dominant readings here were only minimal, and closely resembled those of metonymic primes. Addressing the ongoing debate on the processing and representation of metaphorically vs metonymically polysemous words, the authors thus suggest that 'it seems that metaphorically polysemous words do not have a fixed status in the lexical ambiguity continuum, but rather may be in a transition phase from generated senses to separately stored senses.'

Sense Frequency and Sense Shifting. Structuring their argumentation around polysemes like *book* that allow for a concrete and another, more abstract reading, Frisson (2015) showed through the results of two experiments that sense frequency had no apparent effect on sense switching costs - in contrast to the direction of switching, with especially switches from concrete to abstract interpretations leading to longer fixations on the ambiguous targets.

Arguing that both a traditional sense enumeration (SEL) account as well as relevance theory (RT) approach to polysemic word sense representation would suggest a frequency bias on sense interpretations visible in online reading experiments, they test this hypothesis using both reaction time and eye tracking methodology. In a first experiment, Frisson presented participants with two adjective-noun pairs, asking them for binary sensicality judgements but measuring reaction times. In this setup, neither sense frequency nor the order of sense shifting within a presented pair (from abstract to concrete meaning or vice-versa) had an effect on the participants' judgement reaction times. Acknowledging the stark difference between normal reading and this sensicality judgement task, they then continued by presenting a second experiment with full sentence stimuli resembling co-predication structures including both readings in different orderings. In this setup, they found that 'when a polysemous word was preceded by a neutral context, disambiguating towards either the dominant (abstract) or subordinate (concrete) sense was comparably easy,' as indicated by overall short and closely matched eye fixations on the target regions. When the ambiguous target was preceded by a disambiguating adjective, readers spent more time on the target region - without any notable differences between contexts selecting the primary or subordinate sense interpretation of the target. And if a stimuli introduced both interpretations of the target, processing seemed more difficult in general - but with a larger processing cost linked to switches form the subordinate to the dominant sense.

According to Frisson, both an SEL and RT inspired model would predict that given a neutral context, a reader assigns the most frequent sense to an ambiguous target. Given that no difference in processing time between dominant and subordinate interpretations was observed in either experiment, the author thus argues that the data does not support any of these approaches, but rather supports an under-specification account of polysemic sense representation where readers do not immediately select one of the available sense interpretations. And while both SEL and RT do predict a cost associated with sense switching, Frisson argues that finding a higher cost associated with switching from the concrete (subordinate) to abstract (dominant) is evidence against both of these accounts. Rather, they suggest that this effect can be ascribed to readers 'committing (cf. Frazier and Rayner, 1990) more strongly' to the concrete sense, impeding with a switch to a different (in this case the abstract) interpretation later.

ERPs for Container/Content Alternations. Providing empirical support for a primary interpretation in some metonymic alternations, Schumacher (2013) focused on *container/content* vs *animate/inanimate* alternations to show that while some metonymic extensions involve cost-free meaning selection, others 'engender processing costs associated with re-conceptualisation,' proposing that these targets have an 'original' meaning and a set of contextually appropriate ones derived from it. Given the sometimes contradictory conclusions drawn from previous research (e.g. Frisson and Pickering, 1999, 2001; Frisson and Frazier, 2005), Schumacher proposes to carefully distinguish different types of metonymic alternation: those that have a fully under-specified representation of alternate senses in the mental lexicon, and those that have an inherent meaning that allows for the derivation of contextualised interpretation shifts. Starting with *container/content* alternations, they propose that there are 'relations between ontological types - such as liquids can be contained in physical objects - whose specification is determined by encyclopedic knowledge, and these relations are made available during compositional processing to induce a meaning shift (cf. Copestake and Briscoe, 1995; Dölling, 1995).' Due to the asymmetry in the relation between container and content, the *content-for-container* reading is assumed to rely on 'the application of a general lexical derivation rule, while *container-for-content* interpretations use a variable form the expression's qualia structure.'

Using German question-answer pair stimuli like those in Example (25) asking for the target with a specific restriction on its interpretation, Schumacher tracked participants' event related potentials (ERP) through an EEG experiment. Analysing the grand-average ERPs, they found a more pronounced positive deflection between 550-750ms and between 900-1100ms in the critical region of *container-for-content* than in their controls, but no statistically significant difference in the ERPs of *content-for-container* alternations and their controls. These findings were also mirrored in a pre-test as well as in a post-EEG test asking participants to rate the samples' plausibility, which revealed no differences between *content-for-container* samples and their controls, but reliably lower plausibility for *container-for-content* items than their controls.

(25) **container-for-content**
Was hat Heinz hastig getrunken?
Er hat **den Becher** hastig getrunken.
(What did Heinz drink quickly? He quickly drank **the cup**.)
control
Was hat Rolf wie seinen Augapfel gehütet?
Er hat **den Becher** wie seinen Augapfel gehütet.
(What did Rolf guard jealously? He jealously guarded **the cup**.)
content-for-container
Was hat Asterix an seinem Gürtel festgeschnallt?
Er hat **den Zaubertrank** an seinem Gürtel festgeschnallt.
(What did Asterix fasten to his belt?
He fastened **the magic potion** to his belt)
control
Was hat Miraculix vor dem Eintreffen der Römer gebraut?
Er hat **den Zaubertrank** vor dem Eintreffen der Römer gebraut.
(What did Miraculix brew before the Romans arrived?
He brewed **the magic potion** before the Romans arrived.)

The author proposes to explain the observed differences between the two alternations' acceptability scores as well as processing demands by suggesting a close, natural ontological relation between substances and their respective container (i.e. liquid substances need to be contained in something to be handled). This ontological relation manifests itself in a (prototypical) container becoming available for reference free of processing costs after a liquid is introduced in discourse. The *content* reading for a container on the other hand only is available through a selection from its qualia structure - in this case by shifting to its telic role.

A similar observation was made in a second EEG experiment using sample sentences containing adjective-noun pairs with matching or mismatching animacity. Here they observed an enhanced positivity over posterior electrode sites for mismatched used (e.g. the wooden turtle) over the literal use (e.g. the wooden trunk) between 550-750ms, while the comparison involving animacity-neutral adjectives (e.g. grey dove vs. grey shirt) registered no differences. With an even more clear distinction between primary and derived interpretations, this was taken as additional support for the assumption that late positivity effects are linked to re-conceptualisation, and that therefore *container-for-content* alternations require re-conceptualisation, while *content-for-container* readings do not.

Studies in Support of Sense Enumeration Approaches

Not subscribing to an under-specific one representation approach to the mental representation of polysemes, Klein and Murphy (2001, 2002) and Foraker and Murphy (2012) are among the few to present experimental evidence in favour of sense enumeration approaches. In Klein and Murphy (2001), the authors first introduce a range of five experiments using word pairs like *shredded paper* and *liberal paper* in memory and sensicality judgement tasks to test differences in processing between phrases eliciting the same or different readings. In the first experiment, participants were shown a series of word pairs which they were asked to remember, and subsequently presented test items displayed like *daily PAPER* for which they were asked to decide as quickly as possible whether they had seen the highlighted target word in the previously shown list. Test items all were chosen to be polysemous and divided into three conditions: in the *same phrase* condition, the exact same word pair shown in the initial list was repeated as a test item. In the *consistent sense* condition, the target was shown in combination with a different modifier, which still elicited the same interpretation of the target as in the original phrase. In the *inconsistent sense* condition, a new modifier invoked a different sense of the target than that elicited by the memorisation list. The authors found that same phrase items were judged most accurately (at 79% correct),[12] followed by consistent phrases (64% accuracy) and inconsistent phrases (56%).

A second experiment re-used the same materials, but instead of asking partic-

[12] All test items were indeed shown in the memorisation list, so the correct answer for all test items was *yes*

ipants to remember a list of word pairs, participants were instructed to rate the sensicality of a displayed phrase as quickly as possible. Phrases with the same target word were always shown in succession in order to re-create the *consistent sense* and *inconsistent sense* conditions as in the first experiment. As an example, a phrase like *shredded paper* would be followed by a phrase like *wrapping paper* to create a *consistent sense* condition. In this experiment, the authors measured accuracy and reaction times on the judgements for each second item. Consistent phrases again were rated more accurately than inconsistent phrases (96% accuracy vs. 87%), but reaction times only were reliable in a per-item analysis, where the target was rated more quickly if the prime was consistent.

A third experiment repeated experiment two with targets made up of an equal amount of polysemes and homonyms. The authors here found no significant interaction between consistency and ambiguity type, but consistency again was a reliable factor of judgement accuracy, with consistent phrases being evaluated 85ms more quickly and 12% more accurately than inconsistent phrase pairs.

The fifth experiment finally was designed to test whether priming effects were inhibitory or facilitory, i.e. whether the primes of inconsistent phrases suppressed the interpretation of succeeding inconsistent phrases, whether primes of consistent phrases simplified their interpretation - or whether both effects were active. To test this, Klein and Murphy introduced neutral primes consisting of a blank line followed by the target. Re-using the procedure of experiments two and three, this setup revealed that neutral primes lead to faster processing times of the target in the consistent condition (832ms) than in the neutral (879ms) and inconsistent (938ms) conditions. They also were judged correctly more often in the consistent condition (96% accuracy) compared to neutral (89% correct) and inconsistent (84% correct) phrase pairs. With the neutral condition leading to reliably slower and less accurate judgements than the consistent phrase pairs, but faster and more accurate reactions than for inconsistent pairs, the authors suggest the presence of both, facilitory and inhibitory effects in their priming experiments.[13]

Taken together, the authors suggest that these experiments indicate that there are no signs of significant processing differences between homonymic and polysemic targets, and that invoking different (polysemic) senses - like different homonymic meanings - requires more processing and leads to more inaccurate judgements than

[13] Experiment four was focused mainly on controlling for the modifiers, testing that the modifiers were not priming each other so that observed differences were independent of the target words. The authors however didn't find any reliable priming in the modifiers of the consistent phrase targets that could explain for the previous observations.

when invoking the same interpretation. All of these observations favour a sense enumeration approach, as inconsistent polysemic senses seem to be no more accessible after priming than homonymic meanings are, and priming an inconsistent sense does not facilitate the interpretation of a target in the same way as a consistent prime does. Still maintaining that polysemes can - and should be - distinguished from homonyms on a theoretical basis, Klein and Murphy (2001) therefore offer the notion that different senses of polysemous words are *related* but not *similar*.

In a follow-up series of forced choice experiments, Klein and Murphy (2002) tested whether participants are more likely to group together polysemous targets invoking different interpretations or unrelated words that fulfil the same conceptual or thematic role. In a first experiment, participants saw a target phrase like *wrapping PAPER* and were given two options: *liberal PAPER* and *smooth CLOTH*. In this example, the first option contains the same target word, but the modifier elicits a different interpretation that the *material* sense in the target phrase. The other option presents a different word, which however matches the taxonomic category of the target *paper*. Non-polysemic options were always either matching the taxonomy or the theme of target, with a phrase like *sharp SCISSORS* being an example for a thematic option for target *wrapping PAPER*.

Forced to choose between the two options, participants selected the polysemic option in only 20% of the cases, independent of whether the other option was a category or thematic match. A second experiment included polysemic options that invoked the same sense as the target phrase - which were selected in 70% of the cases - and a third included homonymic options, which were selected only 8% of the time. Given these results, the authors suggest that readers do distinguish between different interpretations of a polyseme, again supporting a sense enumeration approach to their mental representation.

Additional evidence for this hypothesis comes in the form of an eye-tracking study presented in Foraker and Murphy (2012), where participants read late disambiguation sentences invoking dominant or subordinate interpretations of a target polyseme. The target was introduced either in a context more closely related to a dominant reading, a subordinate one, or a neutral one, creating a total of six conditions. Using a subset of the same target words as in Klein and Murphy (2001, 2002), the authors here found that - much like traditionally shown for homonyms (see e.g. Simpson, 1981; Tabossi et al., 1987) - the biased introductions lead to significantly longer fixations on the disambiguating region and increased overall reading times when matched with an inconsistent disambiguation, while the neutral context did

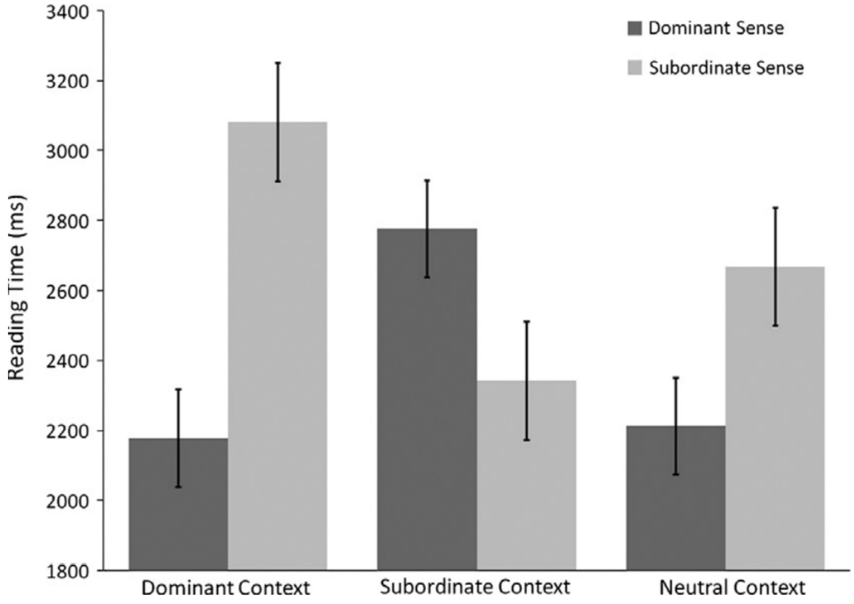


Figure 2.4: Reading times for the disambiguating target sentences in Foraker and Murphy (2012)'s first experiment. Figure replicated from ibid.

lead to comparable reading times for dominant interpretations but slower reading times for subordinate senses (see Figure 2.4). These results again can be taken as evidence that polysemes - much like homonyms - have primary and secondary interpretations, and readers automatically assume the dominant reading on encountering a polysemous target - which in turn indicates that different senses are likely to be represented individually.

The experiments by Klein and Murphy (2001, 2002) and Foraker and Murphy (2012) however are not unchallenged, and in fact have been criticised by a range of subsequent work for their methodology - and especially their materials. Klepousniotou et al. (2008) for example repeated the experiments presented in Klein and Murphy (2001) using ambiguous targets classified as either metonymic polysemes, metaphoric polysemes or homonyms based on either high, moderate or low semantic overlap between their primary (dominant) and a secondary interpretation. This was as a direct reaction (and objection) to the materials used by Klein and Murphy, which according to Klepousniotou et al. contained targets labelled as polysemes that according to a distinction by sense overlap should in part be considered as metonymic or metaphoric polysemes, and to some part as homonyms (e.g. *nail*).[14]

[14] Klein and Murphy (2001) themselves noted that they chose 'senses of words that were fairly distinct, so that we could select each sense with a single word' and picked polysemous words from a pre-established list, but indeed many of them would not be classified as polysemes according to most modern definitions.

Targets in this revised study were highly controlled - matched for meaning dominance, corpus frequency, length in letters and their transitional probability - in order to establish a well-defined methodology intended to provide definite results to supersede the previous inconsistent data. As in the original experiments by Klein and Murphy (2001), participants were shown groups of two word pairs sharing one word, where in the test condition the first word pair would prime the dominant sense (e.g. *marinated lamb*), the subordinate reading (e.g. *baby lamb*) or contain a neutral prime (e.g. ****** lamb*). Participants were asked to rate whether a word pair made sense as quickly as possible, getting feedback after each decision.

When the prime and target pair invoked the dominant interpretation, reaction times on the target pair were significantly faster for ambiguous words with low and moderate sense overlap (i.e. homonyms and metaphoric polysemes). High-overlap (metonymic) targets did not show this effect. Here cooperating and conflicting primes lead to very similar reaction times (788 and 783ms, respectively), both significantly faster than for neutral primes (848ms).[15] For subordinate target pairs the word type×context interaction was not significant, but reaction times were numerically faster for high-overlap targets (846ms) than for both moderate (873ms) and low-overlap target (884ms). Also investigating response accuracy, the authors reported mirrored results: While dominant targets here showed no significant word type×context effects, accuracy for low-overlap words was significantly higher for subdominant targets with matching contexts (97%) than for neutral contexts (88%) or conflicting contexts (79%).

While these results disagree with those of Klein and Murphy (2001) and instead agree with the findings of investigations like their own 2002 study showing metonymic polysemes are processed significantly faster than homonyms (also see Frazier and Rayner, 1990; Frisson and Pickering, 1999; Pickering and Frisson, 2001), the data of this experiment places the processing times of metaphoric polysemy closer to that of homonyms than metonymic polysemes. So while they take their results to show that ‘high-overlap polysemous words differ from moderate- and low-overlap ambiguous words in comparison; nevertheless, there are several potential ways in which they may differ in representation,’ and propose that ‘further work should determine whether more subtle differences in representation and process exist for moderate- and low-overlap ambiguous words.’

[15] The ‘neutrality’ of the neutral primes however was questioned by the authors themselves, acknowledging that in contrast to the original experiments, their asterisks fillers may have been more visually complex and distracting

2.6 Summary

In this chapter we explored the concept of lexical ambiguity and reviewed different approaches of subdividing its manifestations. The theoretical linguistics literature often makes a principled distinction between homonymy and polysemy, teasing apart different word meanings and word senses. Anchoring this distinction in sense relatedness or historic commonality however is challenged by a number of philosophical and linguistic objections, and attempting to show principled processing differences through linguistics tests or behavioural evidence often had to concede that the investigated phenomena were more diverse than anticipated and that the resulting data did not allow for as clear a distinction as aimed for.

Part of the inconsistency in observations of polysemy often is ascribed to the heterogeneity of this type of lexical ambiguity, with a number of scholars presenting different approaches to sub-divide or classify specific polysemic sense alternations. Regularity in metonymic or inherent alternations is often contrasted to the idiosyncrasy of metaphoric polysemic alternations, and sense similarity or relatedness as a result can be seen to span a gradient between identity of sense on the one end, and multiplicity of meaning on the other.

In order to explain differential processing, a range of models of the mental lexicon have been proposed. The most commonly accepted of these models is the so-called one representation approach, which postulates that the different sense interpretations of a polysemic word are stored in a single entry in the mental lexicon. In recent years, a number of additions and revisions were proposed to re-conciliate this model with the expanding empirical evidence related to the online processing of ambiguous word forms, most notably the activation package model proposed by Ortega-Andrés and Vicente (2019), which postulates a more structured representation of polysemic sense driven by an underlying notion of sense relatedness or sense similarity. One of the main objectives of this thesis is to further explore this notion of sense similarity in an attempt to produce empirical evidence contributing to the ongoing discussion.

Chapter 3

Computational Approaches to Word Sense Representation

Having explored (psycho-)linguistic approaches to lexical ambiguity in the previous chapter, in this chapter we will introduce the computational linguistics field of Distributional Semantics, which can be seen as reverse-engineering the language production process by inferring word meaning from large-scale corpus data. We will touch upon seminal literature applying distributional semantics to investigating phenomena of lexical ambiguity (Section 3.1) and introduce Word2Vec, one of the most popular techniques for creating static word embeddings (Section 3.1.1). We will then briefly review word sense disambiguation tasks and some of the most successful approaches for them (Section 3.2), and introduce the concept of graded word sense assignments (Section 3.2.1), which question the application of hard gold standards in the evaluation of word sense disambiguation in line with the observations of behavioural studies presented in the previous chapter.

Finally, in Section 3.3, we will introduce a current generation of so-called Contextualised Language Models that aim to represent the meaning of a word by creating ad hoc vector encodings unique to a specific context. We will briefly show how these models have put their stamp on NLP research by producing state of the art performance scores across a wide range of tasks, while at the same time providing limited explainability. In chapters 4 and 5 we will use contextualised language models ELMo and BERT to encode samples containing different types of polysemic words, aiming to offer some insights into how the differences in the models' representations of ambiguous words correlate with human judgements on different measures of word sense similarity. Finding that especially BERT Large seems to capture at least some aspects of word sense, in Chapter 6 we will experiment with BERT to automatically identify corpus samples exhibiting specific patterns of polysemy.

corpus-based

	dim1	dim2
cut	4	5
cost	1	5
cut cost	4	9

synthetic

	dim1	dim2
CUT COST	5	10

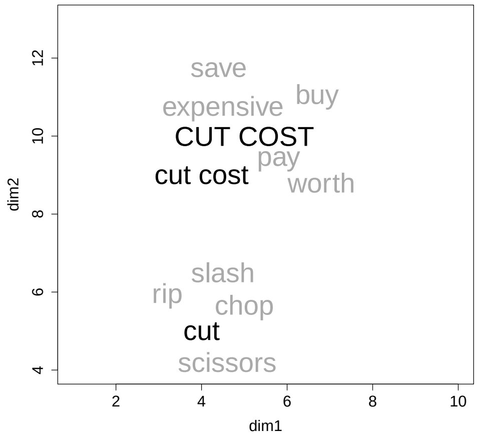


Figure 3.1: Compositional distributional semantics: Illustration with vector addition. Left: The synthetic vector *CUT COST* is built by component-wise addition of the vectors for *cut* and *cost*. Right: The argument *cost* pulls the vector for *cut* towards its abstract use (see nearest neighbours, in grey). The corpus-based vector for *cut cost* can be used to check the quality of its synthetic counterpart. Figure and caption text from Boleda (2020).

3.1 Distributional Semantics

Distributional semantics is based on the Distributional Hypothesis, the assumption that 'similarity in meaning results in similarity of linguistic distribution' (Harris, 1954; Firth, 1957; Erk, 2012; Clark, 2015; Lenci, 2018). In other words, distributional semantics reverse-engineers the language production process by inducing semantic representations from contexts of use. This usually is done by abstracting words and their contexts to vectors in semantic space and measuring the similarity between the vectors of given target expressions.

In traditional approaches to distributional semantics, each word is assigned a single vector, resulting in an abstraction over all its contexts of use, and thus theoretically 'encompassing all the word senses that are attested in the data' (Arora et al., 2016; Boleda, 2020). One way to represent and investigate polysemy under these traditional approaches is by composition, and, in its simplest form, vector addition. Boleda (2020) presents the following example to illustrate compositional distributional semantics: Consider a two-dimensional semantic space in which words are represented, as depicted schematically in Figure 3.1 for ambiguous word *cut*. The overall representation of *cut* is dominated by its physical reading (e.g. *cut the bread*) encoded at (4|5), as attested by its nearest neighbours including *rip*, *chop*, and *scissors* (in grey). *Cut* however also has a more abstract reading, as in *cut costs*. In this example, *cut costs* would be located at (4|9) according to its abstraction from the corpus data. Since deriving all of these contextualised representations from the corpus is infeasible, synthetic representations can be generated instead, in this case by combining the elementary representations of *cut* and *cost.* In our example, the combination of *cut* at (4|5) and cost at (1|5) through vector addition results in a synthetic $CUT\ COST$ representation at (5|10) - relatively close to its corpus-based embedding at (4|9).

Using this vector addition to derive synthetic representations is not too dissimilar to generative processes like those proposed for example in the Generative Lexicon (Pustejovsky, 1995, see Section 2.4.2), with one crucial difference: while the representations of the Generative Lexicon are explicitly structured to allow for the composition of derivative interpretations, in distributional representations the structure is implicitly defined in the space (Boleda, 2020).

In general, Boleda (2020) suggest four ways for distributional semantics research to contribute to linguistics: i) by exploring language data at large scale, ii) by identifying instances of specific language phenomena, iii) as a testbed for linguistic hypotheses, and iv) by aiding the discovery of linguistic phenomena or theoretically relevant trends in data. Experimental investigations of composition methods for example include works like Baroni and Zamparelli (2010); Boleda et al. (2013) and Mitchell and Lapata (2010), where phrase similarity predictions derived from the best composition methods reach Spearman correlation scores with participant data of around 0.4, and (Grefenstette and Sadrzadeh, 2011; Bentivogli et al., 2016) who found that deriving suitable representations of distransitive constructions proves extremely difficult.

3.1.1 Static Word Embeddings

In 2013, Mikolov et al. (2013a,b) presented a new approach to representing words in vector space using their distributional information: word embeddings. Observing that much of the complexity in traditional Feed-forward Neural Net Language Models (NNLM) and Recurrent Neural Net Language Models (RNNLM) stems from the non-linearity in their hidden layers, they proposed two new, log-linear approaches to processing large amounts of corpus data while deriving word representations: Continuous Bag-of-Words (CBOW) and Skip-grams. Using a sliding window determining a target word and a context, with these techniques word embeddings are learned as input to a classifier predicting the probability of the target co-occurring within a given (past and future) context (CBOW), or the probability of a target being surrounded by the selection of context words (Skip-gram), and negative sampling was added to prevent the model from returning perfect probabilities for all proposed combinations (which would yield an initially impressive but ultimately meaningless 100% accuracy). Originally trained on the 6B token Google News corpus with a vocabulary consisting of the 1M most frequent tokens, their approach called Word2Vec displayed promising arithmetic features, like a relatively stable relation between the embeddings of country names and their capitals (e.g. vector(*Madrid*) - vector(*Spain*) + vector(*France*) is closer to vector(*Paris*) than to any other word), and the famous observation that vector(*King*) - vector(*Man*) + vector(*Woman*) results in a vector that is very similar to the vector representation of the word *Queen* (Mikolov et al., 2013c).

A year later, Pennington et al. (2014) presented GloVe (**Glo**bal **Ve**ctors), trained on the 'non-zero entries of a global word-word co-occurrence matrix,' which, according to the authors, provides the 'benefit of count data while simultaneously capturing the meaningful linear substructures prevalent in recent log-bilinear prediction-based methods like Word2Vec.'[1] Being competitive in embedding quality and difficult to compare in terms of training efficiency,[2] both models have been equal contenders for a range of NLP applications in academia and industry in the following years.

3.1.2 Word Sense Embeddings

A crucial shortcoming of any static word embedding approach is that each word is represented by a single vector, which therefore should represent all possible meanings and senses a word can elicit. An alternative approach to representing multiplicity of meaning or even multiplicity of sense explicitly has been proposed in building sense-specific distributional representations, i.e. deriving one vector for each possible interpretation of a word (for a recent survey see Camacho-Collados and Pilehvar, 2018). Sense-specific representations can be generated by encoding only those contexts that invoke a given interpretation of the ambiguous target word, based on the assumption that different uses of a word are reflected by different contexts (see Pedersen and Bruce (1997) and Schütze (1998) for some of the earliest investigations of this approach, and McCarthy et al. (2004); Almuhareb and Poesio (2006); Erk and Padó (2010); Reisinger and Mooney (2010) for more recent contributions). As mentioned in Section 2.4.1, already Kilgarriff (1997) however voiced two principled objections to any sense-based approach: the theoretical difficulty in 'deciding when two senses are different enough to warrant a new entry, and how to represent the

[1] also see `https://nlp.stanford.edu/projects/glove/`

[2] Training efficiency also is a relatively minor factor in this case as both models only need to be run once to provide their static, pre-trained word embeddings that usually are made available online

information that is common to multiple different senses.' Likewise, Hanks (2000) questions whether different senses actually can be represented as disjoint classes defined by necessary and sufficient conditions.

3.2 Word Sense Disambiguation

One of the most common approaches to evaluate distributional models is their performance on Word Sense Disambiguation tasks (WSD). In Word Sense Disambiguation, a model has to select that entry of a provided sense inventory which best represents the meaning of a word in a given context sentence. WSD was developed as one of the aspects of early approaches to automatic machine translation (see e.g. Weaver, 1955) and has been classified as an AI-complete problem (Navigli, 2009). Overall, approaches to WSD either are supervised or knowledge-based, i.e. utilise sense-annotated corpora for training a model (cf. Zhong and Ng, 2010; Iacobacci et al., 2016) or exploit the structure of the provided reference knowledge resource to derive classification rules (cf. Lesk, 1986; Banerjee and Pedersen, 2002; Moro et al., 2014).

The de-facto standard sense reference for WSD is WordNet (Miller et al., 1993; Miller, 1995), which organises English nouns, verbs, adjectives and adverbs into sets of synonyms. Each synonym set is taken to represent a lexicalised concept, and different synonym sets are linked through semantic relations. Annotated by experts and treated as a gold standard for word sense, WordNet however also does not go unchallenged: its inter-annotator agreement (IAA) for example ranges between only 67% and 78% (Fellbaum and Miller, 1998; Mihalcea et al., 2004; Snyder and Palmer, 2004) 'depending on factors such as degree of polysemy and inter-relatedness of the senses' (Erk et al., 2013), which indicates a number of disagreements on sense classifications even among the expert annotators. Similar IAA levels can be found for alternative sense references such as the SemCor and SensEval datasets. Notable exceptions are the SALSA annotations based on FrameNet (Burchardt et al., 2006) and OntoNotes (Hovy et al., 2006), where (also) due to a more coarse annotation scheme annotator agreement often ranges at and over 90%. Erk et al. (2013) on the other hand acknowledge that determining the right level of granularity for the annotation of a WSD task is an important facet of improving model performance, but argue that a theoretically more interesting approach is to explore 'novel annotation tasks that allow us to probe the relatedness between dictionary senses in a flexible fashion, and to explore word meaning in context without presupposing hard boundaries between usages.'

3.2.1 Graded Word Sense Assignment

Erk et al. (2009) therefore make a case for graded word sense assignment, citing the limitations (e.g. Cruse, 2000; Hanks, 2000) and previously mentioned disputed applicability (e.g. Tuggy, 1993; Kilgarriff, 1997, 2001) of discrete sense boundaries in pursuing this deviation from established methodology. With graded annotations, if during annotation a word usage is assigned different sense interpretations, instead of selecting a single sense label through some form of aggregation, a graded sense assignment is established based on the distribution of annotations.

In a pilot, Erk et al. collected two types of graded annotations: WSsim (Word Sense Similarity) and Usim (Usage Similarity). In the first experiment, three participants rated the applicability of different WordNet sense interpretations to a given use of a target word in a context sentence. For each sense, annotators were asked to select the degree to which that sense applied to the presented use. Annotations were collected for a total of 11 lemmas presented in a grand total of 430 context sentences, most of which were randomly sampled from the SemCor (Fellbaum and Miller, 1998) and SenseEval-3 (Mihalcea et al., 2004) corpora, with three lemmas being assigned sample sentences from the LEXSUB data (McCarthy and Navigli, 2007).

In the second experiment, another three participants rated the similarity between usages of the same target word displayed in two context sentences. This experiment covered 34 lemmas selected from the LEXSUB data, including the three lemmas selected from that corpus in the WSsim experiment. For each target word, 10 context sentences were sampled from the LEXSUB data, and each possible combination of context sentences was included in a list of sentence pairs (SPAIR) to be annotated. Annotators here were given the following instructions: ‘Your task is to rate, for each pair of sentences, how similar in meaning the two boldfaced words are on a five-point scale,’ with the scale being labelled as

(26) 1 - completely different
2 - mostly different
3 - similar
4 - very similar
5 - identical

Analysing the resulting annotations, Erk et al. found that in the WSsim judgements the extreme labels 1 and 5 were applied significantly more often than the intermediate values, with label 1 (lowest degree of word sense applicability) making up the lion’s share of these ratings. Annotators nevertheless used intermediate an-

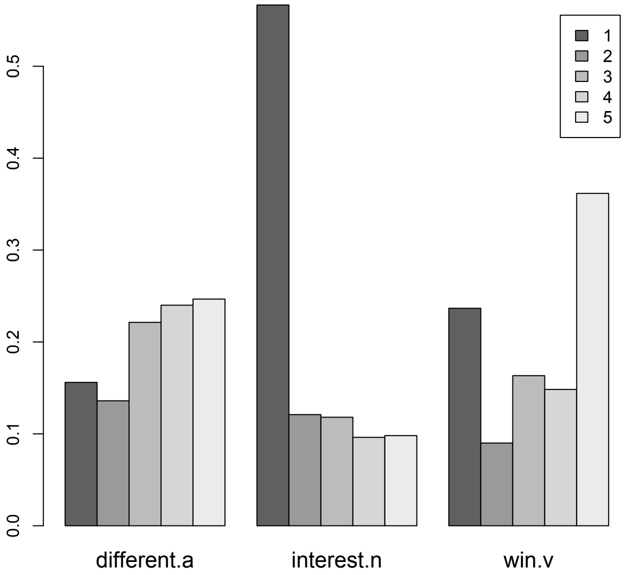


Figure 3.2: Distribution of judgment labels for lemmas *different.a*, *interest.n* and *win.v* in Erk et al. (2009)'s WSsim experiment. Figure replicated from ibid.

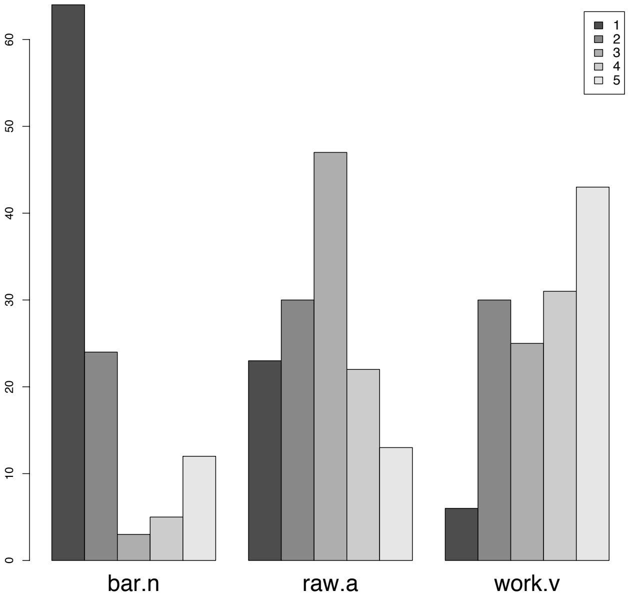


Figure 3.3: Distribution of judgment labels for lemmas *bar.n*, *raw.a* and*work.v* in Erk et al. (2009)'s Usim experiment. Figure replicated from ibid.

notations, and distributions of labels differ significantly between targets (see Figure 3.2). In the Usim annotations, the authors found that annotators used intermediate labels even more often than in the WSsim setting, with only label 1 for completely different uses being assigned significantly more often than others. Here, too, interesting judgement distributions can be observed for different target words: *bar.n* for example receives mostly low Usim ratings, while *work.v* pairs are assigned mostly high similarity scores, and *raw.a* exhibits a peak for mid-level ratings (see Figure 3.3).[3]

Besides showing that - when given the option - annotators do use graded ratings when judging word sense applicability and word use similarity, Erk et al. (2013) also found that the collected Usim annotations obey the triangle inequality. In Euclidean space, the 'lengths of two sides of a triangle, taken together, must always be greater than the length of the third side.' When checking the Usim annotations of sentence triplets with the same target lemma against this principle, they found that over 99% of comparisons did comply with the triangle inequality, indicating that the space spanned by USim annotations indeed is metric and allows for meaningful arithmetic operations.[4]

Finally, given the observation that word sense assignment empirically does not seem to be a binary decision, in Erk and McCarthy (2009) the authors propose a range of evaluation criteria for graded WSD, including (non-parametric) correlation with the human annotations (including proposals to calculate this by lemma, lemma and sense, and lemma and sentence), Jensen-Shannon Divergence, and precision and recall metrics. Here they also propose a simple prototype for a word sense representation through a vector space model, built based on co-occurrences of context words in a window of size 50. This prototype - with different sets of parameter settings - provides promising results on the different correlation and recall and precision metrics when tested on the annotated data.

Datasets of Graded Sense Similarity

Large-scale datasets that capture graded similarity judgements usually do so for word pairs in isolation - often intended to evaluate static word sense embeddings (also see Taieb et al., 2019). Until recently, the only exceptions included the Word in Context (WiC) dataset by Pilehvar and Camacho-Collados (2019), which con-

[3] *bar.n* here indicates the noun *bar* etc.

[4] The authors also mention that this observation can be a useful filter criterion when collecting Usim annotations through crowdsourcing, as all annotations violating the triangle principle could be safely discarded.

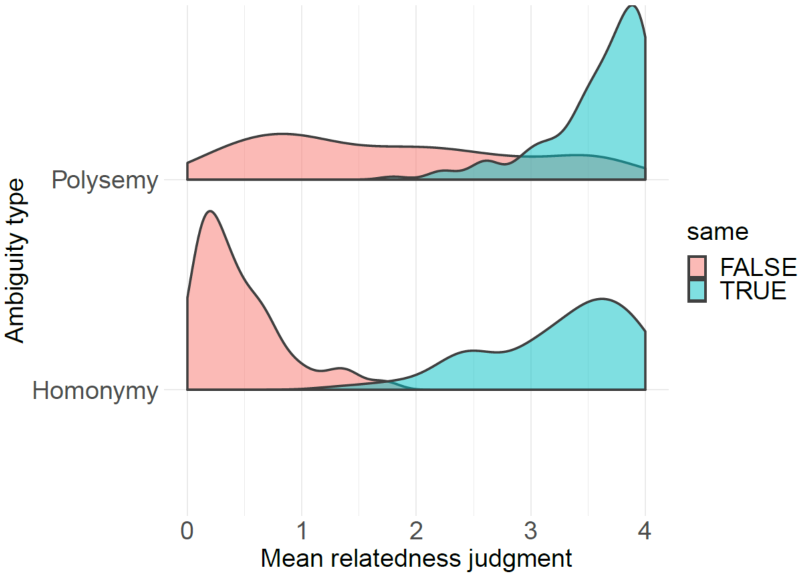


Figure 3.4: Mean relatedness of RAW-C judgements for sentence pairs containing lexically ambiguous words, plotted by by Same Sense (True vs. False) and Ambiguity Type (Homonymy vs. Polysemy). Figure replicated from Trott and Bergen (2021).

tains over 7,000 sentence pairs with an overlapping English word but was annotated based on a binary classification task, and CoSimLex (Armendariz et al., 2020), which contains graded similarity judgements for related words instead of different interpretations of the same word. This meant that at the start of our research project, no annotated resource was available to investigate the notion of word sense similarity for lexically ambiguous words used in different contexts. In parallel to our work however, Nair et al. (2020) recently conducted an investigation of 32 polysemic and homonymic word types extracted from the Semcor corpus (Miller et al., 1993). In their annotation study, participants arranged contextualised samples in a 2D spatial arrangement task (Goldstone, 1994). Investigating only cross-sense samples, they reported polysemic senses to be rated significantly more similar to one another than homonymic samples in both the human annotations and contextualised BERT Base embeddings (Devlin et al., 2019, also see Section 3.3.2), and found a strong correlation between the cosine distance of BERT sense centroids and aggregated relatedness judgements.

In a similar approach, Trott and Bergen (2021) recently presented RAW-C, a dataset of Relatedness of Ambiguous Words, in Context. To create RAW-C, 77 participants annotated a total of 112 ambiguous words, each taken to invoke two different polysemic or homonymic interpretations (38 homonyms and 74 polysemes). Using a 5-point Likert scale, annotators here rated the relatedness of an ambiguous target highlighted in a displayed pair of context sentences. Based on the collected judgements, the mean relatedness for same-sense uses of an ambiguous word was calculated at 3.46, while cross-sense combinations were assigned a mean relatedness score of only 1.31. Furthermore, polysemous cross-sense samples were found to exhibit a much higher variance than the homonym samples while their same-sense distributions did not differ significantly (see Figure 3.4). Investigating contextualised language models ELMo (Peters et al., 2018, also see Section 3.3.1) and BERT Base (Devlin et al., 2019, also see Section 3.3.2) with their newly annotated data, the authors here concluded that 'both language models could differentiate same-sense and different-sense uses of an ambiguous word, but their ability to discriminate between homonymy and polysemy was marginal at best.'

3.2.2 Semantic Change Detection

Besides distinguishing different concurrent meanings of a word, a number of studies also have been investigating lexical semantic change over time. Lexical semantic change is considered to be tightly interlinked with polysemy, with Blank (1997) for example proposing polysemy to be the 'synchronic, observable result of lexical semantic change' (Schlechtweg et al., 2018). The computational modelling of lexical semantic change however has been limited by the unavailability of diachronic sense references nuanced enough to track the development of specific senses. Schlechtweg et al. (2018) recently presented an approach to address this issue, introducing the Diachronic Usage Relatedness (DURel) dataset. To create this resource, five native speakers of German were asked to rate 1,320 pairings of diachronic word uses based on a 4-point relatedness scale. Corpus samples were selected manually based on target words either found to indicate signs of innovation through sense narrowing (Paul, 2002) (19), or reduction due to homonymy (Osman, 1971) (9). Samples were then split into two periods, one with language use recorded from 1750-1800 (EARLIER), and one with samples produced between 1850-1900 (LATER). Comparing ratings given to EARLIER and LATER samples, the authors found three rough types of words: those whose mean relatedness increased, those whose relatedness decreased, and a majority of words for which the mean relatedness remained largely unchanged. If the mean relatedness of a word's senses increased over time, the authors took this as a sign of innovative language use, gaining interpretations unrelated to the original set. A decreasing mean relatedness on the other hand was taken as an indication of the loss of a previously available reading. The authors however also found that their measure was prone to confusing lexical semantic change with polysemy, and words with many interpretations could not reliably be investigated through their overall mean relatedness ratings alone.

3.3 Contextualised Language Models

One thing in common to all traditional static word representations is the central limitation that while a distributional semantics approaches to their development might in fact capture (different) word meanings, they cannot represent their specific use within a given context at test time, and therefore will be unable to encode speaker meaning or intended communicative function (see e.g. Brugman, 1988; Hopper, 1991; Paradis, 2011; Fermann and Lapata, 2016; Westera and Boleda, 2019). For the past few years, the NLP community has been working on a new generation of neural networks to address exactly this limitation of static word embeddings, and developed a range of so-called contextualised language models. Contextualised language models no longer provide a dictionary of pre-learned word embeddings, but instead can be used to derive a representation of a specific word in a specific context based on large-scale pre-training.

In this section we will introduce the most widely used contextualised language models, including ELMo, BERT and the GPT-n series, give an overview of their strengths and applications, and review literature attempting to investigate the inner workings of these models often described as black-boxes.

3.3.1 Embeddings from Language Models (ELMo)

One of the first (remarkably) successful approaches to context-specific representations was presented by Peters et al. (2018) in the form of ELMo, or Embeddings from Language Models. The underlying model is an unsupervised, bi-directional language model (biLM) pre-trained on next (and previous) word prediction. Under the hood, it is made up of a character encoding layer, two LSTM (Long Short-Term Memory) layers, and a simple feedforward neural network combined with a softmax function as an output layer. After pre-training, the contextualised embedding for a target word in a given sentence can be calculated by feeding the sample sentence to the model (with parameters frozen) and extracting the different layers' outputs.

For specific downstream tasks, ELMo embeddings can be derived by concatenating hidden state representations from both, the forward and backward networks, multiplying the concatenated vectors per layer with task-specific weights, and summing the result into a single output vector. In the simplest case, one can however also just select the model's top layer outputs,[5] or the hidden state outputs of one of

[5] as for example in TagLM (Peters et al., 2017) or CoVe (McCann et al., 2017)

the inner layers.[6]

Peters et al. test their ELMo embeddings on a wide range of NLP applications by replacing original model inputs with their novel contextualised encodings. Through this modification alone, they already were able to report state-of-the-art results for tasks like question answering on the Stanford Questions Answering Dataset (SQuAD Rajpurkar et al., 2016), textual entailment on the Stanford Natural Language Inference corpus (SNLI Bowman et al., 2015), semantic role labelling on the OntoNotes benchmark (Pradhan et al., 2013), and coreference resolution on the CoNLL 2012 shared task (Pradhan et al., 2012). Interesting with respect to the representation of ambiguous words is their observation that ELMo embeddings also can be used to predict the sense of a target word using a simple 1-nearest neighbour approach. Based on the SemCor 3.0 training corpus (Miller et al., 1994), the authors calculate the average representation for each of the recorded senses (by averaging the ELMo embeddings of the targets in different reference contexts), and determine the sense of a target word by finding the most similar of these sense embeddings. Using representations from the second LSTM layer only, they report F1 scores just slightly below the current state-of-the-art approach by Iacobacci et al. (2016) on all-words fine-grained WSD (Raganato et al., 2017).

3.3.2 BERT and the Dawn of the Transformers

Since their introduction in 2017, Transformer models (Vaswani et al., 2017) have become ubiquitous in NLP, and by now have effectively replaced (bi-)LM approaches like ELMo in state-of-the-art applications, shattering previous performance benchmarks left and right. Offering a revised model architecture that allows them to efficiently consume immense amounts of unsupervised training data - combined with a previously unthinkable amount of model parameters - Transformer architectures especially showcase an improved capability of modelling long-range dependencies relevant for many downstream tasks requiring an ‘understanding’ of the input text.

One of the most famous Transformer models is BERT, or Bidirectional Encoder Representations from Transformers (Devlin et al., 2019). Fundamentally, the BERT architecture is a stack of Transformer encoder modules consisting of multiple so-called self-attention heads. Each layer of self-attention heads is wrapped with a skip connection, and followed by layer normalisation and a fully-connected intermediate layer to combine and weigh outputs, turning them into the next layer’s inputs. In the BASE model, BERT is made up of 12 layers each consisting of 12 self-attention heads,

[6] see Section 4.3.2 for an pilot on the different layers’ encoding of polysemic word sense given different kinds of contexts.

and creates hidden states of 768 dimensions (for a total of 110 million parameters). BERT LARGE contains 24 layers, each with 16 attention heads and a hidden state representations of size 1024, boasting a total of 340 million parameters.

While officially coined bi-directional, BERT can practically be considered non-directional, as it no-longer processes language input sequentially, but instead encodes entire input sequences simultaneously. In order to allow for this kind of training paradigm, the authors present two pre-training tasks to replace the traditional next (or previous) word prediction: masked language modelling (MLM) and next sentence prediction (NSP). In the MLM task (in linguistics literature often referred to as Cloze task Taylor, 1953), 15% of tokens in an input sequence are replaced with [MASK]. In contrast to traditional left-to-right or right-to-left language models predicting next of preceding words, the masked tokens allow the model to simultaneously consider preceding and succeeding contexts without 'seeing' the target. The NSP training task is especially relevant for tasks such as question answering, where the relationship between different sentences is of central importance. During BERT pre-training, 50% of samples are consecutive sentence pairs in the corpus data, while in the other half the second sentence of a sample is a randomly selected one.

Following an approach previously labelled Universal Language Model Fine-tuning (or ULM-Fit, see Dai and Le, 2015; Howard and Ruder, 2018; Radford et al., 2018), once pre-trained, BERT can be fine-tuned to a specific task relatively inexpensively by feeding it sample pairs relevant to the task at hand, like for example question-answer pairs or hypothesis-premise pairs. Fine-tuned on the GLUE (Wang et al., 2018) benchmark suite for example, the authors report that 'BERT BASE and BERT LARGE outperform all systems on all tasks by a substantial margin, obtaining 4.5% and 7.0% respective average accuracy improvement over the prior state of the art.'

Besides applying masked token and next sentence prediction, it should be noted that BERT operates on so-called word pieces (or sub-word tokens Wu et al., 2016) with a 30,000 token vocabulary. While this greatly increases the amount of possible inputs, it also means that in order to obtain embeddings for specific words that were split into multiple sub-word tokens during pre-processing, a given number of sub-word embeddings need to be combined - usually by simply averaging over the representations of all word pieces involved.

3.3.3 GPT-n, T5 and Other Recent Variants

BERT's main competitor in the race for meaningful contextualised word embeddings is the GPT-n series developed by OpenAI (Radford et al., 2018, 2019; Brown et al., 2020) - for now mostly represented by GPT-2. While also based on a Transformer architecture, GPT (Generative Pre-trained Transformer) models take a slightly different approach on processing input text and as a result more closely resemble traditional language models: under the hood, GPT-2 for example is an auto-regressive stack of Transformer decoders with between 12 and 48 layers. Its auto-regression prevents the model from using the masked word training objective applied in BERT models, but on the other hand again allows it to process in- and outputs sequentially, which - contrary to BERT - enables GPT models to also be used for text generation. GPT-2's hidden state representations range from 768 dimensions in GPT-2 Small, to 1,600 dimensions in GPT-2 Extra Large, giving the latter a total of 1.5 billion parameters - an order of magnitude more than BERT. This number however already has been put to shame by its recently presented successor GPT-3 (Brown et al., 2020), an auto-regressive language model with 175 billion parameters, or, '10x more than any previous non-sparse language model.'

Other notable mentions include Google AI's T5 (Text-To-Text Transfer Transformer Raffel et al., 2020), an 11 billion parameter model based on the novel Reformer architecture (a Transformer model designed to handle context windows of up to 1 million words, see Kitaev et al., 2020); XLNet, a 'generalised auto-regressive pretraining method that enables learning bi-directional contexts' and therefore 'overcomes the limitations of BERT thanks to its auto-regressive formulation' (Yang et al., 2019); as well as BERT variants like RoBERTa (Robustly Optimised BERT Pre-training Approach, Zhuang et al., 2021) and ALBERT (A Lite BERT, Lan et al., 2019), and recent spin-offs like BART (a denoising autoencoder for pretraining sequence-to-sequence models, Lewis et al., 2020b) and MARGE (a Multilingual Autoencoder that Retrieves and Generates, Lewis et al., 2020a). In this thesis we will however focus on the base versions of Word2Vec, ELMo, and BERT only, investigating in a more principled fashion how these models encode contextualised sense similarity rather than attempting to develop a state-of-the-art application.[7]

3.3.4 Probing Contextualised Language Models

While contextualised language models proved to be very successful in a range of downstream NLP tasks, they also provided the community with a dilemma: due to their black-box architecture, not much is known about *how* these models achieve their remarkable performance levels. This lack of knowledge both affects model accountability and explainability, as well as 'limits hypothesis-driven improvement

[7] Due to its conception as a language model, GPT-2 was excluded from our analysis as it doesn't provide contextualised embeddings off the shelf (see Section 4.2.2).

of the architecture' (Rogers et al., 2020). The quest for insights into the inner workings of (among others) contextualised language models therefore spawned a whole new sub-field of NLP research focused on probing and explaining large neural network models.[8]

Among one of the first, Ethayarajh (2019) investigated the vector spaces spanned by the word encodings produced by contextualised language models like ELMo, BERT and GTP-2. Firstly, they found that the word vectors of all of the tested models were anisotropic, forming only a narrow cone in the representation space. They even found that the anisotropy of GPT-2's last layer was 'so extreme that two random word will on average have almost perfect cosine similarity.' Secondly, the similarity between vector representations of the same word in different contexts decreased in upper layers, suggesting that 'upper layers of contextualised language models produce more context-specific representations.' And thirdly, context-specificity was found to manifest differently in different contextualised language models, with for example ELMO representations of words in the same sentence becoming more similar to each other in upper layer representations, while they become more dissimilar in BERT's - and GPT-2 not assigning any different similarity scores to words in the same sentence than to any other two random words.

One of the conclusions that Ethayarajh drew from their observations is that 'the variety of the contexts a word appears in, rather than its inherent polysemy, is what drives variation in its contextualised representations', or, in other words, that contextualised language models do not seem to encode a fixed number of sense representations derived from the corpus data, but instead create idiosyncratic representations for each word occurrence that combine a range of different information derived from context.

Yenicelik et al. (2020) seconded this observation based on a rigorous quantitative analysis of linear separability and cluster organisation in embedding vectors produced by BERT. They found that semantics here does not appear to surface as isolated clusters, but that sense embeddings form seamless structures that are tightly coupled with sentiment and syntax. They however also found that polysemous words had a high variance in their mean standard deviation (providing support for a hypothesis initially put forward by Miller and Charles, 1991), but note that also non-polysemic words, like for example stop words, can have equally high variance, and that variance alone therefore is not a surefire sign of multiplicity of sense.

Reviewing the state of black-box NLP research in 2020, Rogers et al. (2020) cite investigations like Ettinger (2020)'s, finding that BERT can encode information

[8] Also see the BlackboxNLP Workshop Series `https://blackboxnlp.github.io/`.

concerning target words' semantic roles; a study by Tenney et al. (2019b) showing that BERT encodes information about entity types, relations, semantic roles and proto-roles; but also indications of BERT appearing to struggle with other simple concepts like numbers (Wallace et al., 2019). Concerning the roles of the different encoding layers of contextualised language models, the compiled literature suggests that lower layers have the most linear word order information (Lin et al., 2019), syntactic information is most prominent in BERT's middle layers (Hewitt and Manning, 2019), and that the final layers of BERT are the most task specific (Liu et al., 2019) - with semantic information spread across the entire model (Tenney et al., 2019a). Rogers et al. (2020) however also question the use of attention weights as a tool for interpreting deep learning models - an approach which had recently been gaining popularity in the community (also see Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019; Brunner et al., 2020, for different contribution to the ongoing debate). They note that visualising attention weights and similar model-internal metrics limits analysis to exclusively qualitative approaches and should not be interpreted as definite evidence (Belinkov and Glass, 2019).

3.3.5 Contextualised Embeddings for Distributional Semantics

Investigating BERT as a distributional semantics model (DSM), Mickus et al. (2020) find that while BERT shows a tendency towards coherence in its contextualised word representation, it does not fully live up to the expectations of a semantic vector space. In particular, they find that the target word position within a context sentence has a noticeable impact on its embedding and disturbs word sense similarity relationships. Treating BERT as a black-box model, they deliberately only use the outputs of the last layer of a vanilla, pre-trained BERT architecture, and analyse the distribution of silhouette scores (Rousseeuw, 1987) derived from the spacial embedding of different words. Polysemous targets overall tended to have a lower cohesion score in this representation, and a lower silhouette score than monosemes - both compatible with what would be expected of a DSM. Continuing their analysis, the authors however also found that tokens from different sentence positions (even vs. odd) would create significantly different embeddings. Overall, Mickus et al. therefore concluded that the next sentence prediction (NSP) objective used during the pre-training of the model tends to obfuscate its relation to distributional semantics.

3.3.6 Word Sense Disambiguation Revisited

Now having at their disposal a range of models capable of creating contextualised embeddings, a number of scholars started investigating approaches of using contextualised language models and their outputs to improve WSD performance. After observing that ELMo embeddings of target words in similar contexts can form clusters, and that the representations of words with multiple meanings can split into different groups, roughly representing theses different interpretations (Schuster et al., 2019), Chang and Chen (2019) for example started exploring whether contextualised embeddings are sense-informative enough to derive a sense definition given a (target, context) pair. To this end, they encoded all 79,030 meaning definitions from the Oxford dictionary, and trained a classifier to link contextualised embeddings to these definition embeddings. In the `seen` condition (target word and definitions seen during training), the retrieval precision using BERT Base embeddings ranged from 75 P@1 to 85 P@10, and in the `unseen`, zero-shot condition (target word not seen during training) the retrieval precision using BERT Large embeddings ranged from 3.5 P@1 to 15.5 P@10. All of these scores clearly outperform baselines using static word embeddings or static word embeddings together with static context embeddings only. The drop in precision however still clearly shows that the classifier seems to work much better for word sense discrimination, where likely most, or at least the most prominent interpretations were seen during training, and explicitly linked to their respective definition. Word sense induction in the zero-shot condition on the other hand seemed to perform quite poorly still.

At around the same time, Wiedemann et al. (2019) introduced a simple but effective approach to word sense disambiguation (WSD) using a nearest neighbour classification of contextualised embeddings. Applying k-Nearest Neighbours (kNN) clustering with k set to 1, the authors simply classified test set targets based on the nearest train set embedding. While this appeared to work remarkably well for BERT embeddings of train and test samples of the SensEval-2 (Kilgarriff, 2001) and SensEval-3 (Mihalcea et al., 2004) WSD tasks, outperforming the last submissions to these tasks, classification performance dropped notably when this approach was applied to the all-word tasks of SemEval2007 Task 7 (Navigli et al., 2007) and 17 (Pradhan et al., 2007), which both are only comprised of test data. Wiedemann et al. (2019) therefore conclude that the nearest neighbour approach suffers specifically from data sparseness and appears to require reference embeddings of practically each sense to work well.

Visualising the embedding space using t-SNE (van der Maaten and Hinton, 2008), the authors find that 'Flair embeddings hardly allow to distinguish any clusters as most senses are scattered across the entire plot. In the ELMo embeddings space, the major senses are slightly more separated in different regions in the point cloud. Only in the BERT embedding space, some senses form clearly separable clusters.' Based on this observation, Wiedemann et al. conclude that more powerful parametric classification approaches might be able to learn better decision boundaries (see e.g. Vial et al., 2019) than the kNN baseline presented in their paper.

Using a similar approach, Pasini et al. (2020) utilise k Means to cluster BERT's contextualised embeddings, but employ the number of senses registered in BabelNet (Navigli and Ponzetto, 2012) as parameter k for the clustering. While this limits the clustering to detecting only previously recorded interpretations and therefore discretises the problem, it allows the authors to use the BabelNet definitions to automatically disambiguate the resulting clusters. Comparing to a human-annotated gold standard developed by Bennett et al. (2016), their CluBERT approach outperforms the then state-of-the art model based on the Jensen-Shannon Divergence between the predicted distribution of word use definitions and the gold standard. In a very similar vain, Levine et al. (2020) present SenseBERT, noting that they 'focus on a coarse-grained variant of a word's sense, referred to as its WordNet supersense, in order to mitigate [...] brittleness of fine-grained word-sense systems caused by arbitrary sense granularity, blurriness, and general subjectiveness (Kilgarriff, 1997; Schneider, 2014).' This approach however limits them to identifying 45 different supersense categories, 26 of which for nouns, 15 for verbs, 3 for adjectives and 1 for adverbs. Instead of clustering, Levine et al. opt for a self-supervised model to predict soft-label category assignments. This approach out-performed a vanilla BERT baseline on a supersense-based variant of the SemEval WSD test sets (standardised by Raganato et al., 2017) and on the WiC task (Pilehvar and Camacho-Collados, 2019), but a qualitative analysis of some of the classifications revealed that the model still made consistent categorical mistakes.

Taking a different approach, Amrami and Goldberg (2019) experimented with using target substitutions in order to improve the representation of a given sense interpretation within a sample sentence. In order to derive these substitutes, the authors propose to use specific search patterns instead of simply using a language model to replace the target in the sentence, as this naive approach could produce lexically unrelated candidates. For example, when masking *dogs* in the first sample sentence of Example 27, BERT's highest scoring prediction would be *eyes*. So, in order to derive lexically similar substitutes for *dogs* or *brown*, they would use sentences like 27b and c - but report that these patterns do not significantly improve performance on the word sense induction (WSI) task of SemEval 2013.[9]

[9] Examples from Amrami and Goldberg (2019)

(27) a. My *dogs* are brown.
b. My dogs (or even [MASK]) are brown.
c. My dogs are brown and [MASK].

Besides experimenting with substitution techniques to improve WSI performance, they also investigated clustering samples with a dynamic number of clusters as opposed to the fixed number of seven clusters in their previous work (Amrami and Goldberg, 2018). They approached this by setting a relaxed upper bound to the number of clusters, i.e. using soft clustering to assign samples to different degrees to a set of ten clusters. This however also did not improve average WSI scores on the SemEval 2013 dataset.

Blevins and Zettlemoyer (2020) finally highlight the effect of under-representation in the pre-training of large contextualised language models like BERT, specifically on their ability to perform word sense disambiguation (WSD) on words that are either rare or completely unseen during training. They present an end-to-end trained bi-encoder built on top of BERT, designed to improve the performance on rare and zero-shot sentences by jointly learning contextualised word embeddings and a gloss encoder from the WSD objective alone. Applied on the English all-words WSD task introduced in Raganato et al. (2017), this model led to an overall absolute improvement of 15.6 F1 over the next-best previous system, with an 31% error reduction on less frequent senses making up for the vast majority of the improvement gain.

In an ablation experiment, the authors balanced the representation of low-frequency words by weighting the loss for a specific sense by the inverse frequency of the target in the training data. While this slightly improves performance for zero-shot predictions, it reduces the overall F1, showing that this improvement for low-frequency senses comes at the cost of the (rather informative) data bias towards the most frequent sense. Combining their observations, Blevins and Zettlemoyer conclude that with frozen, vanilla BERT models reaching over 94 F1 on samples labelled with the most frequent sense, improving the disambiguation of less common senses should be the main objective of future work on WSD. Data augmentation is then suggested as one factor in this endeavour, but the authors propose that adding a few labelled samples representing rare senses could be more effective than 'simply annotating more data without considering the sense distribution.'

3.3.7 Semantic Change Revisited

Investigating contextualised language models as a tool to analyse lexical semantic change, Giulianelli et al. (2020) showed that predicted similarity shifts correlate well with human judgements. Noting the limitations of a single word representation (Hopper, 1991; Lau et al., 2012; Frermann and Lapata, 2016; Hu et al., 2019) and those of fixed word sense representations (Brugman, 1988; Kilgarriff, 1997; Paradis, 2011) in capturing 'word meaning, which is continuous in nature and modulated by context to convey ad-hoc interpretations,' they suggest the use of contextualised representations to utilise case-by-case context information for a more fine-grained, seamless representation of ad-hoc sense.

As a first step of evaluation, Giulianelli et al. compare the similarity between BERT's embeddings for 16 selected targets with human judgements of word sense similarity. To collect these judgements, they generated a total of 3,285 usage pairs representing target word usages across five 20-year windows within the last century of the Corpus of Historical American English (COHA, Davies, 2012). Annotators on crowd-sourcing platform Figure Eight[10] were shown pairs of target word usages within their original context, and asked to rate their similarity using a 4-point scale, ranging from *unrelated* to *identical* (also see Brown, 2008; Schlechtweg et al., 2018). Judgements from five annotators were then averaged to form a usage pair's similarity score and compared to the cosine similarities between the target's BERT embeddings using the Mantel test (Mantel, 1967) to calculate Spearmans's rank correlation. For 10 out of the 16 targets, the authors determined a significant, positive correlation between human similarity scores and BERT representation similarity, with Spearman ρ coefficients ranging from 0.13 to 0.45.

Encouraged by these results, Giulianelli et al. then used an unsupervised clustering of BERT embeddings[11] to create a *usage type* partitioning of contextualised representations. By measuring the entropy difference, Jensen-Shannon divergence (JSD) and average pairwise distance between periods, the authors quantified the difference in these word type partitionings between periods, correlating with the gold standard partitioning with a coefficient of between 0.27 and 0.28. A subsequent qualitative analysis of the partitionings revealed that 'usage types can discriminate between underlying senses of polysemous (and homonymous) words, between literal and figurative usages, and between usages that fulfil different syntactic roles.' In terms of semantic change across periods, the authors observe cases of *narrowing* as well as *broadening.* If it is said that the meaning distribution of a word narrows over time, this means that one of its senses over time is used less and less frequently, eventually becoming obsolete or disappearing completely. Giulianelli et al. present *coach* as an example for narrowing in their dataset, showing that the distribution of

[10] `https://www.figure-eight.com`, recently acquired by Appen (`https://appen.com`)

[11] In this case k Means with k maximising the silhouette score (Rousseeuw, 1987)

trainer/bus changes from about 70/30 in 1910 to about 15/85 in the 2000s. If the senses of a word broaden over time, it means that an expression develops novel interpretations. An example here is *disk*, which gained an interpretation as a medium for information in the mid-century.

3.4 Summary

In this Chapter we presented different computational approaches to representing and dealing with lexical ambiguity, ranging from static, distributionally motivated word vectors over dedicated sense embeddings to the recent development of contextualised language models that are capable of encoding a specific word in a given context based on a substantial amount of pre-training and a previously inconceivable amount of parameters. We especially focused on an investigation of large-scale datasets of graded word sense similarity intended to give an indication of the relatedness of the uses of an ambiguous target in different contexts. Prior to the data collection presented in this thesis, the only available resources on this issue were preliminary, manually annotated corpora including a few words only. In parallel to the work presented here, a number of other studies however have addressed the same issue, producing datasets similar in nature to the one we will presenting in the next two chapters. The RAW-C dataset developed by Trott and Bergen (2021) exhibits the largest resemblance with the methodology adopted for our experiments (see Chapters 4.3 and 5.2), but includes only two senses for each target word - limiting the analysis of similarity patterns - and oftentimes uses compound noun phrases to disambiguate the target, e.g. *traffic cone* vs *ice cream cone*, or *fruit bat* vs *baseball bat*. We suggest that this can be problematic for polysemic targets, because in this case the expression will no longer allow for an under-specified interpretation, and consequently undermines the function of polysemy as proposed in Chapter 2.3.1.

Having covered the literature on both traditional (psycho-) and computational linguistics, we identified a need for dedicated and reliable data to allow for a more empirical investigation of the notion of (graded) word sense similarity. Once available, this data could contribute to the evaluation of theoretical models of the human language processor and specifically hypotheses concerning mental lexicon, but also could find its use in the development and evaluation of computational models dealing with lexical ambiguity. In the remainder of this thesis we will present a novel, carefully crafted large-scale dataset of graded word sense similarity and co-predication acceptability to indicate the human processing of ambiguous words and allow for an evaluation of computational ones.

Chapter 4

Pilot on Polysemic Word Sense Similarity

Traditionally, deciding whether two uses of an ambiguous word invoke the same interpretation is a discreet, binary question: Either they do invoke the same sense - or they don not. This assumption holds both for word meaning, i.e. the interpretation of homonyms, and word sense, i.e. the interpretation of polysemes. Based on accumulating evidence that these phenomena of homonymy and polysemy however might not be as homogeneous as is traditionally assumed (see Chapter 2.5), in the next two chapters we will investigate the concept of *distance* between interpretations of ambiguous word forms as a more gradual, continuous replacement of that traditional discreet, binary decision. In particular, we will explore graded annotations of explicit word sense similarity and implicit co-predication acceptability as measures of interpretation distance, and show that - depending on their use - polysemic senses can be perceived as ranging from identical in meaning to completely unrelated in their interpretation. We then take these empirical observations as arguments into the ongoing debate on the mental representation of polysemic sense, suggesting that while a perceived identity in meaning between two different polysemic senses is difficult to explain with a sense enumeration approach (what distinguishes these two interpretations to warrant distinct entries in the lexicon), significant differences in the similarity of two polysemic senses are difficult to reconcile with a fully underspecified approach (if two senses are interpreted differently, how can they both be derived from the same entry).

To provide reliable empirical data for an investigation of the notion of distance between word sense interpretations, we collected human-annotated data for different types of word sense similarity through the course of two annotation runs. The first annotation run mainly focused on validating the experiment methodology by collecting and analysing an initial array of judgements for a small number of seminal polysemic nouns. After establishing that the proposed methodology is sensitive enough to record judgements reflecting the fine-grained differences in the interpretation of polysemic senses, the second run was used to generate a comparably large dataset of explicit word sense similarity and co-predication acceptability judgements for a total of 28 regular, metonymic polysemic targets. The collected data can be used for a detailed analysis of polysemic word sense similarity (see Chapter 5.3), as well as a benchmark for evaluating the capability of computational language models to predict human annotations (Chapter 5.3.3).

This chapter is structured as follows: in Section 4.1 we will clarify the motivation and assumptions of our experiments, before introducing the different human annotation measures as well as computational models used in our experiments in Section 4.2. We will then detail the requirements for and generation of the pilot experiment samples in Section 4.3, and present the methodology proposed for the data collection (Section 4.4). Section 4.5 contains a rigorous analysis of the collected pilot data, including a comparison between the different measures and a preliminary investigation of the performance of the computational models in predicting the human judgements. Section 4.6 will complete the chapter with a summary and discussion, concluding the the obtained results provide appropriate validation to continue with a large scale data collection effort based on the established methodology. Parts of this chapter have previously been published in Haber and Poesio (2020a,b).

4.1 Motivation

As we have seen in the previous chapters, ambiguity poses a central challenge to conceptualising the representation of word senses in the mental lexicon. While it is still unclear how exactly brain activation patterns ultimately facilitate our language processing capabilities, illustrating our assumptions about the processing of words through theoretical models is a worthwhile approach to inform strategic hypothesis production and testing. Hypotheses however should be based on reliable empirical data, and be grounded in observations based on representative data samples - not merely introspective arguments based on individual, potentially idiosyncratic examples of language use.

In this thesis we focus on the notion of word sense similarity, or, viewed from a different perspective, the differences or distance in their mental representation. To this end, we assume that word sense similarity and co-predication acceptability ratings - as indicated by a sizeable number of annotators - are derived from and therefore indicate differences in their mental representation. We hope that probing the way we process and understand language through these proxies can provide us with more reliable insights to aid our understanding of the vast and complex phenomenon of language.

The data collection and analysis presented in this and the following chapter are aimed at answering the bulk of the research questions stated in Chapter 1.2, including our primary research question

Q1 Does empirical evidence on word sense distance support the traditional hypotheses of word sense enumeration or fully under-specified mental representation of polysemes in the mental lexicon?

together with its underlying inquiries

Q1a Does empirical evidence on word sense similarity indicate differences in the interpretation of different polysemic senses?

Q1b Does word sense distance form discernible patterns in the interpretation of regular polysemic senses?

While this chapter mainly describes the methodology and presents some validation tests on initial pilot data, Chapter 5 provides an in-depth analysis of the full set of collected data, and presents our findings with respect to these questions. It is also Chapter 5 that presents our results to the second main research question

Q2 Do computational approximations of word sense similarity correlate with empirical indications of word sense distance?

4.2 Annotation Measures

In our first annotation run, we collected three measures of polysemic word sense similarity: explicit word sense similarity judgements, co-predication acceptability judgements, and an experimental measure of word class overlap. Each measure covers a slightly different type of annotation judgement, providing graded, meta-linguistic judgements of explicit word sense similarity in the first case, ecological judgements of sentence acceptability under co-predication, and coarse, multi-hot sense similarity vectors in the word class setting. We then analysed the collected data to determine whether the different measures were sensitive enough to capture different word meanings and word senses, compared the judgements obtained from the three annotation methods with each other, and tested how well a range of computational language models could predict the human judgements.

4.2.1 Human Judgements

As a first measure of word sense similarity, we collected graded annotator judgements explicitly rating the similarity of the interpretation of a given target expression presented in two different contexts. Participants here were shown pairs of sentences like in (28) and asked to rate the similarity of the highlighted words.

(28) 1 The **newspaper** fired its editor in chief.
2 The **newspaper** got wet from the rain.

While at first glance these judgements should provide a straightforward indication of word sense similarity, a central issue with this approach is connected to the proposed function of polysemy: as we explored in Chapter 2.3.1, some scholars suggest that the function of polysemy is to allow the language processor to continue with the interpretation of a sentence without fully specifying the ambiguous expression. Supporting this line of thinking, studies like Swets et al. (2008) suggest that the interpretation of a polysemic expression might be left under-specified in many cases, meaning that a reader either is aware of all possible interpretations and chooses not to decide on a specific one, or - more likely - at that point simply is not aware of the ambiguity of the target word (also see Ferreira et al., 2001, 2002; Logačev and Vasishth, 2016). When in our study an annotator is asked to compare two different interpretations of a polysemic word and rate their similarity, they are indirectly required to resolve any under-specification and select one interpretation for each sample in order to rate their similarity. This means that participants will very likely not process polysemic targets as they would do in a more natural text-processing setting, which might nullify any potential ambiguity advantages. So while these explicit similarity ratings are likely to be a good indicator of any underlying word sense similarities, these annotations present a meta-linguistic signal, with annotators aware of the polysemic targets and actively comparing interpretations.

In order to better assess collected word sense similarity ratings, we will compare them to judgements of co-predication acceptability derived from the same samples. As shown in Chapter 2.5.1, co-predication acceptability is one of the traditional linguistic tests used to distinguish homonyms from polysemes. It usually rests on the assumption that felicitous co-predication indicates that the interpretations of both clauses are based on the same word meaning, while infelicitous co-predication indicates homonymy. To collect co-predication judgements, in our experiments annotators were presented sentences like

(29) The newspaper fired its editor in chief and got wet from the rain.

and asked to rate the acceptability of the sentence.

While as a binary signal co-predication acceptability would be too coarse a measure to assess word sense similarity with the resolution required for our experiments, studies like Lau et al. (2014) and Murphy (2021) suggest that - like other measures of grammaticality - co-predication acceptability might also be perceived as a graded phenomenon. Combined with previous suggestions that some polysemic interpretations appear to fail co-predication tests (e.g. Cruse, 1986, 2004, also see Chapter 2.5.1), graded co-predication acceptability could lend itself particularly well to assess the similarity of different polysemic interpretations. In contrast to explicit similarity judgements, co-predication acceptability judgements provide a more ecological signal of the similarity of the interpretations involved, as annotators here simply judge the acceptability of the resulting structure. This means that they are not necessarily aware of the potential reasons for decreased acceptability, or, in this case, the polysemy of the target word form. Co-predication acceptability on the other hand however also is not a perfect representation of polysemic word sense, as acceptability judgements combine these implicit assessments of sense similarity with compounding factors including grammaticality, sentence length and complexity, and logical and temporal coherence (see e.g. Murphy, 2021, and Chapter 2.5.1).

As a third judgement of word sense similarity, we experimented with a class-based representation of word sense interpretation. Class-based judgements can be represented in multi-hot vectors that allow for a calculation of word sense similarity through determining the overlap of assigned labels. To collect sense labels, we presented annotators with individual sentences and a selection of different classes, from which they were asked to select all applicable ones:

(30) The newspaper fired its editor in chief.

- □ literary work
- □ medium
- □ organisation
- □ person
- □ physical object
- ...

Compared to the previously presented methods, word class overlap is a rather coarse measure of word sense similarity. If, however, co-predication acceptability - as traditionally assumed - only conveys a binary signal as to whether different interpretations are acceptable in co-predication or not, the coarse class overlap might correlate well with co-predication acceptability judgements.

4.2.2 Computational Approaches

Besides collecting different human annotations of word sense similarity, we also intend to investigate a range of computational approaches with respect to their ability to predict or proxy the human judgements. As manual annotations are comparatively expensive to generate for an extensive set of items - and impossible to collect for each possible sample of polysemous sense extension - computational language models could provide a more cost-effective way of establishing large-scale corpora dedicated to polysemic word sense similarity and fine-grained word sense discrimination.

For our analysis, we decided to focus on off-the-shelf versions of the recent line of large, pre-trained contextualised language models such as ELMo, BERT and GPT. As detailed in Chapter 3, these contextualised language models encode words relative to their context, i.e. generate a context-specific word embedding for each occurrence of a target word. While fine-tuning often is shown to significantly improve performance on a specific task (see e.g. Sun et al., 2019; Hao et al., 2020) we do not attempt any fine-tuning here for two reasons: firstly, at this point we are mainly interested in the representation of polysemic word sense that these models develop based on their default pre-training utilising large-scale, non-annotated language resources. And Secondly, we currently do not have sufficient annotated training data to fine-tune these models on distinguishing polysemic senses, a task that usually requires thousands of samples.

4.3 Materials

As shown in Chapter 2.2, polysemes are generally considered to be either regular or irregular, depending on whether or not other word forms share the same set of sense extensions. Considering previous studies indicating that irregular polysemes might be processed differently than their regular counterparts (see Chapter 2.5), we decided to first focus on regular polysemic nouns only. Investigating regular polysemes comes with the advantage that each type of alternation can be investigated through a number of target words that all allow for the same (sub-)set of sense interpretations. Being able to clearly distinguish these different interpretations and investigate them across a number of alternative targets gives us the opportunity to systematically explore the notion of word sense similarity both within the sense interpretations invoked by a specific word form, as well as across word forms that exhibit the same type of alternation.

It should be noted here that our study does not cover - and does not aim to cover - all possible interpretations and potential coercions of a target expression, but instead mostly focuses on the regular sense interpretations that can be found in other expressions as well. As an example, *Magazine* shares the *physical*, *information*, and *organisation* interpretation of *newspaper.* It however also allows for other, idiosyncratic interpretations that include a radio or television show, a type of storage, or an accumulation of food or ammunition stored in it, and each of these interpretations will be either a polysemous or homonymic extension of the others. Investigating regular senses only thus can shed light on the systematic side of polysemic sense alternation, but further research will be needed to fill in the gaps left by interpretations not included here, and to cover the grey areas created by coercion.

4.3.1 Target Expressions

For our first annotation run, we selected ten of the systematic polysemy types compiled in Dölling (2020). The selected types have between two and four clearly distinct but related senses that can be found for at least two target expressions, from which we picked one of the most frequently used ones to represent each class. The resulting list of targets with their respective sense interpretations is

1. **food/event**: lunch
2. **container-for-content**: glass
3. **content-for-container**: wine
4. **work-for-author**: War and Peace
5. **author-for-work**: Hemingway
6. **opening/physical**: door
7. **process/result**: construction
8. **physical/information/organisation**: newspaper
9. **physical/information/medium**: DVD
10. **building/pupils/directorate/institution**: school

With possibly the exception of the *work-for-author* and *author-for-work* alternations, most of the chosen alternations are arguably cases of inherent polysemy,[1] with different senses tied to the denotation of the respective objects. Furthermore, ll alternations are metonymic, representing different facets of the same denoted object (see Chapter 2.2).

[1] According to the definition assumed by e.g. Pustejovsky (1995) and Ortega-Andrés and Vicente (2019).

4.3.2 Sample Contexts

The different meanings or particular senses of an ambiguous expression are elicited by its respective context. In order to compare the different interpretations of the selected target expressions, we thus also require context samples to invoke each of their different sense extensions. Sample contexts can either be selected from corpora of natural language use, or created specifically for a given task. Selecting corpus samples has the advantage of providing potentially more natural context sentences, as - depending on the corpora used - they will be selected from previously recorded utterances or texts written for purposes unrelated to the current study, and therefore limit potential biases introduced when creating new samples specifically for a given task.[2] The downsides of using corpus samples for a study like this include the risk of not finding sample sentences that clearly and unequivocally invoke a certain interpretation - particularly if that interpretation is less dominant or potentially entirely unavailable for a given target - and the overall lack of control over factors such as target word position and function, sentence length and complexity, and other sentence elements beyond the focus of the current experiment. Especially behavioural studies (see Chapter 2.5.2) therefore usually rely on custom samples to control for factors potentially affecting the target signal.

Our study has three specific requirements on sample contexts: i) they need to invoke a single interpretation of the target expression as clearly as possible, ii) the sentences should be able to be rated independently (for class ratings), in comparison to one another (for similarity ratings) and when conjoined into a co-predication structure (for co-predication acceptability), and iii) the sentences need to produce consistent word embeddings when encoded with contextualised language models. As we will show in Section 4.3.2, this puts restrictions on the position and function of target words, as well as overall sample length. To meet all three requirements, we used custom samples in our experiments. The following sections will detail the constraints and requirements that our intended experiments imposed on the sample contexts to be used.

In the following sections, we will present more detail on the different requirements we determined for our sample sentences, as well as which observations these have been derived from.

[2] Although this does not fully eliminate biases, as samples must still be selected from the corpus, which allows the introduction of biases, too.

Requirements for Co-predication Samples

As we intended to use the same set of samples for all three measures of human similarity judgement, we required our sample sentences to be easily conjoinable into co-predication structures. We found that a straightforward way to do so is through conjunction reduction (Chomsky, 1957; Zwicky and Sadock, 1975), where two sentences with the same subject are combined by replacing the subject in the second sentence with a conjunctive (usually *and* or *but*, see Examples (31) and (32)):[3]

(31) a. They saw her duck.
b. They saw her swallow.
c. They saw her duck and (her) swallow.

(32) a. The newspaper fired its editor in chief.
b. The newspaper got wet from the rain.
c. The newspaper fired its editor in chief and got wet from the rain.

Relying on conjunction reduction introduced a first set of two constraints on the construction of sample contexts: firstly, the target expression should be the subject of a context sentence, because otherwise conjunction reduction would be impossible to apply when combining sentences under co-predication. Secondly, individual sample sentences should not contain any conjunctions themselves, as these would likely create more complex and potentially less acceptable constructions in the co-predication setting. To ensure that co-predication structures created from individual sample contexts were logically and temporally consistent, we introduced three additional constraints: firstly, sentences that introduce predications which enable or prevent further interaction with the target (i.e. by damaging it or limiting its availability) should be avoided. An example of a situation where certain interpretations can become unavailable is shown in Example (33), where the two references to *glass* both elicit the same *container* interpretation. Still, co-predication sentence (33)c could receive a low acceptability rating solely due to the fact that a broken cup can no longer be filled to the brim.[4]

(33) a. The glass broke when she dropped it.
b. The glass is filled to the brim.
c. The glass broke when she dropped it and is filled to the brim.

[3] Example (31) from Zwicky and Sadock (1975)

[4] Explicit similarity comparisons might be affected by this effect as well if readers interpret the two sentences as part of the same narrative and therefore consider the glass broken when reading the second sentence.

Additionally, we concluded that the target expression should be introduced with a definite article to prevent potential confusion on reference ('*the* cup fell and *the* cup broke' vs '*a* cup fell and *a* cup broke'), and that the verb phrase should be set in simple past or present tense to allow for temporal dependencies without introducing grammatically complex structures. The tense here should however always depend on the predication itself and potential combinations of predications to minimise its impact on the perceived sense similarity or co-predication acceptability.

Requirements from Contextualised Language Models

A second set of constraints were derived from a preliminary investigation of contextualised embeddings generated by ELMo (Peters et al., 2018).[5] The main focus of this investigation was to determine central dependencies between a context sentence and the encoding of a given target word, especially considering the position and function of the target word in the context sentence, and the amount and relevance of context specified.

We conducted our analyses using the default TensorFlow Hub ELMo implementation,[6] extracting the embedding of a polysemic target after encoding it within a given context sentence. Starting with seminal examples of polysemic expressions, we quickly realised that the position and function of the target within a context sentence had a significant effect on the resulting ELMo embedding, and therefore should be controlled as much as possible (also see e.g. Klafka and Ettinger, 2020). In the generation of our pilot samples both function and position of the target expression however are already fixed through the constraints of applying conjunction reduction to create co-predication structures, which means that these effects will be negligible in our experiments.

As a next step, we wanted to investigate the effect of sample length on the resulting ELMo embedding. We assumed that shorter samples might indicate sense similarity effects stronger than longer ones, as any additional information could dilute the encoding of the relevant disambiguating context. To test this hypothesis, we created a set of sentences that fixed the position and function of the target expression, but continued with one of four contexts varying in length and complexity: 1) the absolute minimal context to invoke a certain sense, 2) a relatively short context, 3) an extensive but descriptive context, 4) an extensive, natural context with potentially tangential information. Using polyseme *newspaper*, with sense interpretations a) *physical*, b) *information*, and c) *organisation*, we generated the following twelve

[5] See Chapter 3.3.1 for an overview]

[6] `https://tfhub.dev/google/elmo/3`

samples according to these guidelines:

(34) 1. The newspaper is folded.
2. The newspaper is boring.
3. The newspaper is famous.

4. The newspaper is lying on the table.
5. The newspaper is listing job openings.
6. The newspaper is struggling financially.

7. The newspaper is made up of 40 sheets of thin, recycled paper, has three columns of text and only a few colour images.
8. The newspaper contains reports on national and international incidents, the daily weather report and sports results.
9. The newspaper fired its editor in chief after her new business strategy caused the company to lose important partners.

10. The newspaper got wet from the sprinklers because the paper boy hadn't thrown it far enough to reach the front porch.
11. The newspaper wasn't very interesting but got the local obituaries and job offers which were read by almost everyone.
12. The newspaper was attacked over its populist coverage of the recent events surrounding the general election in May.

We then calculated the cosine similarities (1-cosine) between the embeddings of the target word *newspaper* for all sentence pairs using the LSTM's first layer's hidden state, the LSTM's second layer's hidden state and the ELMo output embedding. Figure 4.1 displays the results as a heat map. The results indicate that the embeddings of sample sentences 1 through 6 exhibit a much higher similarity to one another than to the rest of the pairwise comparisons in all of the embedding layers. It thus seemed that simply adding the extensive context of samples 7 to 12 caused the target word embeddings to be noticeably different from those of the short context samples. And since we aim to specifically investigate word sense similarity (or the effect of the disambiguating context), we concluded that the contexts of the sample sentences for our experiments should be as short and descriptive as possible to minimise context effects on the contextualised embeddings - which aligns with the initial first requirement that samples should invoke a single interpretation of the target expression as clearly as possible.

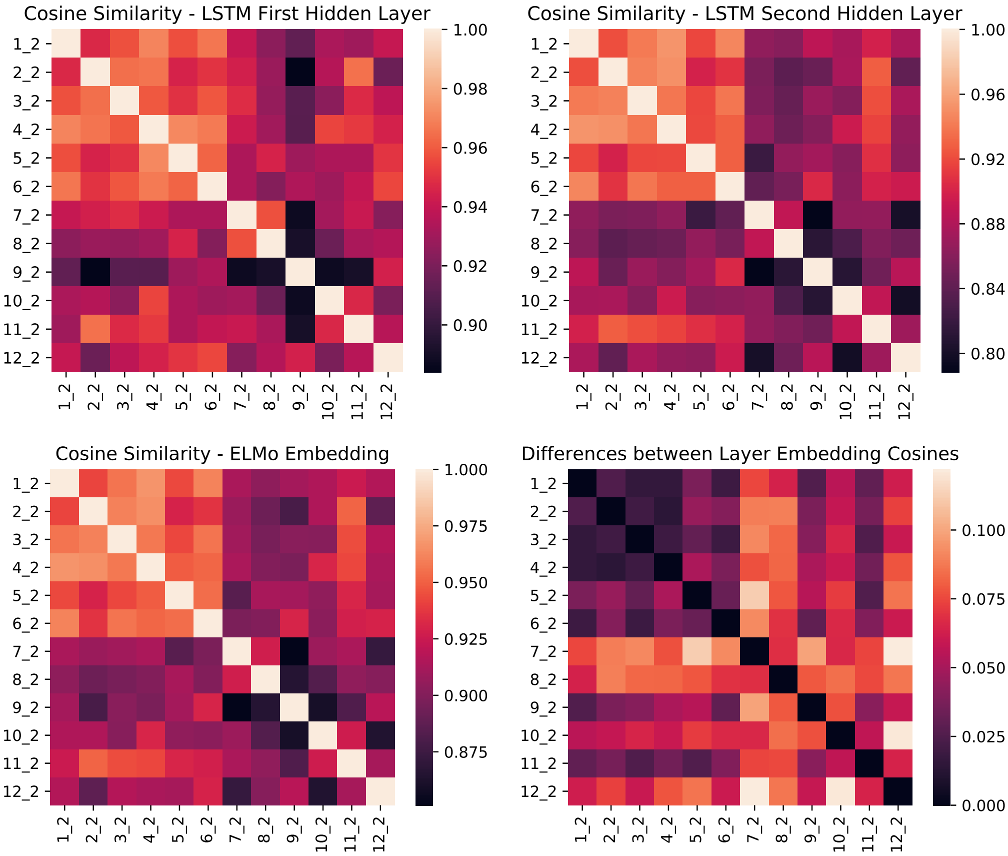


Figure 4.1: Heat maps of the pairwise cosine similarity of target word embeddings using a given ELMo layer, and a heat map of the differences in cosine similarity between the first and second LSTM layers' hidden state representation.

Sample Context Generation

To summarise the constraints and requirements established in the previous sections, context sentences should be created such that i) the ambiguous target expression is the subject of the sentence, ii) the subject is introduced at the start of the sentence, iii) the context is kept as short as possible, and iv) the context invokes a certain sense as clearly as possible.[7] Additionally, v) the target expression should be introduced with a definite article (if applicable), vi) verb phrases should be set in matching present or past tense, and vii) context sentences should not contain any conjunctives.

Collecting judgements of explicit word sense similarity or co-predication acceptability requires two samples for each comparison - either to be displayed as a pair to collect judgements of explicit word sense similarity, or combined into a co-predication structure to obtain acceptability ratings. To be able to also collect similarity and acceptability judgements for contexts eliciting the same interpretation in both samples, we therefore required exactly two sample sentences for each of the senses of our target words. The resulting full list of 54 target contexts can be found in Appendix A.1. As an example, consider the six sample sentences for polyseme *newspaper*, two each for its three senses (1) *organisation* or *institution*, (2) *physical object* and (3) *information* or *content*:

(35) 1a. The newspaper fired its editor in chief.,
1b. The newspaper was sued for defamation.
2a. The newspaper lies on the kitchen table.,
2b. The newspaper got wet from the rain.
3a. The newspaper wasn't very interesting.,
3b. The newspaper is rather satirical today.

Using this notation, comparing samples with the same number identifier results in what traditionally would be considered a same-sense scenario, and combining samples with different number identifiers results in a cross-sense comparison. For co-predication, two contexts are combined into a single sentence by the previously mentioned conjunction reduction (Zwicky and Sadock, 1975). As an example, contexts 1a and 1b are combined into co-predication sample 1ab as follows:

(36) 1ab. The newspaper fired its editor in chief and was sued for defamation.

We also created an additional sample set comprised of 15 common homonyms, with two sentences invoking their most dominant senses each, and a set of 30 sample

[7] Contexts should invoke a certain sense without mentioning that sense explicitly, as in 'The school is an old building.' for sense *building*

sentences containing 15 different pairs of synonyms. We initially included these samples to be used for filtering, but later realised that the judgements for these items were useful to better position annotations collected for the polysemic targets. The full lists of homonymic and synonymic samples can be found in Appendix A.1.1 and A.1.2, respectively. All sample sentences were rated to be acceptable by annotators recruited from Amazon Mechanical Turk (AMT)[8] in a validation experiment.

4.4 Method

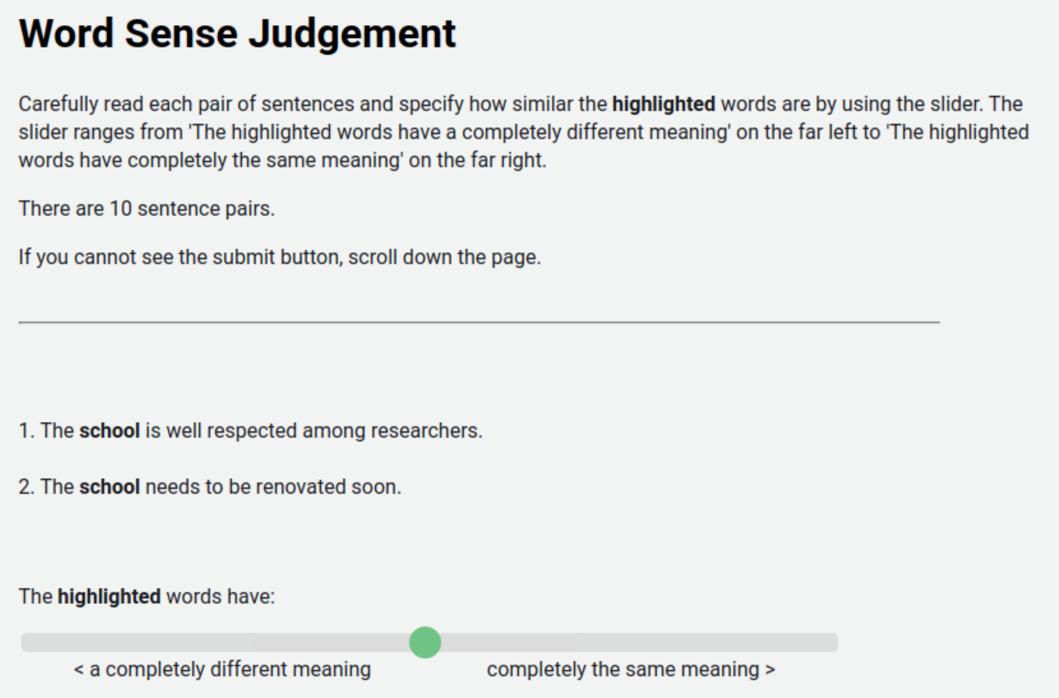
All human annotations were collected through the crowd-sourcing platform Amazon Mechanical Turk. Tasks were presented to annotators in the form of questionnaires, labelled Human Intelligence Tasks (HITs) on AMT. We asked for no prior experience or qualifications in annotating linguistics data, but required annotators to have obtained a US high school degree to indirectly filter for English native speakers, and limited annotators to those who reached the 'AMT Master' qualification to reduce annotation noise and improve judgement quality.[9] Since all samples were manually constructed, we could exclude the possibility of any explicit, abusive, offensive or otherwise harmful content, and could open the task for all interested annotators without a need for content warnings. Annotators were paid 0.35 USD for every completed questionnaire, for an average expected hourly rate of 7.00 USD, and (trough a technical malfunction) were not limited to completing just one questionnaire.

4.4.1 Word Sense Similarity Judgements

We collected explicit word sense similarity judgements by combining context sentences for a given target expression into pairs invoking all possible combinations of sense interpretations (including same-sense and cross-sense alternations). This resulted in four test items for polysemes with two senses, nine items for polysemes with three senses, and 16 for those with four, and a grand total of 75. In each test item, the target expressions were highlighted in bold font. Test items were distributed over 15 questionnaires so that target expressions appeared only once in any one questionnaire (and in one case twice for target *school* with 16 items). The questionnaires then were augmented with one of the 15 homonym and synonym samples each, and filled up to a total of ten items with filler samples randomly pairing discarded sentences with unmatched target words. Item order finally was

[8] `https://www.mturk.com/`

[9] According to AMT's website, '[t]hese Workers have consistently demonstrated a high degree of success in performing a wide range of HITs across a large number of Requesters,' `https://www.mturk.com/worker/help`



Word Sense Judgement

Carefully read each pair of sentences and specify how similar the **highlighted** words are by using the slider. The slider ranges from 'The highlighted words have a completely different meaning' on the far left to 'The highlighted words have completely the same meaning' on the far right.

There are 10 sentence pairs.

If you cannot see the submit button, scroll down the page.

1. The **school** is well respected among researchers.

2. The **school** needs to be renovated soon.

The **highlighted** words have:

< a completely different meaning ● completely the same meaning >

Figure 4.2: Screenshot of the the AMT interface for the explicit word sense similarity annotation task.

randomised within each questionnaire. Participants were given the following set of minimal instructions:

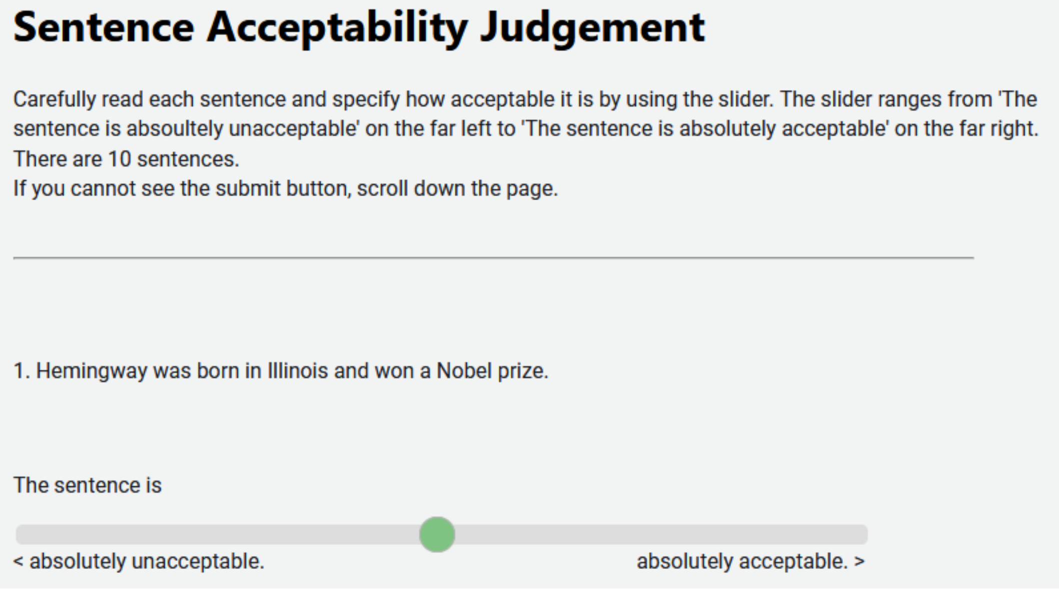
(37) Carefully read each pair of sentences and specify how similar the highlighted words are by using the slider. The slider ranges from ‘The highlighted words have a completely different meaning’ on the far left to ‘The highlighted words have completely the same meaning’ on the far right.

There are 10 sentence pairs.

A screenshot of the the AMT interface for this task is displayed in Figure 4.2. The submitted slider positions were translated to a 100-point similarity score ranging between 0 and 1, and stored in combination with an anonymised annotator ID.

4.4.2 Co-predication Acceptability Judgements

We collected graded annotator judgements rating the acceptability of co-predication structures combining different pairings of target word samples through conjunction reduction as described above. We manually inspected the co-predication structures for any inconsistencies that might have emerged through the conjunction and corrected issues with the least evasive measures possible, i.e changing the tense of a verb or adding or deleting temporal indicators. The samples were distributed over



Sentence Acceptability Judgement

Carefully read each sentence and specify how acceptable it is by using the slider. The slider ranges from 'The sentence is absoultely unacceptable' on the far left to 'The sentence is absolutely acceptable' on the far right. There are 10 sentences.
If you cannot see the submit button, scroll down the page.

1. Hemingway was born in Illinois and won a Nobel prize.

The sentence is

Figure 4.3: Screenshot of the the AMT interface for the co-predication acceptability annotation task.

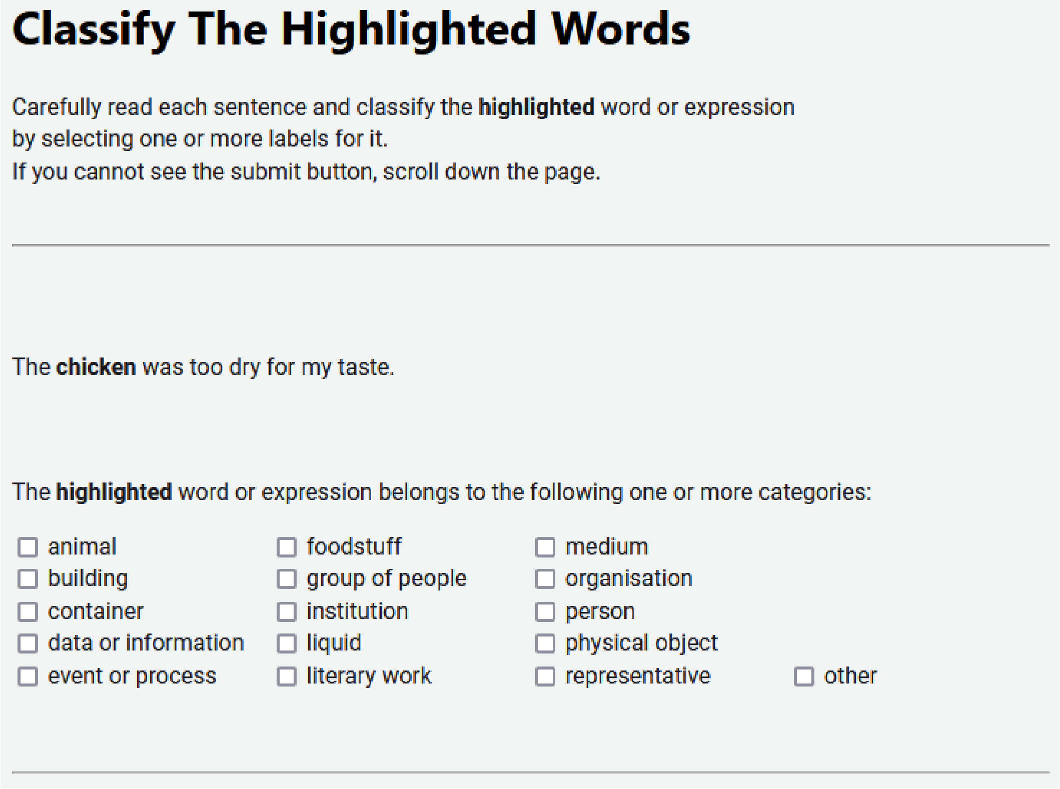
15 questionnaires in the same way as described above, and we again added homonym, synonym and filler items. Since conjunction reduction drops the second target mention, the synonym items however lost their effect in this setup. In the sentence acceptability task, the following instructions were shown to the participants:

(38) Carefully read each sentence and specify how acceptable it is by using the slider. The slider ranges from 'The sentence is absolutely unacceptable' on the far left to 'The sentence is absolutely acceptable' on the far right.
There are 10 sentences.

A screenshot of the AMT interface for this task is displayed in Figure 4.3. The submitted slider positions here were translated to a 100-point acceptability score ranging between 0 and 1, and again stored in combination with a unique but anonymised annotator ID.

4.4.3 Word Sense Class Annotations

To collect sense class labels, annotators were presented with individual sample sentences together with a list of 16 sense class labels. Class labels were derived from the descriptions of the ten polysemes' different interpretations as used in Dölling (2020) and included an 'other' category label. The resulting full list of class labels presented to annotators was



Classify The Highlighted Words

Carefully read each sentence and classify the **highlighted** word or expression by selecting one or more labels for it.
If you cannot see the submit button, scroll down the page.

The **chicken** was too dry for my taste.

The **highlighted** word or expression belongs to the following one or more categories:

□ animal
□ building
□ container
□ data or information
□ event or process
□ foodstuff
□ group of people
□ institution
□ liquid
□ literary work
□ medium
□ organisation
□ person
□ physical object
□ representative
□ other

Figure 4.4: Screenshot of the the AMT interface for the word sense class annotation task.

(39)
0 animal
1 building
2 container
3 data or information
4 event or process
5 foodstuff
6 group of people
7 institution
8 liquid
9 literary work
10 medium
11 organisation
12 person
13 physical object
14 representative
15 other

Class labels were presented without any examples to allow for a subjective interpretation of the given classes. A screenshot of the the AMT interface for this task is displayed in Figure 4.4. As here we were only interested in the labels assigned to polysemous test items, we did not include any homonym, synonym or filler items in the sense class annotation experiment. Target expressions again were highlighted in bold font, and distributed over 15 questionnaires with 10 items each. Annotators were given the following set of minimal instructions:

(40) Carefully read each sentence and classify the highlighted word or expression by selecting one or more labels for it.

As annotators were asked to classify the highlighted target expression by selecting all applicable labels, submissions were stored in 16-dimensional, binary multi-hot vectors indicating the selection of labels together with the worker's unique ID.

4.4.4 Computational Approaches

In order to assess word sense similarity encoded in contextualised embeddings, we extracted target word embeddings from the different disambiguating contexts and calculated their pairwise cosine similarity (1-cosine). In the pilot run, we tested embeddings generated with ELMo (Peters et al., 2018), BERT Base (Devlin et al., 2019), GPT-2 (Radford et al., 2018) and a static baseline based on Word2Vec (Mikolov et al., 2013a,b).

ELMo. In our experiments we used the default, pretrained ELMo implementation available on TensorFlow Hub[10] and extracted vectors from the LSTM's second layer hidden state (which has previously been shown to best represent semantic information, see e.g. Ethayarajh, 2019) at the index of the target to represent the target expression's embedding. In case the target expression spanned multiple words, we averaged the vectors of all words included to generate a single vector.

BERT. For the pilot, we used the default, cased BERT Base (12 layers, hidden state size of 768) implementation available at TensorFlow Hub.[11] We extracted three different types of contextualised embeddings from BERT: the pooled sentence embedding (SE), the final layer encoding at the indexes of the target expression (WE), and the embedding of the special classification token (CLS). In case a target expression consisted of multiple sub-word tokens, we again averaged their vectors to obtain a single vector representation in the case of word-level embeddings.

GPT-2. In the data collection pilot we also tested a pretrained implementation of GPT-2 (see Chapter 3.3.3), but excluded this model from our analysis, as due to its traditional left-to-right processing, it produced the same embedding for all context samples using the same target expression. This is because all of our samples start with 'The *[target]...*', mentioning the target without any preceding context, leading the left-to-right model to produce identical target embeddings before processing the subsequent context. This limitation could be overcome by either applying GPT-2 bi-directionally, encoding a sample sentence left-to-right and right-to-left, or by adding future token attention (see e.g. Lawrence et al., 2019), but since we were mainly

[10] `https://tfhub.dev/google/ELMo/3`, see Chapter 3.3.1 for more information

[11] `https://tfhub.dev/tensorflow/bert_en_cased_L-12_H-768_A-12/4`, see Chapter 3.3.2

interested in the off-the-shelf capabilities of the contextualised language models, we excluded GPT models from our evaluation for now.

BASELINE. Lastly, we established a baseline computational word sense similarity score by averaging over the static Word2Vec encodings (see Chapter 3.1.1) of all words in a sample context to create a naive contextualised embedding from static vectors. For our experiments, we used the pre-trained Word2Vec vectors stored in the Gensim implementation of the model.[12]

4.5 Analysis

We report the collected data and the results obtained from our experiments in three steps: first, we analyse in detail the collected human annotations of sense similarity, investigating whether the three tested measures are sensitive to expressing nuances in the annotators' interpretation of polysemous expressions. Detecting measurable differences in polysemic interpretations is an important first step in further investigating sense similarity as an underlying factor of the mental representation of polysemes, and could be used in the testing of hypotheses proposing distance based clustering or grouping of polysemic sense extensions. In a second step we will then analyse how well the different contextualised language models correlate with the human annotated scores, and how well they perform in predicting the different types of measures. In step three - by means of a sanity check - we analyse to what degree the different annotation-based and computational similarity metrics can predict whether the two interpretations of a target expression given in two different sample sentences invoke different meanings (i.e. homonymy) or different senses (i.e. polysemy) to test their sensitivity to this relatively stark distinction.

4.5.1 Word Sense Similarity Judgements

We collected 20 judgements for each explicit word sense similarity questionnaire. 65 individual annotators contributed to the study, with HITs taking an average of 133 seconds (median=90s). Through filtering out any submissions that rated at least two filler samples higher than 0.66 or the synonym sample lower than 0.33, we removed 9 submissions and retained at least 18 judgements per item.

As a first step, we calculated the overall means of word similarity judgements for all polyseme, homonym, synonym and filler sentence pairs in the dataset to determine any principled differences among these groups. In order of decreasing

[12] `https://radimrehurek.com/gensim/models/word2vec.html`

mean similarity, synonym sentence pairs obtained a mean similarity rating of 0.9040 (std=0.1622), polyseme sentence pairs a mean of 0.8711 (std=0.2162), homonym pairs a mean of 0.1274 (std=0.2908) and filler sentence pairs a mean of 0.0280 (std=0.0805). We then used Student's t-test to compare the distributions of judgements, which indicated that the polyseme and synonym distributions each are significantly different from all other distributions ($p<0.05$). This means that annotators rated synonyms (i.e. different words with similar meaning) to be overall more similar to each other than polysemes (i.e. identical words with different sense interpretations). The t-tests revealed no significant difference in the distribution of homonym and filler item ratings, but both of these sample types were rated significantly lower in sense similarity than the synonymic and polysemic items tested in this study. Together, these observations provide preliminary support for a representation of polysemes occupying a unique middle ground between identity of meaning on the one hand, and homonymy (and expressions with unrelated meanings) on the other.

Next, our analysis focused on the polysemic test items, and we split judgements for sentence combinations invoking the same sense interpretation (same-sense samples), and those that invoke different sense interpretations (cross-sense samples). Figure 4.5 visualises the distribution of explicit sense similarity judgements for polysemic items after this split, as approximated by a Kernel Density Estimate (KDE). A KDE considers a given window (also called bandwidth) and produces a local distribution of the amount of encountered data points within that window. The smaller the window, the more detailed is the local distribution (but prone to over-fit the data); the larger the window, the more interpolated is the resulting estimate (but likely to under-fit the data). By default, we use Scott's rule (Scott, 1992) to automatically derive the window size for the KDE plots, but we also manually reduce the window size (i.e. lower the bandwidth smoothing) to show a more fine-grained estimate.

When considering all individual annotations (top), the two distributions are relatively similar, with cross-sense samples exhibiting a marginally wider tail below the 0.9 similarity mark. When considering annotation means (bottom), the differences between the two distributions are more clearly pronounced. Here, same-sense context combinations exhibit a defined spike close to the perfect similarity rating at 1, followed by additional modes around 0.9, 0.8 and, interestingly, 0.6. The modes of the cross-sense distribution are less stark, indicating a wider distribution with modes around 0.95, 0.75 and 0.65.

Because the twelve different polysemous target expressions used in this study each represent a different type of regular polysemy, we next split the collected

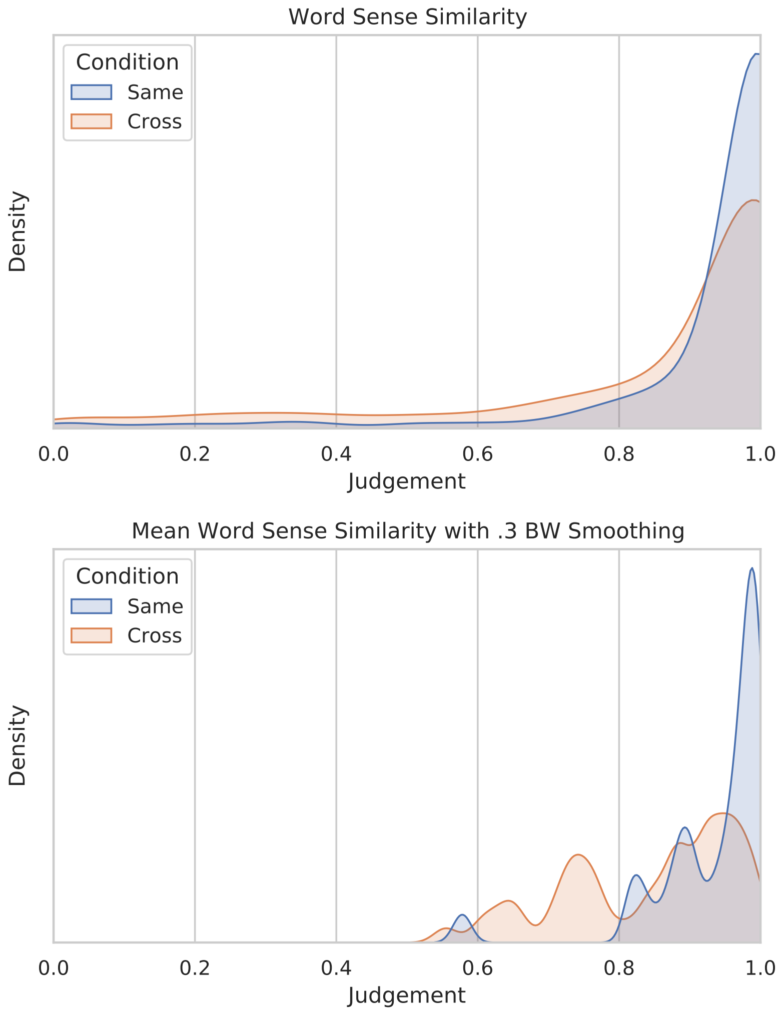


Figure 4.5: Top: Kernel Density Estimate (KDE) of the distribution of explicit word sense similarity judgements for the pilot polyseme samples. Ratings for same-sense co-predication structures are shown in blue, cross sense structures in orange. KDE bandwith determined by Scott's rule. Bottom: KDE of the distribution of mean word sense similarity judgements under manual bandwidth smoothing of 0.3.

	All		Same Sense		Cross Sense	
Polyseme	**Mean**	**std.**	**Mean**	**std.**	**Mean**	**std.**
Newspaper (3)	0.8599	0.2616	0.9922	0.0222	0.7915	0.2998
Hemingway (2)	0.9213	0.2030	0.9647	0.1582	0.8800	0.2304
War and Peace (2)	0.9506	0.1784	0.9960	0.0251	0.9285	0.2134
Lunch (2)	0.8968	0.1900	0.9683	0.1098	0.8236	0.2240
Door (2)	0.9649	0.1258	0.9855	0.0591	0.9432	0.1670
DVD (3)	0.9136	0.1786	0.9554	0.1274	0.8920	0.1965
Chicken (2)	0.7312	0.2437	0.7447	0.2462	0.7183	0.2406
School (4)	0.9059	0.2054	0.9691	0.0774	0.8844	0.2296
Wine (2)	0.9527	0.1376	0.9924	0.0385	0.9141	0.1814
Glass (2)	0.7079	0.3599	0.7482	0.3381	0.6666	0.3766
Construction (2)	0.7773	0.2937	0.8885	0.2033	0.6633	0.3266
Overall	0.8711	0.2162	0.9277	0.1277	0.8278	0.2442

Table 4.1: Polysemic target expression (number of regular senses), together with means and standard deviations of all pairwise sense similarity judgements, and same-sense and cross-sense samples only.

judgements based on the target expression and calculated the mean sense similarity judgements for same-sense and cross-sense sentence pairs. Table 4.1 displays these numbers, showing that same-sense means are consistently higher than the cross-sense ones, and except for *chicken*, *glass* and *construction* range above 0.95 (i.e. higher than the synonym mean). This means that barring these three outliers, the generated same-sense pairs were rated as invoking an almost identical interpretation of the polysemic target expression. The average similarity of cross-sense pairs often ranges between 0.8 and 0.9, showing a high similarity still, but indicates that not all cross-sense pairs seem to be perceived as invoking the same sense.

Qualitative Analysis

Turning to a more qualitative analysis, we investigated the similarity ratings obtained for sentence pairs containing a specific target expression to assess whether the collected data provides any evidence for sense clustering as proposed by Ortega-Andrés and Vicente (2019). Since it is difficult to collapse results over the different types of polysemes tested, we here exemplify our analyses through a summary of the observations concerning polyseme *newspaper* and draw parallels to other test items where possible.

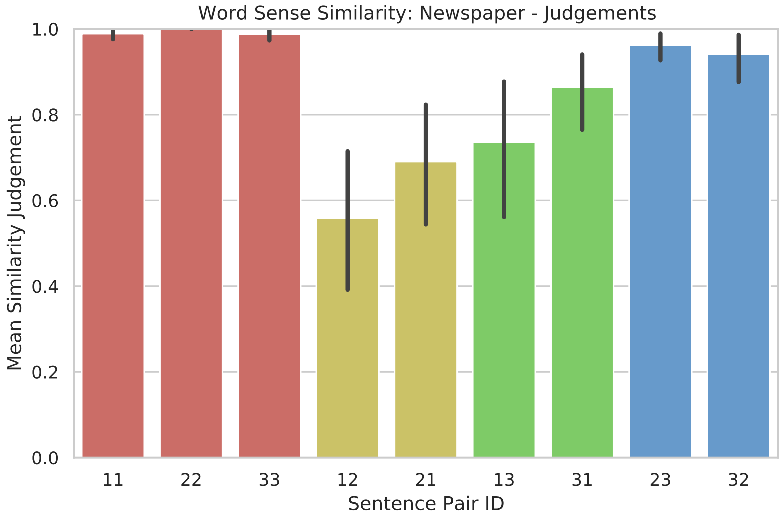


Figure 4.6: Similarity judgements for sentence pairs containing the polyseme *newspaper*. The two numbers in the sentence pair IDs indicate the combination of senses. The first three bars thus indicate same-sense pairs, the other three groups the different variations of cross-sense samples. Senses: 1-organisation, 2-physical, 3-information.

As mentioned above, polyseme *newspaper* was taken to invoke three distinct but related senses; (1) *organisation/institution*, (2) *physical object* and (3) *information/data*. Creating all combinations of senses generates the following nine sense pairs indicated by their sense number:[13]

(41) 11 organisation/organisation
22 physical/physical
33 information/information

12 organisation/physical
21 physical/organisation

13 organisation/information
31 information/organisation

23 physical/information
32 information/physical

Figure 4.6 shows the mean word similarity judgements for these nine sentence pairs. The three same-sense pairs 11, 22, and 33 (red) receive mean similarity ratings close to 1, indicating that in these cases annotators indeed perceive the target word contexts to invoke exactly the same sense in both sample sentences. This effect can be observed for all tested polysemes except for *glass*, where one of the same-sense pairs

[13] see Appendix A.1 for the full list of sample sentences.

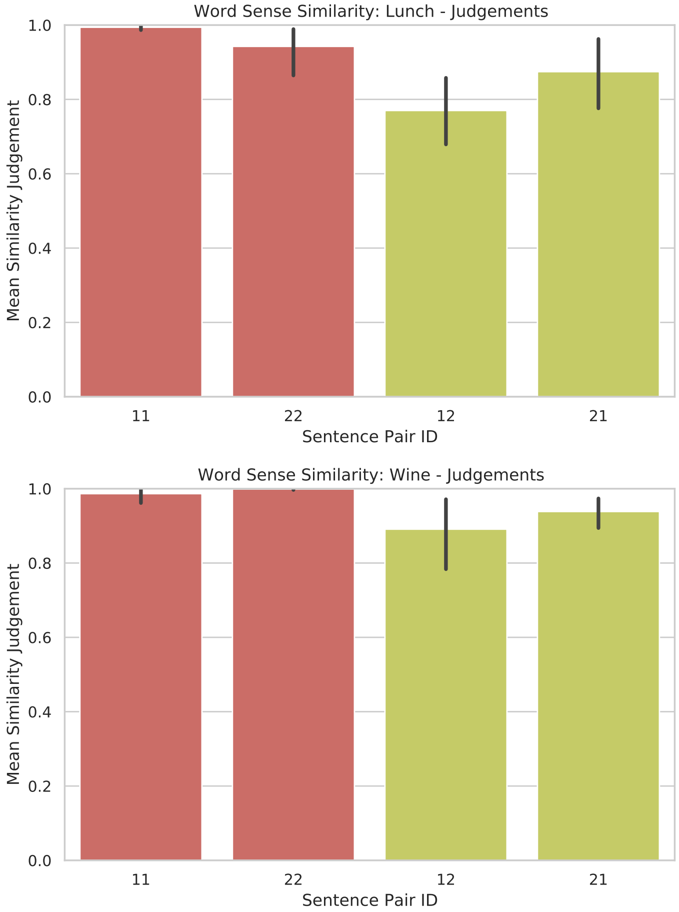


Figure 4.7: Similarity judgements for sentence pairs containing the polysemes *lunch* and *wine*, respectively. Same-sense pairs in red (left), cross-sense pairs in yellow. *Lunch* senses: 1-food, 2-event. *Wine* senses: 1-container, 2-content.

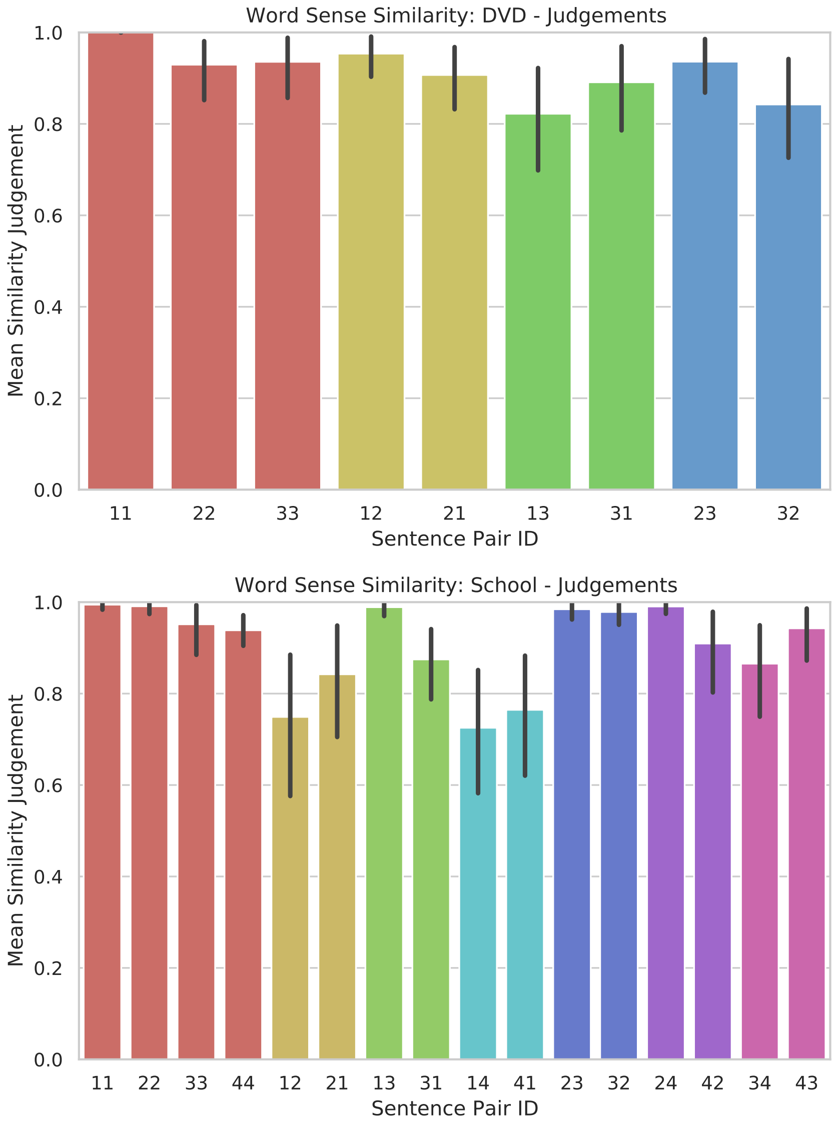


Figure 4.8: Similarity judgements for sentence pairs containing the polysemes *DVD* and *school*, respectively. Same-sense pairs in red (left), cross-sense pairs in other colours. *DVD* senses: 1-physical, 2-content, 3-medium. *School* senses: 1-building, 2-administration, 3-institution, 4-students/faculty.

does not actually seem to elicit the same sense (rated at a similarity of 0.48) and a same-sense pair for *construction* which only received a similarity score of 0.82 (being higher still than the cross-sense pairs). Returning to *newspaper*, all six cross-sense pairs receive lower ratings than the same-sense pairs: both, the *organisation/physical* sentence pairs 12 and 21 (yellow), and the *organisation/information* sentence pairs 13 and 31 (green) receive significantly lower similarity ratings than the same-sense pairs. The similarity ratings for the *physical/information* pairs 23 and 32, (blue) are ranging between 90 and 100, being significantly higher than the ratings for pairs 12, 21, 13, but significantly lower than same sense-sense pair 22. This indicates that at least between the *organisation* and *physical* sense interpretation there seems to be a notable difference in meaning, while the *information* readings are judged to be relatively similar to either - however not to a level that same-sense sample pairs are similar to each other.

We see a similar but less pronounced effect for most of the other tested polysemes, where cross-sense samples usually are rated to be less similar to each other than same-sense samples, but except for significant differences between the *building* and *administration* and *institution* senses of polyseme *school*, none of these differences reach significant levels. Figure 4.7 contains a visualisation of the word sense similarity ratings for two target expressions with two regular senses each, *lunch* and *wine*, and Figure 4.8 displays the mean judgements for targets *DVD* and *school*, with three and four polysemic sense interpretations, respectively.

Predication Order

Returning to the *newspaper* samples, a second point of interest are the notable although non-significant differences in similarity ratings for sentence pairs 12 and 21, and 13 and 31, respectively. Since these sentence pairs were created to invoke the same pair of (cross-sense) interpretations, it is noteworthy that their ratings differ so much. This difference can be the result of two factors: i) the sentence pairs contain different sample sentences, which within the same sense interpretation could evoke interpretation differences, and ii) the order of presentation for the two sentence pairs is different, and presentation order is known to induce biases and affect acceptability in co-predication studies. To control for the latter, we repeated our experiments with the same set of samples, but inverting the presentation order within the sentence pairs. Based on an average of ten judgements, only one of the 67 test items' similarity ratings changed significantly, indicating that the observed difference in similarity ratings is not an effect of presentation order, but indeed due to subtle interpretation differences in the contexts used to elicit a certain sense.

This means that even after spending a considerable amount of effort on creating samples that as clearly as possible invoke a certain reading, participants sometimes interpreted samples (at least slightly) differently than intended.

4.5.2 Co-predication Acceptability Judgements

We collected 30 annotations for each questionnaire containing ten co-predication structures. 76 individual annotators contributed to the study, with HITs taking an average of 146 seconds (median=93s). Through filtering out any submissions that rated at least four filler samples higher than 0.66 or the synonym sample lower than 0.33, we removed 14 submissions and retained at least 20 judgements per item.

As a first step, we again calculated the overall means of the acceptability judgements for all polyseme, homonym, synonym and filler co-predication structures in the dataset to determine any principled differences among these groups. In order of decreasing mean acceptability, co-predication structures combining synonym samples obtained a mean acceptability rating of 0.8869 (std=0.1909), polyseme context combinations a mean acceptability of 0.7060 (std=0.3052), filler items a mean of 0.3658 (std=0.3591), and co-predication structures with homonym contexts a mean of 0.2936 (std=0.3442).

A range of pairwise Student's t-tests indicated that each distribution was significantly different to all other distributions (all $p<0.05$). This means that also in the co-predication setting, annotators rated the acceptability of structures containing synonyms to be higher than that of structures containing polysemes. Note however that due to the method of conjunction reduction used to create co-predication structures from individual sample contexts, the resulting structures only contain one target word, which means that synonym items here could be considered as same-sense items:

(42) 1 The computer suddenly turned off.
2 The PC needs to be replaced soon.
12 The computer suddenly turned off and needs to be replaced soon.

For the distribution of co-predication acceptability ratings, the t-tests also indicated a significant difference between the distribution of homonym and filler item ratings, which means that here filler items were rated significantly more acceptable than co-predication structures containing homonyms. Together, these observations indicate the effect of sense similarity on the co-predication acceptability ratings: synonyms produce the highest acceptability scores, with both contexts eliciting the same sense - originally for different targets, but here as the result of conjunction

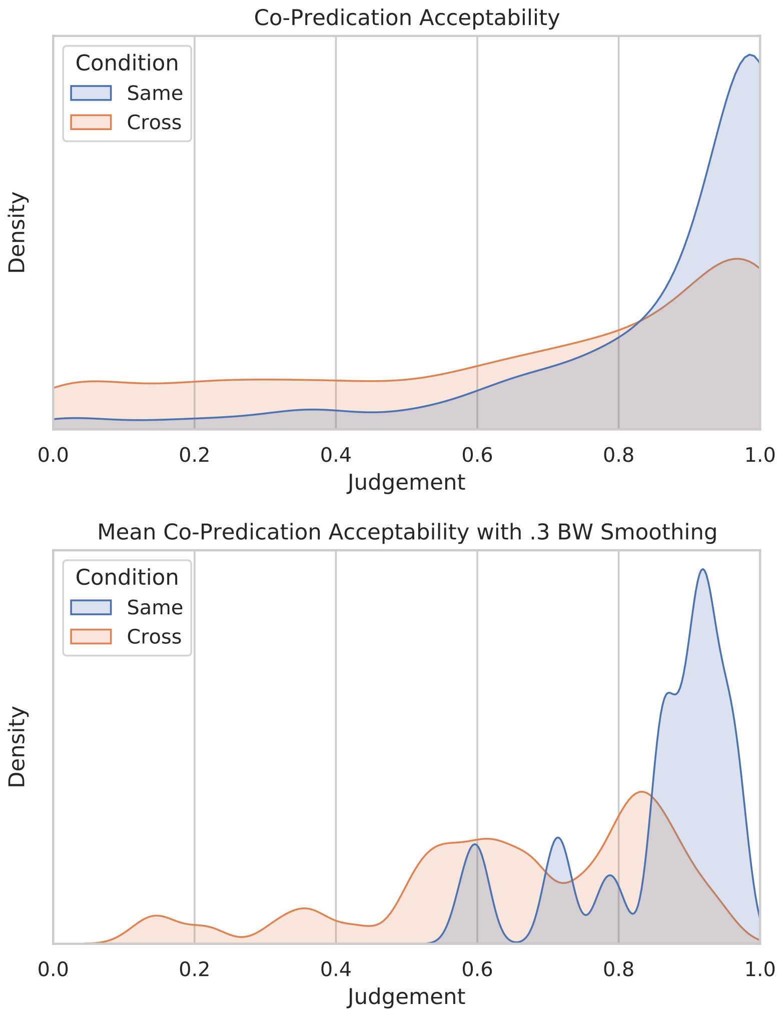


Figure 4.9: Top: Kernel Density Estimate (KDE) of the distribution of co-predication acceptability judgements for the pilot polyseme samples. Ratings for same-sense co-predication structures are shown in blue, cross sense structures in orange. Bottom: KDE of the distribution of mean co-predication acceptability judgements under bandwidth smoothing of 0.3.

reduction for a single target. Polysemes again seem to not fully duplicate this effect of identity of meaning, indicating that some combinations of polysemic sense interpretations can lead to relatively lower acceptability judgements. Co-predication structures containing two different interpretations of a homonym finally are overall rated lowest, which is in line with the original use of the co-predication test in determining homonymy as infelicity under co-predication. Using a graded rating scale, our results even suggest that homonymic items are less acceptable under co-predication than filler items which combine random sample contexts.

Focusing on the polysemic items, we again separated judgements for co-predication structures invoking the same sense interpretation in both predications from those

	All		Same Sense		Cross Sense	
Polyseme	**Mean**	**std.**	**Mean**	**std.**	**Mean**	**std.**
Newspaper	0.6861	0.3346	0.8379	0.2239	0.5906	0.3567
Hemingway	0.7989	0.2932	0.9529	0.0878	0.6583	0.3407
War and Peace	0.8549	0.2144	0.9343	0.1316	0.8208	0.2333
Lunch	0.7648	0.2634	0.7402	0.2665	0.7884	0.2583
Door	0.6427	0.3400	0.7900	0.2714	0.5599	0.3466
DVD	0.6863	0.3240	0.8315	0.2220	0.6041	0.3432
Chicken	0.3952	0.3936	0.5605	0.4060	0.2376	0.3076
School	0.6812	0.3239	0.7872	0.2814	0.6320	0.3305
Wine	0.8551	0.2251	0.9276	0.1447	0.7857	0.2633
Glass	0.6468	0.3387	0.6794	0.3241	0.6150	0.3494
Construction	0.7542	0.3058	0.7467	0.3765	0.7597	0.2408
Overall	0.7060	0.3052	0.7990	0.2487	0.6411	0.3064

Table 4.2: Polysemic target expression (number of regular senses), together with means and standard deviations of all pairwise co-predication acceptability judgements, and same-sense and cross-sense samples only.

that invoke different ones. Figure 4.9 visualises the distribution of co-predication acceptability judgements after this split. When considering all individual annotations (top), it becomes immediately clear that the cross-sense distribution here has a much wider tail then the same-sense one. When considering annotation means (bottom), the differences between the two distributions again are more clearly pronounced, showing the overall lower mean acceptability of cross-sense samples. Here we also observe spikes in mean acceptability around the 0.6 and 0.7 mark in the same-sense distribution, indicating that some of the context pairs intended to invoke the same meaning result in relatively low acceptability scores for the resulting co-predication structure.

We next split the collected judgements based on the target expression and calculated the mean co-predication acceptability judgements for same-sense and cross-sense sentence pairs. Table 4.2 displays these numbers, showing that same-sense means are higher than the cross-sense ones for ten of the targets, with *lunch* and *construction* forming the exception. In three cases, the same-sense mean is higher than the synonym mean of 0.8869 (*Hemingway*, *War and Peace* and *wine*). The average similarity of cross-sense pairs usually ranges between 0.6 and 0.8, with a mean similarity of 0.2376 for the cross-sense reading of *chicken* presenting the lowest over-

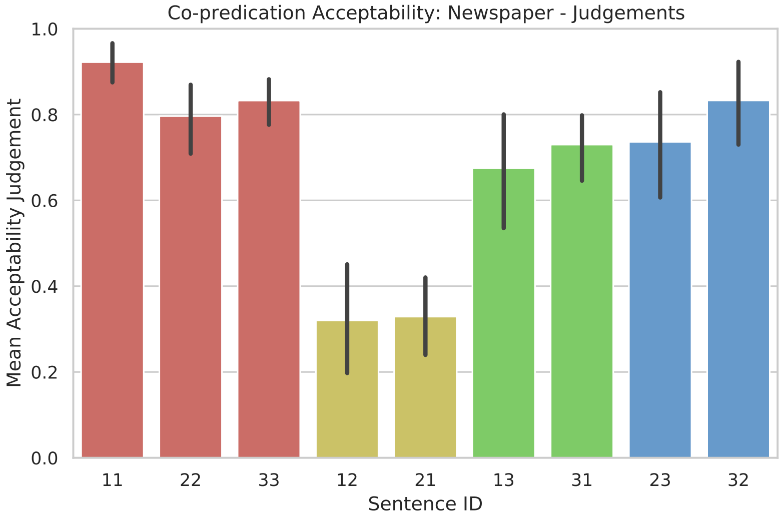


Figure 4.10: Acceptability judgements for co-predication structures containing the polyseme *newspaper*. Senses: 1-organisation, 2-physical, 3-information.

all mean acceptability score. Compared to the explicit similarity ratings presented in the previous section, mean acceptability scores are overall lower and exhibit a clearer distinction between same-sense and cross-sense means, with a difference of about 15 points between the overall same-sense and cross sense means here, and a difference of 10 points in the explicit similarity means.

Qualitative Analysis

Moving to a more qualitative analysis of the collected co-predication acceptability judgements, Figure 4.10 shows the mean acceptability judgements for the co-predication structures combining *newspaper*'s different sense interpretation (compare to Figure 4.6 showing the samples' explicit similarity ratings). The three same-sense structures 11, 22, and 33 (red) here are rated less consistently than in the word sense similarity setting, receiving acceptability ratings roughly between 0.8 and 0.9. With these scores, they are however still rated as more acceptable than most of the cross-sense structures, and significantly so in some cases. *Organisation/physical* cross-sense samples 12 and 21 again stand out at close to 0.3, receiving much lower acceptability ratings than any of the other cross-sense pairings, which all score above 0.65. This replicates our previous observation that at least between the *organisation* and *physical* sense interpretation there seems to be a notable difference in meaning and therefore reduced felicity under co-predication, while the *information* readings are judged to be relatively acceptable combined with either *physical* or *organisation* predications - and in some cases not significantly less so than same-sense structures.

We observe similarly pronounced drops in acceptability for some sense combi-

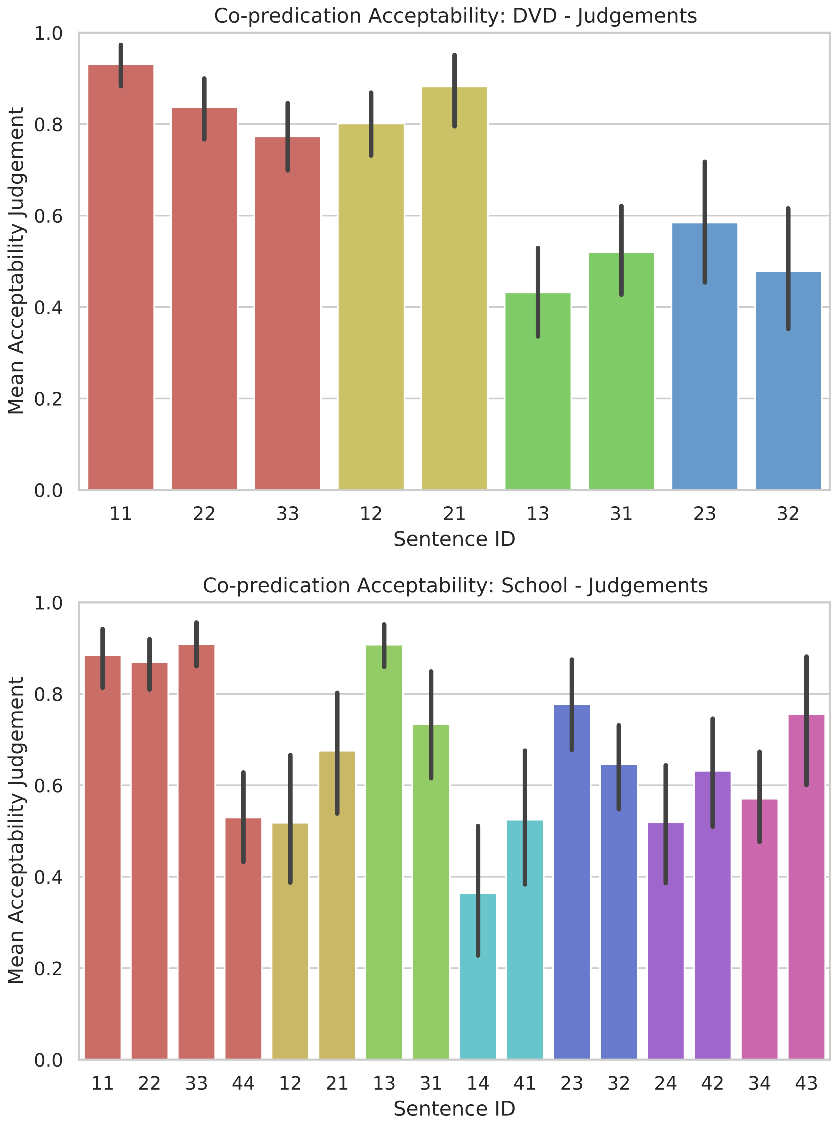


Figure 4.11: Mean acceptability judgements for co-predication structures containing the polysemes *DVD* and *school*, respectively. *DVD* senses: 1-physical, 2-content, 3-medium. *School* senses: 1-building, 2-administration, 3-institution, 4-students/faculty.

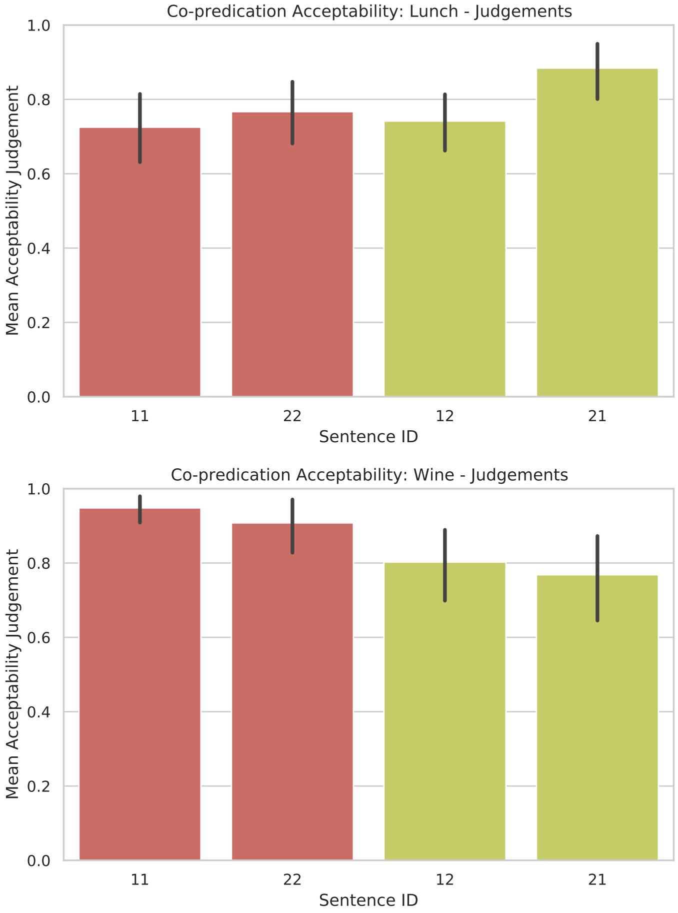


Figure 4.12: Mean acceptability ratings for co-predication structures containing the polysemes *lunch* and *wine*, respectively. *Lunch* senses: 1-food, 2-event. *Wine* senses: 1-container, 2-content.

nations of multi-sense targets *DVD* and *school*; in *DVD* for the *physical/medium* combinations (13 and 31) and the *content/medium* combinations (23 and 32), and for school most notably for any combinations including *student* interpretations (sense 4). Overall, the acceptability ratings for these targets correspond well with the explicit similarity judgements (see Figure 4.8 for reference), but seem to be more pronounced here than in the explicit sense comparison.[14]

Revisiting *lunch* and *wine* (Figure 4.12), we observe that *lunch*'s same-sense co-predication structures receive relatively low acceptability ratings, with the cross-sense samples 12 and 21 receiving a comparable and even a higher acceptability score. Example (43) shows the co-predication structures 11 and 22 annotated in the pilot experiment. While sample 22 arguably exhibits a temporal mismatch in its two predications reducing the structure's overall acceptability, we cannot directly pinpoint any issues with sample 11.

(43) 11 Lunch was exceptionally delicious today but got cold while we waited for someone.

22 Lunch took more than an hour yesterday and is great for socialising and networking.

The acceptability judgements for target *wine* for example again more resemble the ratings collected in the word sense similarity setting (compare Figure 4.7) with same-sense means consistently higher than cross-sense means.

Predication Order

Like with the explicit sense similarity judgements, we investigated whether the order of the two sample contexts combined into a co-predication structure had any effect on its acceptability. Given previous observations of predication order affecting co-predication acceptability (see e.g. Murphy, 2019), we expected that here sample ordering effects might be larger than for the explicit sense similarity judgements, but hoped that the careful crafting of sample contexts would have kept its impact to a minimum. To test for predication order effects, we again also collected co-predication acceptability judgements for co-predication structures with the same predications as in the original data, but presented in inverse order. As an example, *newspaper* item `11` was changed as follows:

(44) **Original**

11 The newspaper fired its editor in chief and was sued for defamation.

[14] Note that for *DVD*, some cross-sense samples receive higher acceptability scores than the same-sense ones. We will investigate these outliers in more detail in Chapter 5.3.1.

	All		**Same Sense**		**Cross Sense**	
Polyseme	**Mean**	**std.**	**Mean**	**std.**	**Mean**	**std.**
Newspaper (3)	0.5904	0.4077	0.9860	0.0071	0.3926	0.3632
Hemingway (2)	0.7640	0.2228	0.9853	0.0014	0.5427	0.0369
War and Peace (2)	0.9457	0.0547	0.9939	0.0050	0.9216	0.0523
Lunch (2)	0.7447	0.2477	0.9800	0.0052	0.5095	0.1095
Door (2)	0.9983	0.0017	0.9983	0.0017	0.9983	0.0017
DVD (3)	0.9510	0.0447	0.9919	0.0049	0.9306	0.0417
Chicken (2)	0.7154	0.2804	0.9832	0.0138	0.4476	0.1167
School (4)	0.7284	0.2870	0.9693	0.0252	0.6480	0.2896
Wine (2)	0.8722	0.1198	0.9908	0.0013	0.7536	0.0239
Glass (2)	0.9262	0.0413	0.9520	0.0272	0.9004	0.0365
Construction (2)	0.7828	0.2162	0.9781	0.0007	0.5876	0.1315
Overall	0.8200	0.1750	0.9826	0.0085	0.6939	0.1094

Table 4.3: Polysemic target expression (number of regular senses), together with means and standard deviations of all pairwise word class overlaps, and same-sense and cross-sense samples only.

Inverse

11 The newspaper was sued for defamation and fired its editor in chief.

We then compared the distribution of ratings given to an original item with those assigned to the inverse, finding that in only 8 of the 67 comparisons the inverted items had received a significantly different rating. We therefore again concluded that our samples were constructed sufficiently carefully to allow a focus on the impact of the actual predication and combination of senses without too much influence of confounding factors.

4.5.3 Word Class Similarity Judgements

We collected 15 word sense class annotations for each sample sentence, incidentally provided by exactly 15 individual workers - i.e. each individual worker completed all 15 questionnaires. HITs took an average of 178 seconds (median=107s). Classification results were not filtered, but averaged per item in order to create word sense class vectors. Pairwise sense class similarity was then calculated through the cosine similarity (1-cosine) between the different combinations of sense interpretations, i.e. the overlap in their averaged multi-class assignments.

Table 4.3 contains the mean class overlap scores for all combinations of polysemic

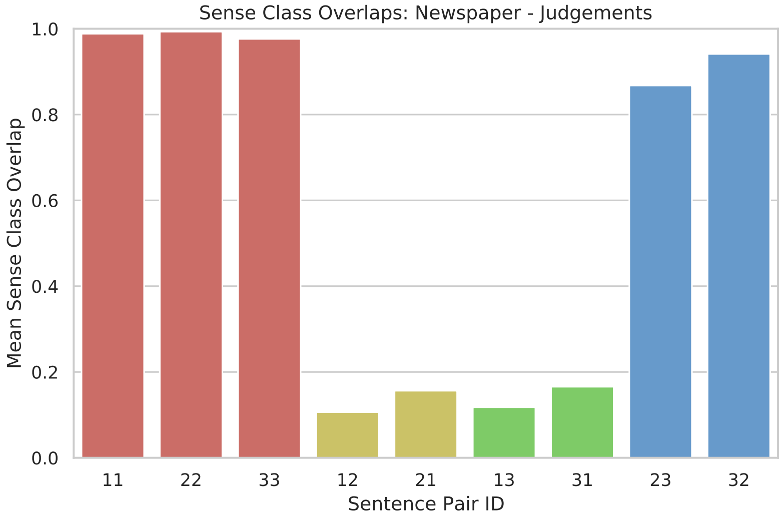


Figure 4.13: Word class overlap for sentence pairs containing the polyseme *newspaper*. Senses: 1-organisation, 2-physical, 3-information.

sample sentences, as well as those of only the same-sense and cross-sense pairings. Overall, word class overlap displays the highest difference between same-sense and cross-sense means of the three measures collected, indicating a 27 point gap. Same-sense pairings received consistently close to perfect overlap scores (lowest overlap score is 0.95 for *glass*), and cross-sense samples display a high variance, with mean overlap scores ranging from as low as 0.39 for *newspaper* to a perfect overlap score for the cross-sense samples of *door*.

Qualitative Analysis

Investigating these differences in more detail, Figure 4.13 shows the pairwise overlap scores for different sample combinations containing target *newspaper*, and Figure 4.14 and 4.15 display the class overlap scores for targets *DVD* and *school*, and *lunch* and *wine*, respectively.[15] Starting with *newspaper*, we note the close to perfect overlap scores for the same-sense comparisons 11, 22 and 33, as well as the very low overlap scores for cross-sense comparisons containing the *organisation* reading 1 (all below 0.2) that stand in stark contrast to the relatively high overlap scores for the *physical*/*information* combinations 23 and 32 with calculated overlap scores of well over 0.8. The class overlap ratings for target *DVD* in Figure 4.14 do not exhibit these stark differences, with all three cross-sense combinations receiving overlap scores above 0.8, and the *content*/*medium* comparisons 23 and 32 getting near-perfect scores similar to the same-sense items. The overlap ratings for *school* on

[15] As overlap calculations are based on the cosine similarity of the averaged class assignments (i.e. a single calculation) we cannot compute variance intervals for these ratings.

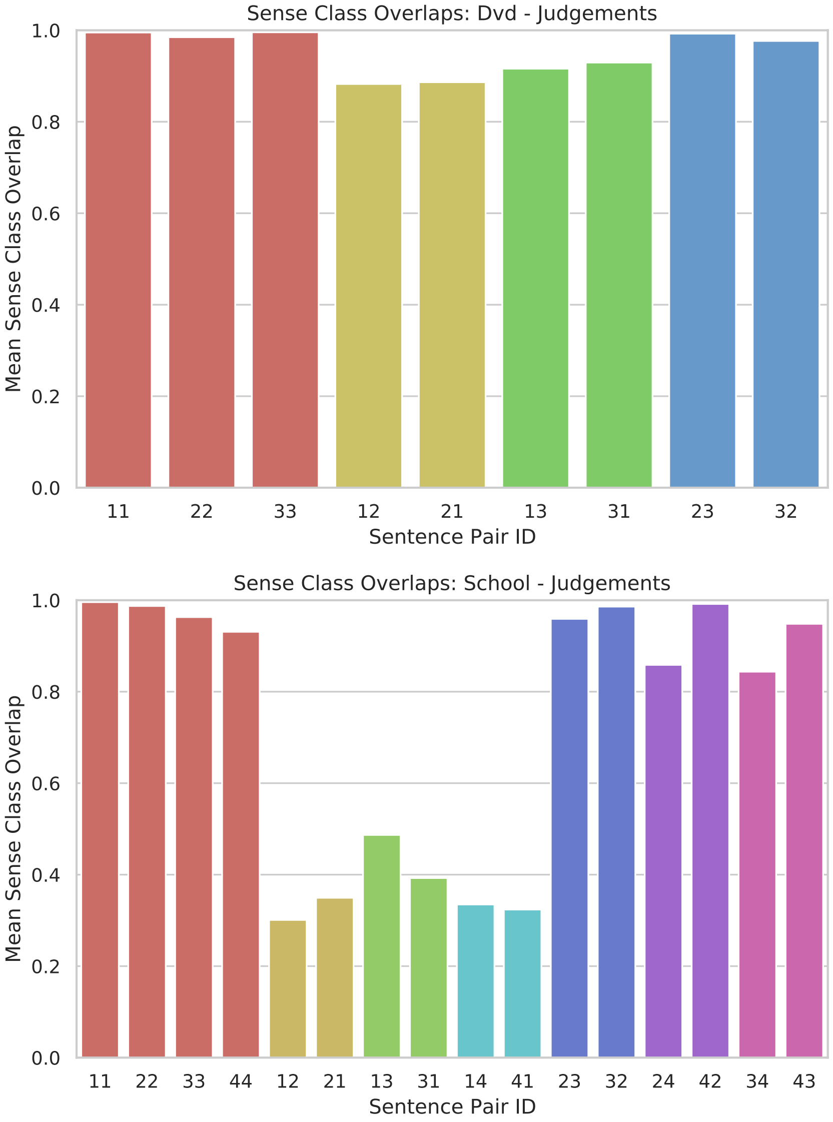


Figure 4.14: Word class overlap for sentence pairs containing the polysemes *DVD* and *school*, respectively. *DVD* senses: 1-physical, 2-content, 3-medium. *School* senses: 1-building, 2-administration, 3-institution, 4-students/faculty.

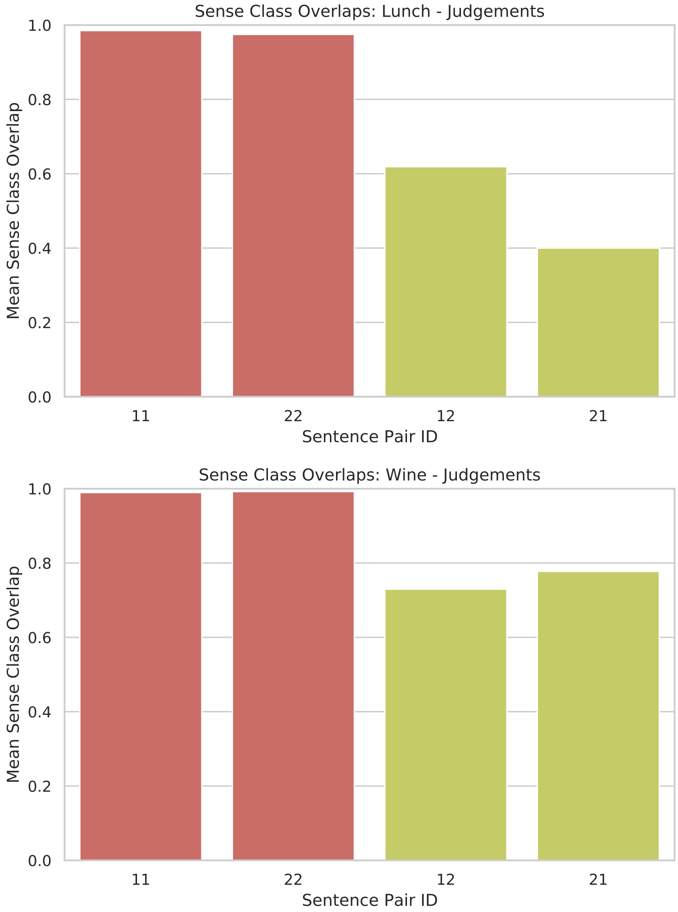


Figure 4.15: Word class overlap for sentence pairs containing the polysemes *lunch* and *wine*, respectively. *Lunch* senses: 1-food, 2-event. *Wine* senses: 1-container, 2-content.

Combination		**Correlation**		**OLS Regression Analysis**				**Prediction**	
First Measure	**Second Measure**	**r**	**p**	**Coef.**	**R^2**	**F-stat.**	**Prob.**	**MSE**	**R^2**
Similarity	**Acceptability**	0.529	2.08E-06	0.910	0.280	26.855	2.08E-06	0.040	0.208
Similarity	**Sense Class**	0.539	1.21E-06	1.091	0.291	28.320	1.21E-06	0.057	0.162
Acceptability	**Similarity**	0.529	2.08E-06	0.308	0.280	26.855	2.08E-06	0.014	0.149
Acceptability	**Sense Class**	0.563	3.21E-07	0.662	0.317	32.015	3.21E-07	0.050	0.301
Sense Class	**Similarity**	0.539	1.21E-06	0.267	0.291	28.320	1.21E-06	0.014	0.175
Sense Class	**Acceptability**	0.563	3.21E-07	0.479	0.317	32.015	3.21E-07	0.037	0.258
BERT WE	**Similarity**	0.211	0.077	0.762	0.045	3.226	0.077	0.018	-0.214
BERT WE	**Acceptability**	0.482	0.000	2.991	0.233	20.936	0.000	0.041	0.204
BERT WE	**Sense Class**	0.221	0.064	1.614	0.049	3.553	0.064	0.069	-0.007
BERT CLS	**Similarity**	-0.038	0.756	-0.390	0.001	0.097	0.756	0.019	-0.298
BERT CLS	**Acceptability**	0.271	0.023	4.832	0.073	5.448	0.023	0.049	0.033
BERT CLS	**Sense Class**	0.051	0.672	1.075	0.003	0.181	0.672	0.073	-0.051
BERT SE	**Similarity**	-0.007	0.955	-0.067	0.000	0.003	0.955	0.020	-0.322
BERT SE	**Acceptability**	0.011	0.929	0.181	0.000	0.008	0.929	0.058	-0.162
BERT SE	**Sense Class**	-0.016	0.895	-0.317	0.000	0.018	0.895	0.073	-0.067
ELMo WE	**Similarity**	0.295	0.012	1.191	0.087	6.600	0.012	0.018	-0.188
ELMo WE	**Acceptability**	0.178	0.138	1.233	0.032	2.257	0.138	0.051	-0.015
ELMo WE	**Sense Class**	0.323	0.006	2.630	0.104	8.022	0.006	0.065	0.063
Word2Vec SE	**Similarity**	0.053	0.662	0.085	0.003	0.193	0.662	0.020	-0.305
Word2Vec SE	**Acceptability**	0.245	0.039	0.681	0.060	4.423	0.039	0.051	-0.006
Word2Vec SE	**Sense Class**	0.249	0.036	0.813	0.062	4.555	0.036	0.070	-0.026

Table 4.4: Correlations between the three different metrics of word sense similarity based on annotation judgements, and correlation between computational proxies of word sense similarity as compared to the human judgements. The first set of columns displays pairwise correlation based on Pearson's r, the second set shows the key statistics obtained from their OLS regression, and the third set contains the mean regression scores based on 5-fold cross validation.

the other hand do again show a clear pattern of differences in sense class overlap, with all combinations including the *building* sense 1 receiving significantly lower overlap scores than any of the other cross- or same-sense comparisons. A second noteworthy observation here are the clear differences between the overlap scores of 24 and 42, and 34 and 43, respectively. As samples were rated individually, there are no potential order effects here, and differences in class overlap must be attributed to the sample sentences invoking slightly different sense interpretations, or invoking their intended sense interpretation to varying degrees.

4.5.4 Comparison of Human Annotation Measures

In order to establish a measure of correlation between the three human annotation metrics, we consider all six combinations of metrics and i) calculate their Pearson's

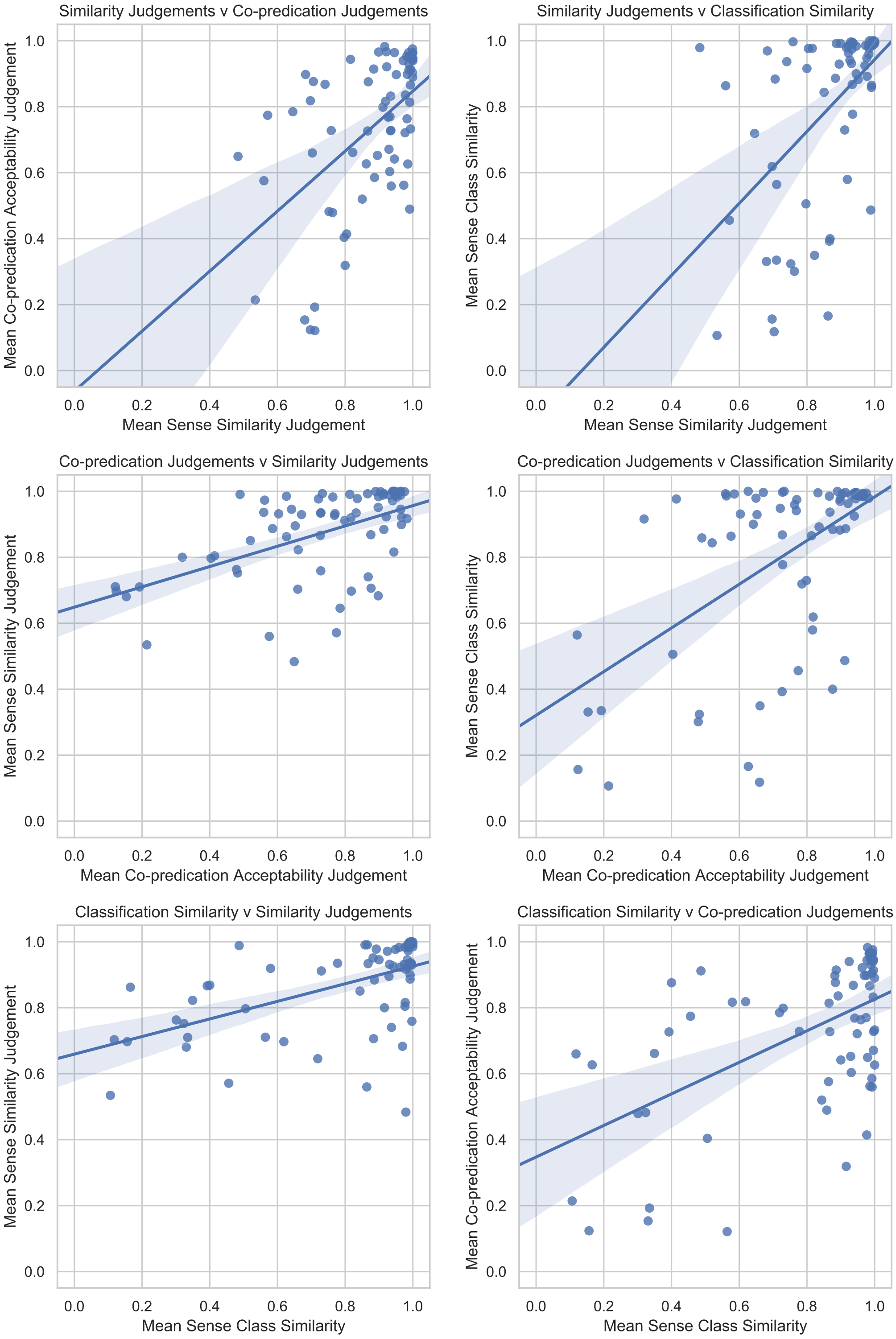


Figure 4.16: Correlations between polysemic target word pairs based on the three collected judgements of word sense similarity, together with their best linear fit.

r, ii) perform an ordinary least squares (OLS) regression, and iii) calculate the mean squared error (MSE) of OLS predictions under five-fold cross validation. The results of these calculations are displayed in the top part of Table 4.4, and visualised in Figure 4.16. Overall, we find a moderate but significant correlation between the three human annotation metrics. Similarity judgements and co-predication acceptability judgements show the lowest correlation in the set (Pearson's r of 0.529), while acceptability judgements and categorical class similarity achieve the highest correlation (Pearson's r of 0.563). These results indicate that categorical class boundaries between referent interpretations might have a more direct influence on whether two different senses can felicitously be co-predicated than their graded similarity score. The correlation graphs in Figure 4.16 again display the coverage of judgements obtained for the three human annotation metrics, indicating that class similarity ratings - like co-predication acceptability, span over the full scale - while similarity judgements only cover the top half. Here however this means that predicting acceptability scores from similarity ratings is more difficult than the inverse, leading to a higher error rate in the prediction of low-similarity items, and an overall higher mean squared error (MSE; 0.014 to 0.04). The same holds for predicting similarity class labels from similarity judgements, which is more difficult than predicting similarity judgements based on class similarity.

Overall, the three tested measures of word sense similarity show a similar pattern: as expected, same-sense sample pairs usually received almost perfect explicit sense similarity, co-predication acceptability and word sense class overlap scores. Cross-sense samples in some cases can receive ratings that are comparable to the same-sense ones, but in some cases also can exhibit significantly lower scores - even when compared to each other. Co-predication acceptability and word sense class overlap exhibit the largest variance in scores assigned to cross-sense samples - and as a result show the highest correlation with each other. Explicit word sense similarity judgements often exhibit a similar pattern as the other two scores, but annotated scores vary less, and seem to be overall higher than for the other two ratings.

When comparing explicit similarity scores with co-predication acceptability ratings, judgements seem to better align towards the upper end of the rating scale. This indicates that while these two types of judgements seem to assign comparable scores to more similar sense combinations, ratings for low similarity samples can differ significantly - which in turn hints at a potential distinction in the sensitivity to lower similarity cross-sense samples. To further investigate this observation, we inspected the annotations of polyseme *newspaper*, which exhibits both low-scoring and high-scoring cross-sense samples. As mentioned before, in our experiments we

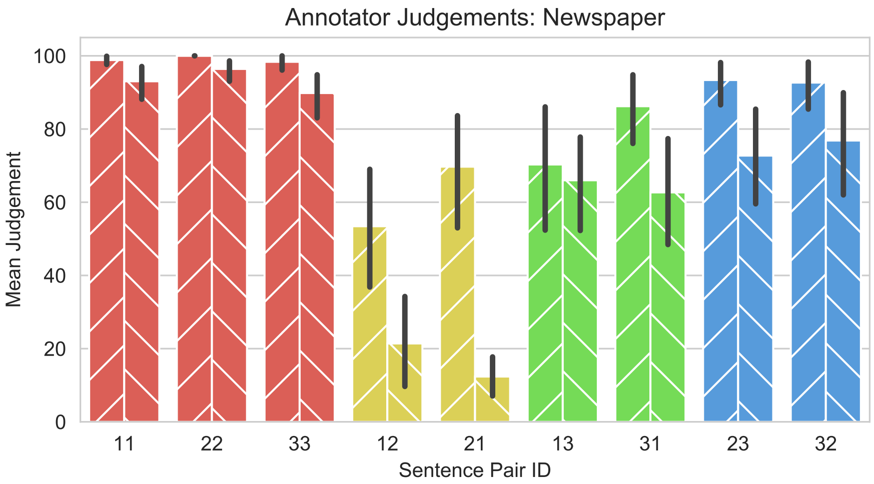


Figure 4.17: Mean similarity ratings (left, ascending hatch) and co-predication acceptability ratings (right, descending hatch) for the nine sense interpretation pairs of polyseme *newspaper*. The first three bars represent same-sense pairs, the other three groups the different combinations of cross-sense readings, respectively.

assume that *newspaper* has three distinct but related sense interpretations: (1) *organisation/institution*, (2) *physical object*, and (3) *information/data*. Figure 4.17 shows the mean similarity and acceptability ratings for the nine combinations of sense interpretations: The first three bars represent same-sense pairs 11, 22 and 33, the other three groups the different combinations of cross-sense pairs. The figure reveals that the three same-sense pairs receive equally high similarity and acceptability ratings, but while similarity ratings for cross-sense pairs decline in a continuous fashion down to 0.53 for sentence pair 12 combining the *organisation* and *physical* readings, acceptability ratings roughly decrease in two steps, separating similarity and acceptability scores more strongly for lower-rated samples. These results indicate that explicit similarity ratings might be more finely graded than co-predication acceptability, which appears to assign significantly lower scores to readings perceived to be infelicitous. As this significant drop in acceptability is also reflected by the class overlap scores, co-predication acceptability might be especially sensitive to class overlap, and assigned low-similarity scores largely reflect class mismatches, while explicit sense similarity judgements are less indicative of this aspect.

4.5.5 Computational Predictions

The bottom part of Table 4.4 displays the results of predicting human judgements of polyseme sense similarity based on the different contextualised encodings produced by ELMo, BERT and the Word2Vec baseline. Overall, only six of the pairwise

	Newsp.	**H.way**	**W&P**	**Lunch**	**Door**	**DVD**	**School**	**Wine**	**Glass**	**Constr.**
BERT WE	0.383	0.692	**0.235**	**0.899**	0.079	0.409	0.259	0.459	-0.739	**0.623**
BERT SE	0.591	**0.999***	-0.159	0.316	**0.449**	0.355	0.092	0.458	-0.973*	-0.115
BERT CLS	0.317	0.960*	0.017	0.152	-0.202	**0.517**	0.084	0.216	-0.933	-0.492
ELMo WE	**0.919***	0.916	-0.310	-0.278	0.018	-0.167	**0.332**	0.442	-0.666	0.648
Word2Vec SE	0.576	0.126	0.089	-0.923	0.177	0.361	-0.310	**0.795**	-0.614	0.117

Table 4.5: Correlations between human sense similarity judgements and the similarities in the representations derived from different contextualised word embedding techniques as measured with Pearson's r. Highest correlating model output in bold font, significant correlations ($p<0.05$) starred.

comparisons between the similarity scores calculated with a contextualised language model and the mean judgements provided by our annotators reached significance - which we suspect is mainly due to the small number of only 67 test items annotated in the pilot.[16] The highest correlation any of the tested models achieves in comparison with one of the human annotations is BERT Base's target word embedding similarity compared to co-predication acceptability ratings, which reaches a moderate correlation with Pearson's r of 0.48, and thus only a few points below the correlation between explicit similarity ratings and co-predication acceptability judgements (Pearson's r of 0.53). The best predictor of explicit sense similarity ratings here seems to be ELMo with a correlation of 0.26, strongly outperforming the Word2Vec baseline at 0.05, and ELMo's embeddings also display the best match with class-label overlap scores at a Pearson's r of 0.32. The Word2Vec baseline performs on par or better than the remaining contextualised approaches in predicting those two measures, with both correlations close to 0.25. Based on these preliminary results, it appears that the actual target word embeddings overall perform best in predicting the human annotations of word sense similarity, and that especially BERT seems to perform relatively well in predicting co-predication acceptability specifically.

One of the central findings in the analysis of the collected human annotations were the - sometimes significant - differences in similarity or acceptability scores assigned to different cross-sense pairings of our polysemous test items. Considering a potential application of contextualised language models to proxy or replace human judgements in future work, we next investigated whether the computational models' predicted similarity scores replicated specifically these findings. Table 4.5 displays the per-target correlations between the collected explicit sense similarity ratings and

[16] See Table 5.2 for the comparison of computational approximations and human judgements on the extended, complete data. Here indeed all comparisons do clearly reach significance.

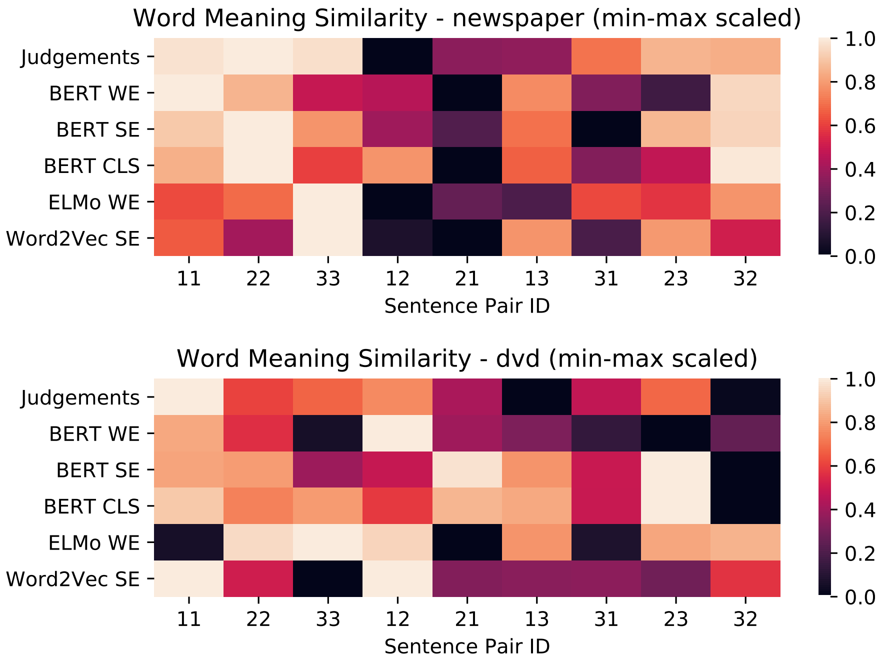


Figure 4.18: Comparison of word sense similarity ratings based on annotator judgements and ELMo and BERT context-sensitive word embeddings for targets *newspaper* and *DVD*, min-max normalised to amplify the visibility of effects. Brighter indicates higher similarity.

the cosine similarities of the target expressions (or sentences) given these different contextualised embedding techniques. With only a fraction of the correlations reaching significance,[17] none of the embedding techniques appears to consistently capture the similarity patters observed in the human judgements. With the exception of *door* and *glass* - which generates a negative correlation for all computational approaches - BERT Base however seems to produce the best predictions of pairwise similarities.

Qualitative Analysis

Moving to a more qualitative analysis of the contextualised language models' embeddings, we created heat maps to display the similarity patterns for the different polysemic expressions tested. The resulting heat map for *newspaper* is shown in Figure 4.18, displaying on a more accessible level the difference in similarity scores assigned by human annotators and the different contextualised approaches.[18] Scores are min-max normalised in these Figures, with darker colours indicating lower similarity scores, and brighter colours indicating higher similarity. In the case of *newspaper*, some of the contextualised similarity scores seem to reflect the human judgements -

[17] Note that the compared similarity vectors are of length 4-16 only

[18] The heat maps for the full set of tested polysemes can be found in Figure A.4

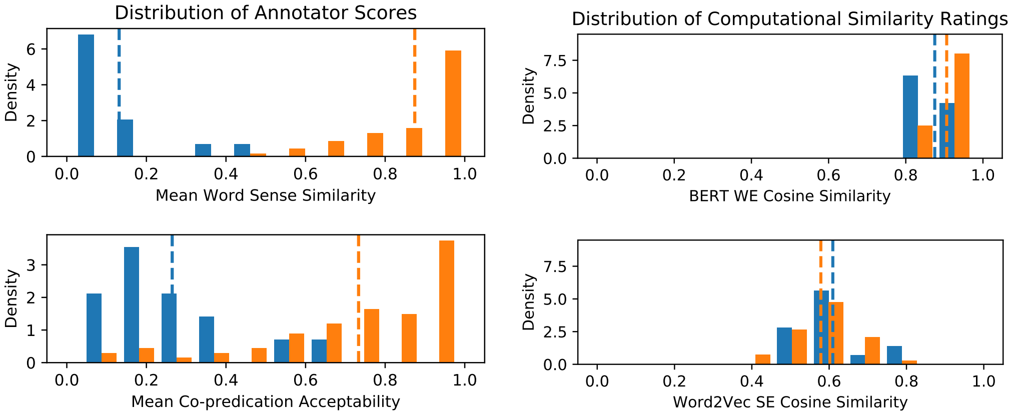


Figure 4.19: Distribution of human annotation ratings and computational similarity ratings for homonymic (blue) and polysemic (orange) sentence pairs, together with their means.

and especially so for sense interpretations rated to be highly similar (e.g. 11, 22 and 32) or dissimilar (12, 21) - but overall the differences in embeddings do not appear to consistently resemble the human judgements, as exemplified by the heat map for *DVD*.

The min-max scaling in the heat maps was necessary to better visualise the similarity scores produced by the computational approaches because the overall embedding similarity of different samples was significantly higher here than in the human annotations. With the exception of some sample combinations for *glass* and *school*, all of BERT's similarity scores were above 0.9 - both for same-sense as well as for cross-sense samples, and all of ELMo's similarity scores were well above 0.8. This finding reflects an observation made earlier by Ethayarajh (2019), suggesting that the embeddings of contextualised language models only occupy a relatively small cone in the embedding vector space, where in some cases even random sample pairings can achieve close to perfect similarity scores. We will provide a more detailed investigation of this phenomenon in Chapter 5.3.3 based on an extended sample set including homonymic alternations to contrast the similarity scores for polysemic targets.

Predicting Ambiguity Types

By means of a sanity check, we were interested to see whether the different measures of word sense similarity we had collected so far - both human and computational - were able to correctly classify polysemic and homonymic alternations. The two left-hand graphs in Figure 4.19 show the overall distribution of human annotation ratings for the similarity and co-predication acceptability of homonymic (blue) and polysemic (orange) sentence pairs, together with their means. Both annotation metrics clearly separate the modes of the distributions, but while co-predication acceptability judgements for the tested polyseme pairs occupy the entire rating scale, explicit word sense similarity ratings only span the upper half (lowest score = 0.48). Conversely, co-predication ratings for homonym pairs reach up to 0.67, while the highest-scoring homonym pair only reaches a similarity score of 0.44. This impacts the distribution means, which are closer to each other in the acceptability ratings than in the similarity scores. As mentioned earlier, the computational approaches to rating word sense similarities overall return relatively high scores for all combination of samples, usually only utilising the top 20% of the scale. As a result, the means of their distributions are significantly closer, as exemplified by the distributions of BERT word embedding similarity ratings for polyseme and homonym pairs in the top right graph of Figure 4.19. The primitive Word2Vec sentence embeddings lastly even assign a higher mean similarity score to homonym pairs than to polysemes (last graph).

Because co-predication acceptability judgements show a higher overlap between the distributions of homonym and polyseme ratings than the similarity ratings, we expected explicit similarity ratings to be a stronger predictor in classifying target pairs as either homonyms or polysemes. To validate this intuition, we classified items through a support vector machine (SVM) with linear kernel under five-fold cross-validation (Cortes and Vapnik, 1995). An SVM is a basic but robust supervised learning method that uses labelled training samples to derive a multi-dimensional decision boundary aimed to separate the different classes encountered in the training data as clearly as possible. In order to do so, it attempts to maximise the distance between the decision boundary and the data points closest to it.

As our dataset is skewed towards polysemy samples, baseline performance is an accuracy of 0.825, achieved by assigning all samples to the polysemy class. Sample classification based on both similarity ratings and co-predication ratings outperform this baseline, with an accuracy of 0.988 for explicit similarity ratings, and 0.895 for co-predication acceptability ratings, respectively. Figure 4.20 shows the optimal decision boundary between homonym samples (blue) and polyseme pairs (orange) calculated for the two human annotation metrics. These figures indicate that the higher overlap in homonym and polyseme ratings indeed prevents a clear delineation between the two ambiguity types when assessed based on their co-predication acceptability. None of the computational metrics manages to outperform the baseline, with each of them consistently applying max-class labels based on the SVM's decision boundary. Neither combining the two human annotated metrics, nor combining any

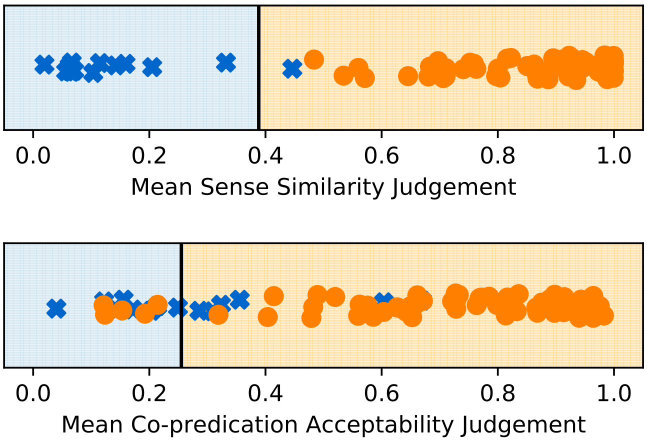


Figure 4.20: Classification of homonym (blue) and polyseme (orange) sample pairs based on pairwise similarity ratings and co-predication acceptability judgements.

of the computational metrics improves their respective classification performance over the best individual score.

4.6 Discussion

In this chapter we presented an annotation pilot for collecting human judgements of the similarity of polysemic sense extensions, an investigation of the data collected through the pilot, and a preliminary analysis of a number of contextualised language models with respect to the annotated data.

We collected three different measures of human sense similarity judgements: explicit sense similarity judgements, co-predication judgements and discrete word class overlap scores. Annotations were collected through online crowd-sourcing, with layperson annotators rating custom samples invoking different sense extensions of a set of ten seminal polysemes. All three metrics show that in some cases polysemic cross-sense samples can be perceived as significantly less similar than their same-sense counterparts - and in some cases even significantly less similar than other cross-sense combinations. Explicit similarity judgements seem to be the most graded measure of the three tested types of annotations, with co-predication acceptability and word sense class overlap showing signs of gradedness especially in the upper end of the rating scale but displaying stark drops in ratings when sentences are considered infelicitous or class overlap is considered minimal. Co-predication acceptability and word sense class overlap also exhibit the strongest correlation with each other, indicating that co-predication acceptability might at least partially depend on the target expression's sense class overlap.

In order to investigate how well contextualised language models could replicate the human judgements, we compared similarity scores based on the cosine between contextualised target embeddings produced by ELMo, BERT and a Word2Vec baseline. None of the tested approaches seemed to consistently capture the human ratings of any annotation type, but BERT Base produced a moderate correlation with acceptability judgements and seemed to perform quite well for a selection of targets when inspected more qualitatively.

Based on the results obtained from the data collection pilot, we concluded that all three human measures were providing interesting data for an empirical, data-driven investigation of polysemy, showing perceived word sense distances as a gradual phenomenon. Moving forward, we however decided to discontinue the collection of class labels due to their comparably coarse sensitivity and high correlation with the co-predication acceptability judgements.

With respect to our primary research question, the differences in the similarity labels assigned to some polysemic cross-sense samples provided us with some initial evidence that word sense distance might play a role in the mental representation of polysemes - challenging traditional one representation models. Equally, we found that some cross-sense pairs were rated at similarity levels identical to those of same-sense pairs, which is difficult to consolidate with a sense enumeration approach to the mental lexicon. But since the validation data from the pilot run only covered a single target word for ten different classes of regular metonymic polysemes, it didn't yet allow for any remakrs on the ubiquity of this phenomenon, nor on its regularity. A straightforward and crucial next step therefore was the expansion of the dataset with alternative targets and additional types of alternations. More data was equally thought to improve the soundness of our analysis of contextualised language models where similarity scores matched the human annotations only occasionally, and would allow us to conduct a more rigorous analysis of inter-annotator agreements.

Moving forward, we decided to discontinue the collection of class labels due to their comparably coarse sensitivity and high correlation with the co-predication acceptability judgements, and dropped the analysis of BERT sentence embeddings and the special classification token [CLS] due to their noisy signals. Instead, we opted to include BERT Large to the repertoire of tested models, as these larger models had been found to markedly improve downstream task performance. As Chapter 5.3.3 will show, BERT Large ultimately outperformed any of the approaches tested in the pilot, and proved a promising tool in investigating phenomena of polysemy on a larger scale - leading to its use in the detection heuristics presented in Chapter 6.

Chapter 5

Similarity Patterns in Regular Polysemy

The data collection pilot presented in the previous chapter indicated that both explicit ratings of word sense similarity and co-predication acceptability judgements were sensitive enough to capture subtle differences in the interpretation of polysemous expressions. Both measures produced overall lower ratings for polysemous cross-sense samples than for their same-sense counterparts, and in some cases the differences in assigned scores reached significant levels.

Because an in-depth analysis was still limited by the small amount of data collected so far - especially so when investigating the performance of contextualised language models - in this chapter we present a second annotation effort aimed at augmenting and extending the pilot data. For this second annotation run we focused on collecting annotations of additional samples allowing for the same set of sense alternations as the target expressions in the first run, i.e. additional words exhibiting the same type of polysemic alternations as those in the pilot. The resulting dataset contains similarity and co-predication acceptability judgements for a total of 28 seminal and experimental polysemous targets representing ten different types of regular, metonymic polysemy, and contains a total of close to 18,000 annotations.

The extended dataset then allows us to carry out a set of rigorous analyses, including i) an investigation of similarity patterns within and across polysemy types, ii) performing a more detailed investigation of the correlation between human judgements and sense similarity scores calculated by contextualised language models, and iii) attempting a clustering of word sense interpretations based on their contextualised embeddings.

This chapter follows roughly the same setup as the previous one, introducing first the extended set of materials and changes in methodology (Sections 5.1 and 5.2), followed by an in-depth analysis of the full set of annotated data with a focus on word sense similarity patterns. We then revisit the investigation of contextualised language models in Section 5.3.3, now obtaining much better and more stable performance scores indicating that especially BERT Large might capture word sense sufficiently well to prove useful as a proxy of human annotations in future work. We conclude this chapter with an investigation of similarity patterns in regular metonymic polysemes (Section B.2). Parts of this chapter have previously been published in Haber and Poesio (2021).

5.1 Materials

The materials used for our second annotation run added a range of new, seminal and experimental target words, as well as a few additional sense extensions. We otherwise adhered to the same procedure of creating custom sample sentences based on the template presented in Chapter 4.3.2, but experimented with the inclusion of control items to support the filtering of noisy annotations during analysis.

5.1.1 Target Words

From among the target expressions of the pilot run, only two target expressions resulted in multi-word or multi-token embeddings averaged to create a target encoding: *Hemingway* and *War and Peace*. These two targets also were the only proper nouns included in the study, and especially *War and Peace* seemed to lead to unstable contextualised embeddings - potentially due to the averaging over contrasting concepts *war* and *peace*, and the encoding of syntactical information related to *and*. In order to produce even clearer results, we therefore decided to exclude any proper nouns or multi-word expressions in the second annotation effort. We then selected 18 additional targets, each allowing for the same alternations as one of the initial single-word targets in the pilot run in order to investigate potential patterns in their distribution of sense interpretations. Targets were chosen to be either highly representative of the respective type of alternation (seminal targets), or to represent a fringe case of the phenomenon (experimental targets). By including experimental targets along seminal ones we aimed at getting an even better indication of the range and consistency of any potential patterns in the interpretation of polysemous word forms.

We collected annotations for sample pairs and co-predication structures containing the 18 new targets, re-collected annotations for targets *chicken* and *glass* due to the high level of noise observed in the data for these targets obtained from the first annotation run, and included the original data of the other eight initial targets. The modified and expanded dataset now contains explicit similarity ratings and copredication acceptability judgements for different interpretations of the following set of logical metonymic, polysemic targets:

1. **animal/meat**: lamb, chicken, pheasant, seagull;
2. **food/event**: lunch, dinner;
3. **container-for-content**: glass, bottle, cup;
4. **content-for-container**: beer, wine, milk, juice;
5. **opening/physical**: window, door;
6. **process/result**: building, construction, settlement;
7. **physical/information**: book, record;
8. **physical/information/organisation**: newspaper, magazine;
9. **physical/information/medium**: CD, DVD;
10. **building/pupils/directorate/institution**: school, university

5.1.2 Sample Contexts

Since the custom template developed to create sample contexts appeared to be suitable for eliciting both explicit word sense similarity judgements as well as copredication acceptability judgements, we applied exactly the same approach to create most of the materials for the second data collection effort. Besides polysemic alternations, some of the new targets however also allowed for homonymic alternations (e.g. *magazine*, which like *newspaper* allows for different sense interpretations related to the print medium, but also allows for at least one other, homonymic interpretation as a storage type). Instead of adding specific homonymic targets, this time we included these homonymic interpretations of our polysemic targets to the materials to allow for an even better perspective on the results obtained.

Control Items. For the second annotation run we abandoned including synonym or filler items in order to optimise the proportion of usable test data, and opted for including two control items in each questionnaire to still allow for filtering spurious submissions. In the explicit word sense rating setting, one of the control items would present virtually the same sentence twice, with one of the sentences changing a minor detail irrelevant to the interpretation of the target:

(45) 1. The **bat** flew in through the open window.
2. The **bat** flew in through the door.

The second control item contains two completely unrelated sentences of the same format as the test items:

(46) 1. The **match** ended without a clear winner.

2. The **bass** managed to get off the hook.

In the co-predication acceptability setting, one test item displayed a sentence of roughly the same length as the test items which also would contain a conjunction - but both phrases would introduce and refer to different subjects:

(47) A group of boys were playing Frisbee in the park and a girl tried to balance on a slack line.

The other control sample would start off as a regular test sample, but end in a random permutation of words:

(48) The match ended without a clear winner and the off the managed bass hook get to.

We expected very high ratings for each first control item and very low ratings for each second, and intended to use the scores assigned to these items as a filtering criterion.

Context⊕ and Context⊖ Samples. Inspecting the data collected through the annotation pilot, we observed that even after spending considerable effort in creating highly controlled, clear context sentences, in some cases the predication itself still had a significant effect on the similarity and acceptability ratings given to an interpretation pair (see Chapter 4.5.1). After we established that this effect was not due to predicate ordering in the item (see ibid), we decided to include some additional, experimental items to our second annotation run in order to create more data allowing us to specifically investigate predication effects. Besides the original, neutral context sentences used to invoke one certain interpretation of an ambiguous target as clearly as possible, in the second annotation run we thus also included some sample contexts that - while clearly invoking one particular sense - were designed to either support or impede with a coercion of the target's interpretation into an alternative reading. Using a concrete example, in the *animal/meat* alternation for target *chicken*, we added a `context⊕` sample intended to facilitate a *meat* interpretation while invoking the *animal* reading by mentioning the breeding of the animal (for meat production) rather than a depicting the target as an animate entity:

(49) The chicken was bred by a well-known family of poultry breeders.

Conversely, we added some `context`⊖ samples that were aimed at impeding sense shifting, usually by focusing on aspects that clearly distinguish different sense interpretations. As an example, for polysemic target *bottle* allowing for *content/container* alternations, we added a context sample specifically focusing on the physical attributes of the container:

(50) The bottle was made out of recycled glass fished from the ocean.

We added a total of 13 of these experimental sample contexts to the materials, and this time directly included both possible orderings of any sample pair in our questionnaires. In total, we created 20 questionnaires with 18 test items and 2 control items each. The full list of samples can be found in Appendix B.1.

5.2 Method

In our second data annotation run we collected crowd-sourced judgements of explicit word sense similarity and co-predication acceptability, and computational predictions of word sense similarity based on Word2Vec, ELMo, BERT Base and BERT Large.

5.2.1 Human Annotations

To collect word sense similarity judgements we again asked participants on Amazon Mechanical Turk to rate the similarity in meaning of a target word shown in two different contexts. We followed the pilot method in highlighting target expressions in bold font and asked annotators to rate the highlighted expressions using a slider labelled with 'The highlighted words have a completely different meaning' on the left hand side and 'The highlighted words have completely the same meaning' on the right. Co-predication acceptability ratings were obtained for the same samples combined into a single co-predication structure through conjunction reduction. The slider shown to participants here again was labelled with 'The sentence is absolutely unacceptable' on the left and 'The sentence is absolutely acceptable' on the right.

We slightly updated the instructions given to participants, now mentioning that there were control items in each questionnaire that would allow us to detect and withhold the reward of spurious annotators. The instructions for the explicit similarity rating task for example now read as follows:

(51) Carefully read each pair of sentences and specify how similar the highlighted words are by using the slider. The slider ranges from 'The highlighted words

> have a completely different meaning' on the far left to 'The highlighted words have completely the same meaning' on the far right.
>
> There are 20 sentence pairs.
>
> The survey contains a number of test items that can be used to determine whether you are carefully reading the sentences or are submitting random answers. Submissions that fail the test items will be rejected.

Annotators this time were paid 0.70 USD for a completed questionnaire with 20 items, for an average expected hourly rate of 7.00 USD.[1] To improve annotation quality, we this time required annotators to be located in the US (replacing the previous high school graduate criterion to select for English native speakers), and have completed at least 5000 previous surveys with an acceptance rate of at least 90% (replacing the relatively expensive AMT Master criterion).

5.2.2 Contextualised Language Models

In order to assess word sense similarity encoded in contextualised embeddings, we again extracted target word embeddings from the different disambiguating contexts and calculated their cosine similarity (1-cosine). For ELMo we used the pretrained model on TensorFlow Hub[2] and extracted target word vectors from the LSTM's second layer hidden state. We used the pretrained BERT Base (12 layers, hidden state size of 768) and BERT Large (24 layers, hidden state size of 1024) from the Huggingface transformers package.[3] As suggested by Devlin et al. (2019) and Loureiro and Jorge (2019), we this time experimented with both the last hidden state and the sum of the last four hidden states as contextualised representation of a target word. Finding that summing over the last four layers significantly improved correlation with the human annotations, we settled for the latter in most parts of the analysis. Lastly, we again created a baseline score by averaging over the static Word2Vec encodings of all words in a sample context to create a naive contextualised embedding.

5.3 Analysis

In our analyses of the collected data we focused on three aspects: First, we again computed graded similarity and acceptability ratings for different polysemic alternations based on the collected annotations, specifically investigating the notions of

[1] Annotators were paid irrespective of their ratings of control items.

[2] `https://tfhub.dev/google/ELMo/3`

[3] `https://huggingface.co/transformers/pretrained_models.html`

word sense distance and similarity patterns. We then analysed how well the different contextualised language models' target embeddings correlated with either of the human annotations measures, and to what degree they replicated the patterns of word sense similarity observed in the human annotations. Lastly, we analysed the contextualised embbedings themselves for a preliminary assessment of how well these 'off-the-shelf' word sense encodings fare in clustering samples based on their sense interpretation.

5.3.1 Word Sense Similarity Judgements

We collected an additional 8,980 explicit sense similarity judgements through 449 surveys completed by a total of 220 unique AMT participants rating the similarity of highlighted target words in different contexts. In order to reduce annotation noise, we then filtered out submissions from participants who failed to rate the two control items included in each survey according to a set of custom criteria.

As mentioned before, one control item contains a (potentially ambiguous) target word interpreted in exactly the same way in both sentences, with only minimal, insignificant changes to the context (control-same):

(52) 1. The **mole** dug tunnels all throughout the garden.
2. The **mole** dug tunnels under the flower bed.

The second control item contains two sentences with completely unrelated targets (control-random):

(53) 1. The **model** wore a new dress designed by Versace.
2. The **seal** indicated that the letter had never been opened.

Submissions were excluded from analysis if either the *control-same* item was rated below 0.7 similarity, or the *control-random* item was rated above 0.2 similarity. After filtering, we retained a total of 5,862 judgements, including those obtained in the initial data collection, with an average of 16.5 annotations per item (minimum 7) and a per-questionnaire inter-annotator agreement rate of 0.62 (Krippendorff's alpha, Artstein and Poesio, 2008) - a decent rate considering the fine-grained continuous scale provided to our annotators.

Given that we this time directly included both possible orderings of sample sentences in our questionnaires, we first investigated whether sample ordering had any effect on a test item's similarity ratings. To do so, we applied a Mann-Whitney U test (Mann and Whitney, 1947) comparing the ratings assigned to one ordering of a test item with those given to the other. Confirming our preliminary observation that

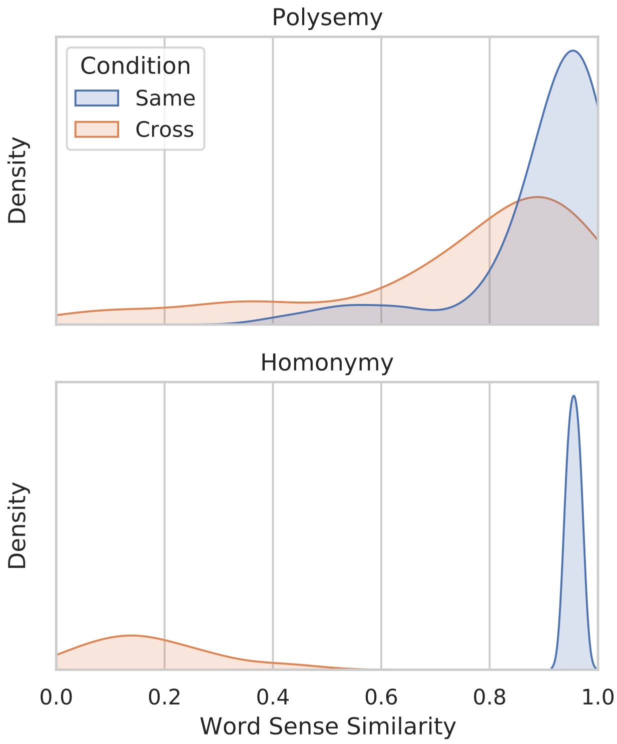


Figure 5.1: Normalised distributions of explicit word sense similarity ratings given to same-sense (blue) and cross-sense (orange) samples with polysemic and homonymic alternations.

ordering effects were non-significant (see Section 4.5.1), only 22 of 229 pairwise tests yielded p-values < 0.05, and none passed Bonferroni correction, i.e. no comparison was deemed significantly different after adjusting the p-value threshold with respect to the number of tests conducted. We therefore concluded that - as expected - order effects are negligible for explicit word sense similarity ratings, and combined results for further analysis.

Figure 5.1 shows the overall distributions of word sense similarity ratings collected across all target words, separated on whether or not there is a sense alternation in the sample (same- vs cross-sense), and whether this alternation is traditionally considered to be polysemic or homonymic in nature. Homonymic cross-sense samples obtained a mean similarity rating of just 0.17, significantly lower than the overall same-sense mean of 0.89 (p-value < 0.05). Polysemic cross-sense samples received a mean similarity score of 0.73, which is significantly lower than the same-sense mean, but significantly higher than the homonymy cross-sense mean (see Table 5.1, row 1). These results support the traditional view that polysemy occupies a distinctive middle ground between identity of meaning and homonymy (see e.g. Pinkal, 1995;

Poesio, 2020), and do not replicate the findings of Trott and Bergen's analysis of the annotations collected for the RAW-C dataset. While our data was collected in a very similar way to theirs, we suggest that the observed difference in the distribution of ratings stems from the presentation of the ambiguous targets in the experiment materials. While in our samples a target expression always was followed by a verb phrase disambiguating its interpretation, RAW-C oftentimes presents targets within compound noun phrases, like *traffic cone* vs *ice cream cone*, or *fruit bat* vs *baseball bat.* We argue that this is especially problematic for polysemic targets, as in this case the expression will no longer allow for an under-specified interpretation. As a result, we would classify their compound polyseme samples as h-type ambiguous following Pinkal's definition, placing them closer to homonymy than polysemy proper.

Next, we grouped the collected similarity data based on target words, and performed pairwise comparisons on all ratings given to their cross-sense interpretations. A large number of significant comparisons would indicate a high variance in the assigned ratings; a low percentage of significant differences indicates a consistent rating of samples. Due to the large number of tests, we then again carried out a Bonferroni correction on the obtained results to establish a corrected, more conservative significance level and determine an upper bound on this statistic. Comparing all combinations of same-sense pairings for example, 20 of 58 tests yielded significantly different results (p-values < 0.05), but only 4 entries passed Bonferroni correction (6.90%), indicating that same-sense samples were quite consistently rated to invoke very similar interpretations.

All 52 pairwise comparisons between homonymic cross-sense and same-sense ratings passed Bonferroni correction, meaning that all homonymic cross-sense samples were rated significantly lower than any cross-sense sample. 14.71% of the 34 pairwise comparisons among homonymic cross-sense samples passed Bonferroni correction, as did 23.44% of the 337 pairwise comparisons between ratings for polysemic cross-sense samples. Ratings for cross-sense samples therefore are less consistent than same-sense ratings, and polysemic alternations are rated more inconsistently than homonymic ones. Observing this variance in similarity scores again proves the importance of offering annotators a graded rating scale instead of a binary classification task during annotation. And with almost a quarter of the similarity ratings for polysemic sense alternations showing significant differences to those of other sense pairings, these results also contribute to the accumulating empirical evidence that is difficult to explain when assuming a uniform treatment of polysemic senses in the mental lexicon.

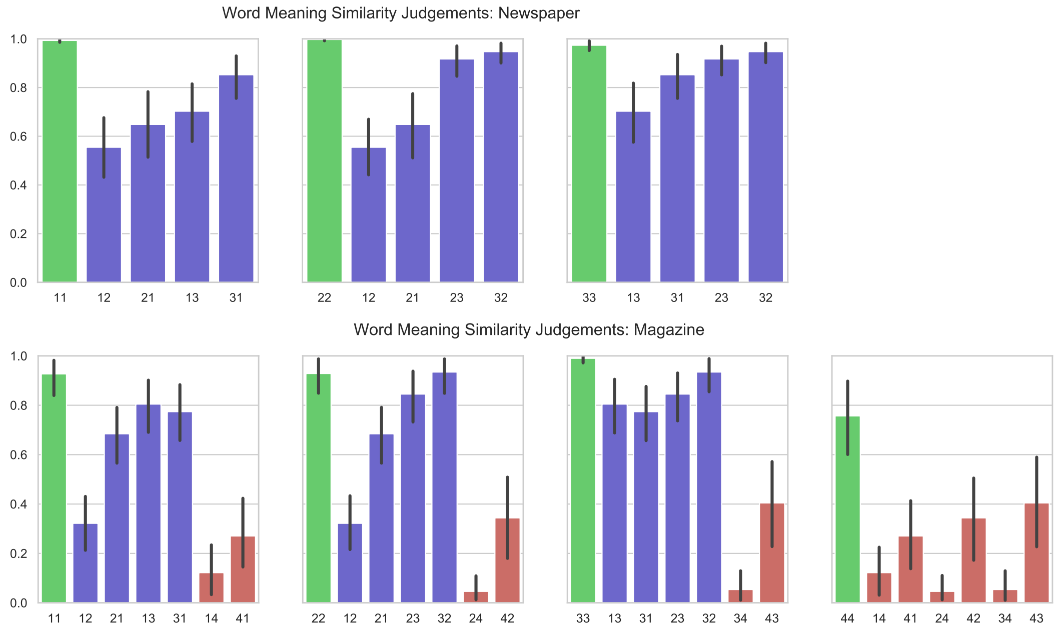


Figure 5.2: Mean word sense similarity ratings given to same-sense (green) and cross-sense (blue) sentence pairings eliciting different interpretations of targets *newspaper* and *magazine*. Additional homonymic comparisons are shown in red. Senses: 1-organisation, 2-physical, 3-content, 4-storage type.

Qualitative Analysis

Figure 5.2 shows the mean similarity ratings for sample context pairs eliciting different combinations of interpretations of the ambiguous targets *newspaper* and *magazine*. Like in all of the following figures of this kind, ratings for sample pairs eliciting the same sense are shown first, in green colour, ratings for polysemic cross-sense samples are shown next, in blue, and additional homonymic cross-sense pairings are shown last, in red colour. The indicators on the x-axis show the combination of sample contexts for the target as indicated by their context identifiers. In this case, context 1 elicits the *organisation* reading, sense 2 a *physical* reading, sense 3 a *content* reading, and sense 4 a homonymic interpretations as a type of storage. As mentioned earlier, results for both possible sample orderings within each test item were combined because we found no significant order effects. Keeping this in mind, the ordering of context identifiers depicted in the figure does not relate to the ordering of the sample contexts in the test item, but rather to which of the two context samples created for each sense was used in the comparison. As an example, the label 12 indicates that the sample uses the first context sentence for the first sense interpretation of the target (in this case an *organisation* reading, and the second sample sentence for the second reading (in this case *physical*). Label 32 indicates that the sample uses the first sentence for sense three (*information*) and the second sentence for sense two (*physical*), etc. The y-axis displays mean similarity on a scale from 0 (complete unrelatedness) to 1 (identity of sense). Ratings are grouped into different sub-plots by sense, with some being repeated for clarity (i.e. the rating for pairing `12` in the first sub-plot is identical to the `12` rating displayed in the second).

Newspaper and Magazine. A first observation that can be made throughout the similarity ratings for different targets - and here for *newspaper* and *magazine* - is that same-sense ratings usually are rated to be highly similar, often close to perfect sense identity. This is reflected by the previously mentioned overall same-sense mean of 0.89, and only 7% of significantly different comparisons among same-sense pairings. The same-sense rating for magazine's homonymic fourth interpretation as a type of storage coincidentally is one of the samples contributing to this 7% share. The context sentences used in this comparison were

(54) 4a The **magazine** contained all kinds of defunct WW2 weaponry.
4b The **magazine** was originally designed for storing ballistic missiles.

and at 0.76 were rated to elicit significantly less similar interpretations of the ambiguous target than any of magazine's polysemic same-sense pairs. This is surprising as same-sense ratings in principle should not be affected by type of ambiguity - and in this case even less so, as *magazine* has both, polysemic and homonymic interpretations. We see two potential explanations for the low same-sense similarity rating of the *storage* same-sense contexts: Firstly, *magazine*'s print media related interpretations appear to be more salient than its storage type meaning, which could cause annotators to struggle in deriving the storage reading for at least one of the context sentences and consequently failing to properly assess the similarity of the depicted sample contexts. A second reason could be vagueness or ambiguity in the materials, with one of the two context samples invoking its intended interpretation not clearly enough to result in a perceived identity of sense when comparing the samples. Considering the very low similarity ratings for comparisons of sample sentence 4b with any of the print-medium readings (red bars `14`, `24` and `34` in *magazine*'s similarity ratings in Figure 5.2), and the significantly higher ratings for comparisons with 4a (`41`, `42` and `43`) the culprit here might be sentence 4a, which potentially could be coerced into an *information* reading as a magazine containing articles or information about all kinds of defunct WW2 weaponry. This again highlights the importance of carefully crafted context samples for an analysis as detailed as ours, and shows that even after a considerable amount of effort has been put into making contexts as clear as possible, readers still might derive other interpretations than intended.

Returning to the set of polysemous interpretations relating to the print medium depicted in blue, it is noteworthy that the spread of pairwise similarity ratings is large, ranging from 0.56 to 0.95 for *newspaper*, and even 0.32 to 0.93 for *magazine*. For *magazine*, the low end of this spectrum falls in the range of similarity ratings obtained for cross-sense pairings with unrelated, homonymic interpretations, and the top end for both targets reaches levels similarity scores close to those of same-sense comparisons. This means that for targets *newspaper* and *magazine*, at least some of the different polysemic interpretations related to the print medium are perceived to be significantly dissimilar to one another. Specifically, the second bar chart with similarity ratings for *newspaper* in Figure 5.2 reveals that the physical reading (sense 2) appears to be more similar to the content reading (sense 3, see comparisons `23` and `32`) than to the organisation reading (sense 1, see comparisons `12` and `21`). These differences between polysemic sense extensions can be seen as tentative support for the existence of some form of grouping of polysemic senses in the mental lexicon.

A third observation relates to the inconsistency of similarity ratings assigned to items that contain the same combination of senses invoked by different context sentences. Take for example items `12` and `21` for target *magazine*, which are the following comparisons:

(55) 1a The **magazine** lost a court battle against a former pop star.
2b The **magazine** was covered in paw prints after a cat sat on it.

(56) 1b The **magazine** got into serious money problems last year.
2a The **magazine** just kept falling off the small living room table.

while contexts 1a and 1b, and 2a and 2b are rated to be highly similar in meaning (both at 0.93-see comparisons `11` and `22` in the *magazine* ratings of Figure 5.2), the comparison between 1a and 2b (item `12`) receives significantly lower ratings (mean of 0.32) than the comparison between 1b and 2a (item `21`, mean of 0.68). In contrast to the the same-sense homonym caparison discussed above, here we struggle to identify a clear cause for the divergent annotation scores.

Animal/Meat: Lamb, Chicken, Pheasant and Seagull. The full data set contains annotations for four different targets allowing for an *animal/meat* alternation in their interpretation: seminal examples *lamb* and *chicken*, less frequent *pheasant* and experimental *seagull.* The similarity scores obtained for the different sense interpretations of these targets are visualised in Figure 5.3, with sense `1` representing the *animal* reading and sense `2` the *meat* one. Like for *newspaper* and

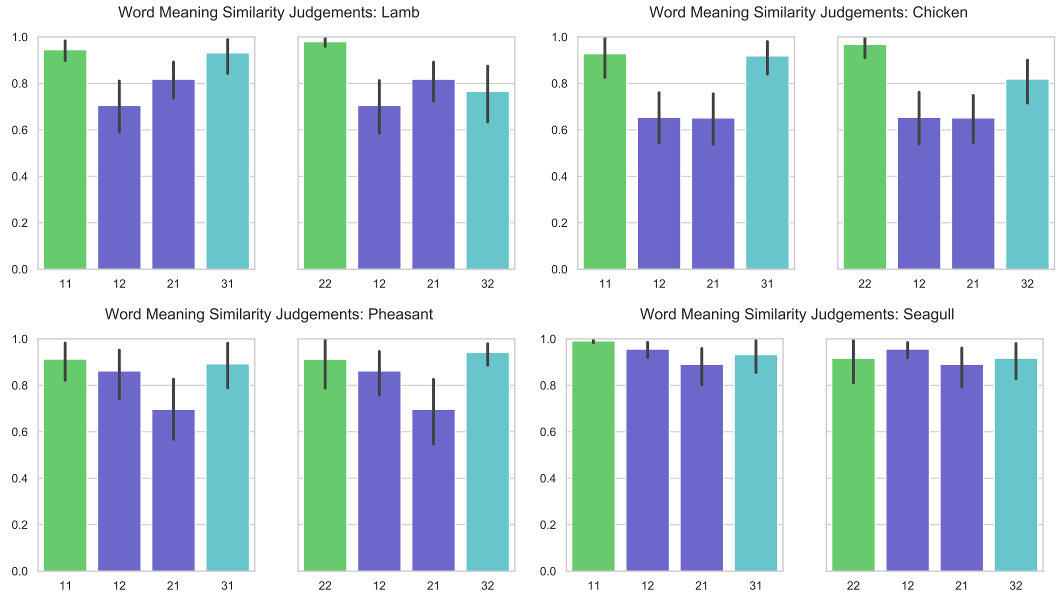


Figure 5.3: Mean word sense similarity ratings given to same-sense (green) and cross-sense (blue) sentence pairings eliciting different interpretations of targets *lamb*, *chicken*, *pheasant* and *seagull*. Experimental context modifications in teal. Senses: 1-animal, 2-meat, 3-animal⊕.

magazine, all same-sense comparisons were rated to be highly similar, indicating that the different context samples developed for these comparisons indeed invoke similar interpretations. For the seminal targets, cross-sense comparisons were rated significantly lower than the same-sense comparisons. For the less common *pheasant* (which is listed with both an *animal* and *meat* reading in WordNet[4]) and the experimental *seagull* (which according to WordNet only has an *animal* reading[5]) - while cross-sense ratings are lower - the differences are not always significant.

In our materials we used the following context samples for the targets *chicken* and *seagull*:

(57) 1a The **chicken** pecked for some food pellets in the new feeder.
1b The **chicken** sat on the roof of the coop all afternoon long.
2a The **chicken** was served with steaming hot potato wedges.
2b The **chicken** tasted like it had been marinated for at least 12 hours.

(58) 1a The **seagull** kept circling over a food stall near the promenade.
1b The **seagull** stole a sandwich from an unsuspecting beachgoer.
2a The **seagull** definitely tasted better than anyone could have imagined.
2b The **seagull** had been roasted on a spit over a makeshift campfire.

[4] `http://wordnetweb.princeton.edu/perl/webwn?s=pheasant`

[5] `http://wordnetweb.princeton.edu/perl/webwn?s=seagull`

One approach to explain the observed differences in their cross-sense comparisons is that in the mental representation of our annotators, *seagull* only has an *animal* reading, or at least a strong bias towards the *animal* reading. This in turn would mean that when processing the sample contexts designed to elicit a *meat* or *food* reading, the interpretation of the target still mostly represents the animal, and what was designed to be a cross-sense reading is interpreted like a same-sense reading. In addition, this effect could be amplified if the samples themselves allowed for some vagueness or ambiguity. The predication *tasted* in sample (58) 2a however clearly calls for a *meat* reading, while *has been roasted* in 2b could be interpreted as the animal still, and might explain the slightly lower similarity score for comparison 21 (samples 2a and 1b) than for 12. Still, both of these similarity scores are significantly higher than those for the *chicken* cross-sense pairings, with all sample contexts for the *animal* reading arguably unambiguously calling for an animate subject (*pecked*, *sat*, *kept circling* and *stole*).

A second approach at explaining why the *animal* and *meat* reading for *chicken* are perceived more dissimilar than those of *seagull* ties the ease of disambiguation to the frequency of use and the prototypicality of the *animal-for-meat* interpretation of *chicken*. That is to say that the mental representation of *chicken* as an animal and the representation of *chicken* as meat or food could be quite clear - and clearly distinct - while there is no pre-formed, clear representation of *seagull* as food. In this case, when encountering *chicken* in the different contexts of Example (57), annotators either directly think of an Orpington hen or a golden-brown chicken breast, which leads to lower similarity ratings in the cross-sense samples, while *seagull* only elicits a (coerced) animal reading, leading to overall higher similarity ratings in the cross-sense setting.

Some support for this second consideration comes from the similarity ratings obtained for the experimental *context*⊕ samples included in the materials. As described in Section 5.1.2, we included some samples that - while clearly eliciting one sense interpretation - were designed to facilitate coercion or sense-shifting to another sense, labelled as *context*⊕. For *chicken* and *seagull*, the developed *animal*⊕ contexts, i.e. contexts invoking an *animal* reading while facilitating a *meat* reading, were

(59) 1c The **chicken** was bred by a well-known family of poultry breeders.
1c The **seagull** was the only thing they were able to catch that day.

These samples were each combined with the second sample for the target's respective *animal* and *meat* interpretations to create items 31 and 32 (see the teal bars in Figure

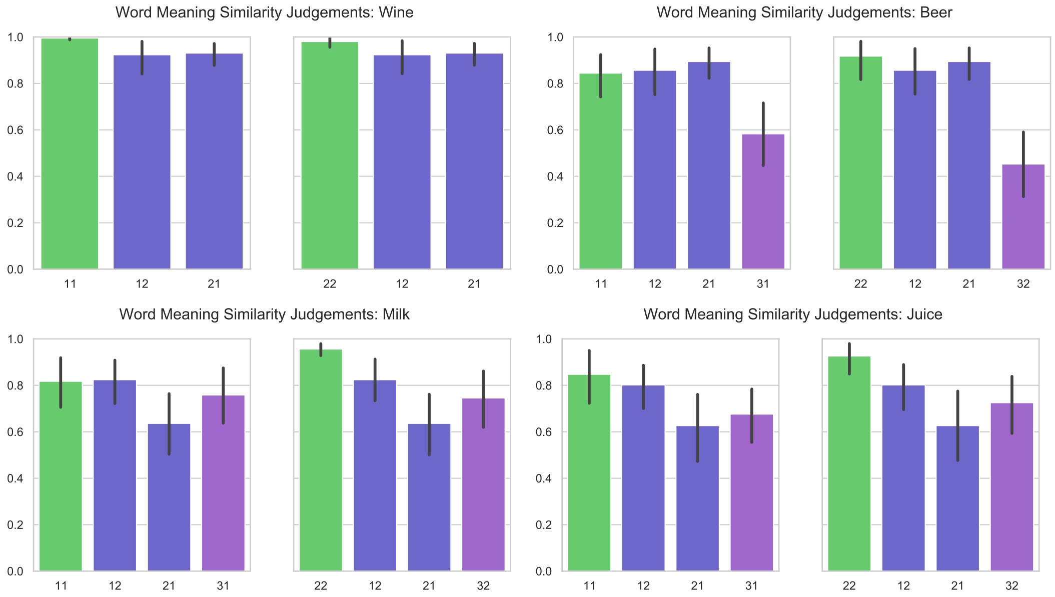


Figure 5.4: Mean word sense similarity ratings given to same-sense (green) and cross-sense (blue) sentence pairings eliciting different interpretations of targets *wine*, *beer*, *milk* and *juice*. Experimental context modifications in purple. Senses: 1-container, 2-content, 3-container⊖.

5.3). While for seminal target *chicken* comparison 32 receives a significantly lower rating than the 31 comparison but a significantly higher similarity rating than either of the original cross-sense pairings 12 and 21, for *seagull* none of the differences are significant. This observation could be taken to mean that for *chicken* - where we assume the representations for the *animal* and *food* readings to be more distinct - facilitating a shift between the two senses by introducing the animal as something that will be turned into a foodstuff increased the perceived similarity of the two interpretations, but for *seagull* we observe no such effect as the food reading still is not readily available.

Content-for-Container: Wine, Beer, Milk and Juice. The collected data includes annotations for seminal *content-for-container* targets *wine* and *beer*, as well as for experimental targets *milk* and *juice*. The average similarity ratings of different combinations of their sense interpretations are visualised in Figure 5.4. For both *wine* and *beer*, the cross-sense comparisons receive similarity ratings resembling those of the same-sense comparisons, while for *milk* and *juice* at least some of the cross-sense comparisons are rated significantly lower in similarity than the same-sense pairings.

At a first glance, this behaviour seems to be the inverse of what has been observed for the *animal/meat* alternation, where the seminal, more frequent targets displayed bigger differences in the ratings for same-sense and cross-sense pairings than the experimental targets. We expect however that the underlying effect could be quite similar: due to the prototypicality and frequency of targets *wine* and *beer* being used in their *container* interpretation, this coerced reading could more readily available for the seminal targets than for the experimental ones, which here facilitates an under-specified interpretation of the ambiguous target and consequently leads to higher similarity judgements when comparing different samples. With the *container* interpretation less available in the processing of the less prototypical targets, the shift in use would be more obvious to annotators, and lead to overall lower similarity scores. This hypothesis again gains some support from the experimental context samples included in our materials, in this case the *context*⊖ variant of the content reading. The *container*⊖ variant here explicitly focuses on the physical properties of the *container* reading, and was intended to impede a *content* interpretation of the target. The full set of sample contexts for *beer* and *juice* used in this study were:

(60) 1a The **beer** left a ring of condensed water on the cardboard coaster.
1b The **beer** was filled much lower than the fill line on the label.
1c The **beer** was made of recycled glass fished from the ocean. (`container`⊖)
2a The **beer** tasted exactly like Sue had remembered it.
2b The **beer** thoroughly refreshed Ben after his 10k evening run.

(61) 1a The **juice** was just too large to fit into the fridge's door compartment.
1b The **juice** had drawings of exotic fruits Sue had never seen before.
1c The **juice** got squished when they dropped it from the shelf. (`container`⊖)
2a The **juice** was made out of 100% fresh, sun-riped fruit.
2b The **juice** was sweetened by naturally occurring fructose only.

For seminal target *beer*, the *container*⊖ reading 1c significantly reduced the similarity rating when compared to *content* reading 2b (items `12` and `32` in the *beer* sub-plot of Figure 5.4), while this had a less drastic effect for *milk* and *juice*, where in both cases the *container*⊖ rating falls between the two original cross-sense variants.[6] While this again is an observation based on only a small number of experimental samples, it could indicate that focusing on the physical aspects of the *container* reading impedes with an under-specified representation of the seminal target and forces it into two more clearly distinct interpretations that lead to a higher dissimilarity in their comparison. Because the less prototypical targets didn't allow for an under-specified interpretation in the first place, this impeding focus has less of an effect on

[6] As we included these experimental targets only in the second annotation run, we do not have a rating for the *container*⊖ reading of *wine*.

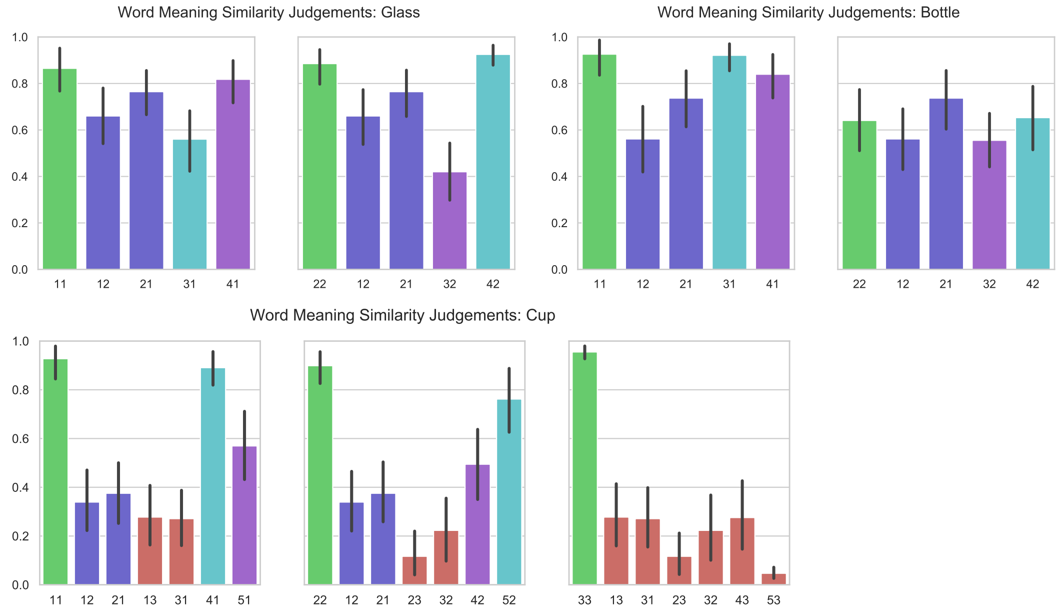


Figure 5.5: Mean word sense similarity ratings given to same-sense (green) and cross-sense (blue) sentence pairings eliciting different interpretations of targets *glass*, *bottle*, and *cup*. Homonymic interpretations in red, experimental context modifications in teal and purple. Senses for *glass* and *bottle*: 1-container, 2-content, 3-container⊖, 4-content⊖. Senses for *cup*: 1-container, 2-content, 3-trophy, 4-container⊖, 5-content⊖.

the perceived similarity of the different context samples.

Container-for-Content: Glass, Bottle, Cup. Analogous to the *content-for-container* targets described in the previous section, we also collected similarity judgements for a number of words allowing for a *container-for-content* reading. We here focused on seminal targets *glass* and *bottle*, and included less frequent target *cup* which also allows for a range of homonymic interpretation, including one as *trophy*. The mean similarity judgements for different combinations of context samples containing these lexically ambiguous targets are visualised in Figure 5.5. While cross-sense items (blue) were overall rated lower than the same-sense comparisons, here these differences again were not always significant for the seminal targets but very clear for experimental target *cup*, where cross-sense samples were rated similarly to homonymic meaning comparisons (depicted in red).

In the materials for *content-for-container* targets we included two additional experimental context samples: *container*⊖ and *content*⊖, each focusing on either the *physical* properties of the container or the *liquid/beverage* properties of the content, intended to impede with a cross-sense shifting of the target's interpretation:

(62) **Glass**

3 The **glass** chipped when they accidentally hit it with a billiard cue. (container⊖)

4 The **glass** seemed to be some kind of high-caffeine energy drink. (content⊖)

Bottle

3 The **bottle** was made out of recycled glass fished from the ocean. (container⊖)

4 The **bottle** was a fruit spirit produced by a family-run distillery. (content⊖)

Cup

4 The **cup** had a beautiful handle shaped to look like a snake. (container⊖)

5 The **cup** was made out of 100% fresh, sun-riped fruit. (content⊖)

We expected that the *container⊖* samples should lead to relatively high similarity ratings when compared with the original *container* samples (sense 1) - as this creates a de-facto same-sense pairing - and produce relatively low similarity scores when compared with a *content* reading (sense 2), likely even lower than the rating for the original cross-sense comparisons. This effect was expected to be inverted for the *content⊖* samples.

For seminal target *glass*, we found that the similarity between the *container⊖* sample and the original *container* sample (sense 1, comparison 31) was rated lower than the original cross-sense comparisons, and lower than the *content⊖/container* comparison 41. However, *container⊖/content* comparison 32 was rated even lower than this, and *content⊖/content* comparison 42 received a similarity score higher than the original same-sense comparison. This means that while comparisons 32 and 42 yielded expected results, both 31 and 41 did not. One reason for this observation could be a problematic sample context for the *container* reading 1, but here seems unlikely as the same-sense comparison 11 elicited a high similarity rating, which to some degree validates the used context samples. Considering this, for now we cannot readily explain the unexpected results for comparisons 31 and 41.

Inspecting the ratings for *bottle*, we find that same-sense comparison 31 has been rated higher in similarity than the original cross-sense comparisons 12 and 21 - but so is *content⊖/container* comparison 41. None of the comparisons containing a coerced container-for-content reading (sense 2) was rated significantly different than any other, with same-sense, cross-sense and experimental same-sense and cross-sense items all fluctuating around the 0.6 mark. Here the low same-sense similarity scores could indicate that at least one of the sample contexts for sense 2 was not as clear as required for a comparison as sensitive as this one - and given that comparison 21 is rated highest, the culprit here is likely to be sample 2b, which is not used in this comparison but in all others. The sample contexts used for the *content* sense 2 in our materials were

(63) 2a The **bottle** tasted exactly like Sue had always imagined.
     2b The **bottle** made them talk a lot louder than they normally did.

By referring to the effect of the alcohol contained in the content of the container, sample context 2b here potentially created too complex of a bridging reference to clearly invoke either a *container* or the intended *content* reading required for the experiment.

Experimental target *cup* lastly displayed very clear results: cross-sense comparisons 12 and 21 were rated significantly lower than the same-sense comparisons 11 and 22, and only slightly higher than the homonymic pairings. *Container⊖/ container* comparison 41 here was rated similar to same-sense comparison 11, and *content⊖/container* comparison 51 significantly lower than that - while still significantly higher than any of the original cross-sense and homonymic comparisons. *container⊖/content* comparison 42 on the other hand was rated slightly higher than the original cross-sense comparisons, and experimental same-sense comparison 52 (*content⊖/content*) received a mean similarity rating significantly lower than the original same-sense pairing 22, but also significantly higher than any of the cross-sense or homonymic comparisons.

Summarising our observations for these three targets, we find that for the *container-for-content* alternations tested in our study, seminal targets *glass* and *bottle* led to equivocal results, while the less frequent *cup* shows a clear distinction in the similarity ratings given to same-sense and cross-sense comparisons. This indicates that here the two different interpretations of the ambiguous target either are clearly distinguished, or that the *content-for-container* reading is less available for the less prototypical target.

Other alternations. The *event/food* alternation of seminal targets *lunch* and *dinner* showed similar effects as the three types of alternations discussed above. Same-sense sample combinations overall were being rated significantly higher than the cross-sense samples, with the lowest rating for cross-sense pairings obtaining a

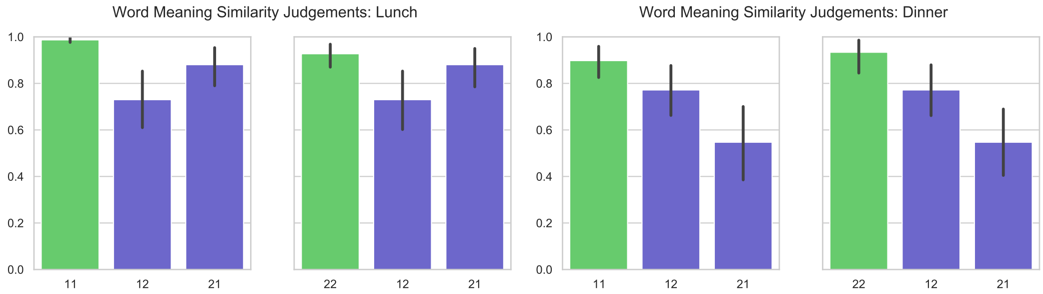


Figure 5.6: Mean word sense similarity ratings given to same-sense (green) and cross-sense (blue) sentence pairings eliciting different interpretations of targets *lunch*, and *dinner*. Senses: 1-food, 2-event.

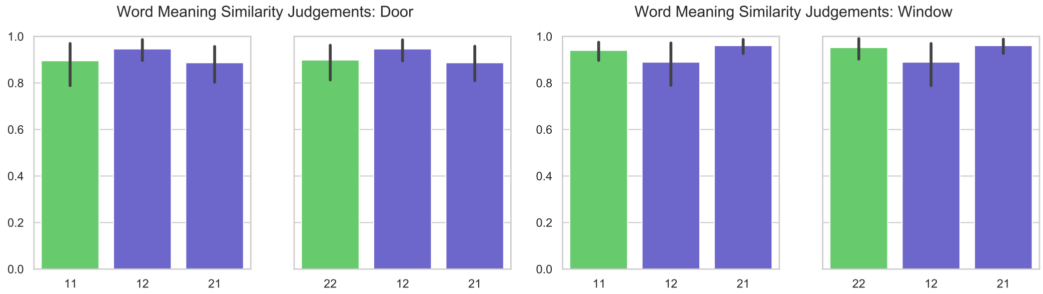


Figure 5.7: Mean word sense similarity ratings given to same-sense (green) and cross-sense (blue) sentence pairings eliciting different interpretations of targets *door*, and *window*. Senses: 1-opening, 2-physical.

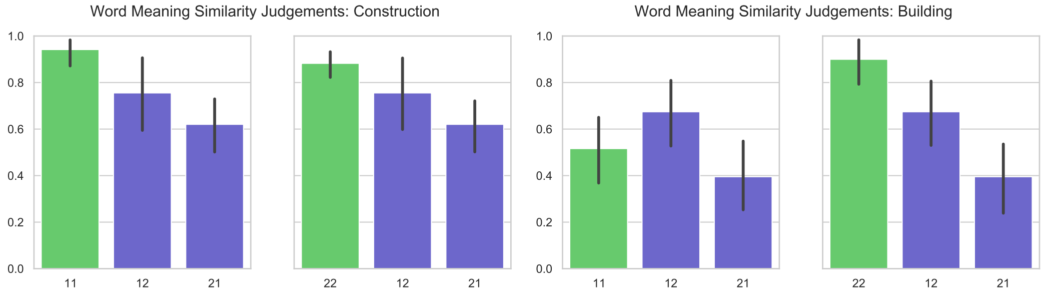


Figure 5.8: Mean word sense similarity ratings given to same-sense (green) and cross-sense (blue) sentence pairings eliciting different interpretations of targets *construction*, and *building*. Senses: 1-process, 2-result.

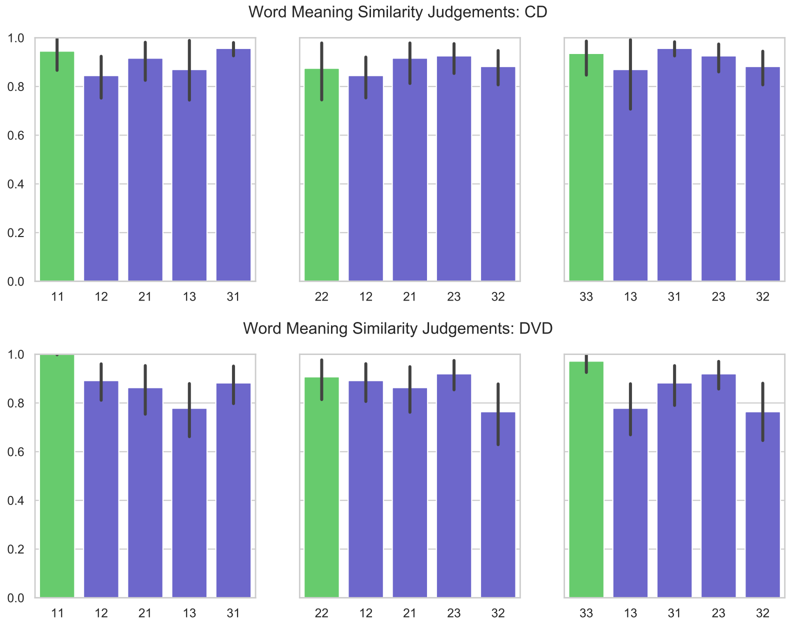


Figure 5.9: Mean word sense similarity ratings given to same-sense (green) and cross-sense (blue) sentence pairings eliciting different interpretations of targets *CD*, and *DVD*. Senses: 1-physical, 2-medium, 3-content.

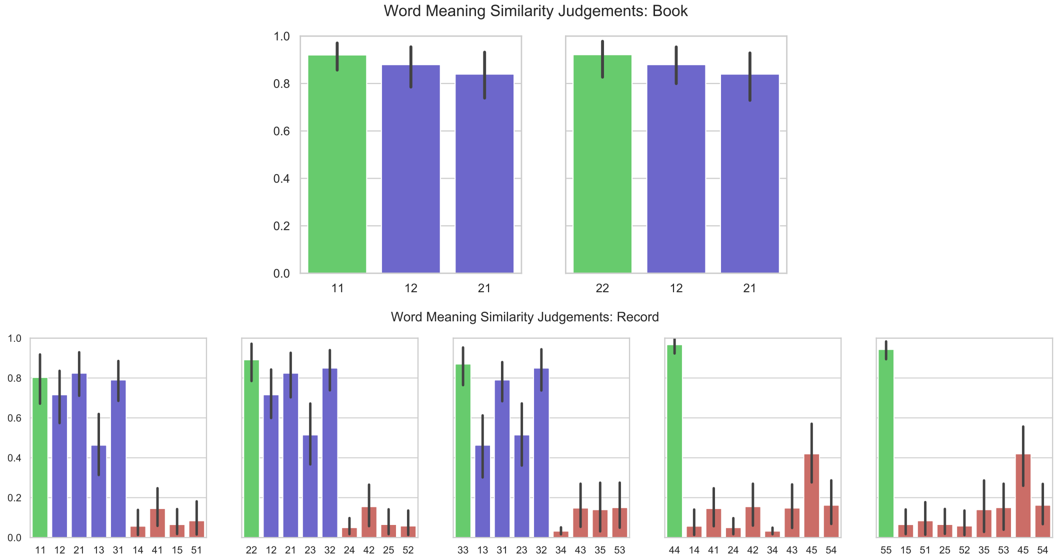


Figure 5.10: Mean word sense similarity ratings given to same-sense (green) and cross-sense (blue) sentence pairings eliciting different interpretations of targets *book*, and *record*. Homonymic interpretations in red. Senses: 1-physical, 2-content, 3-album, 4-paperwork, 5-achievement.

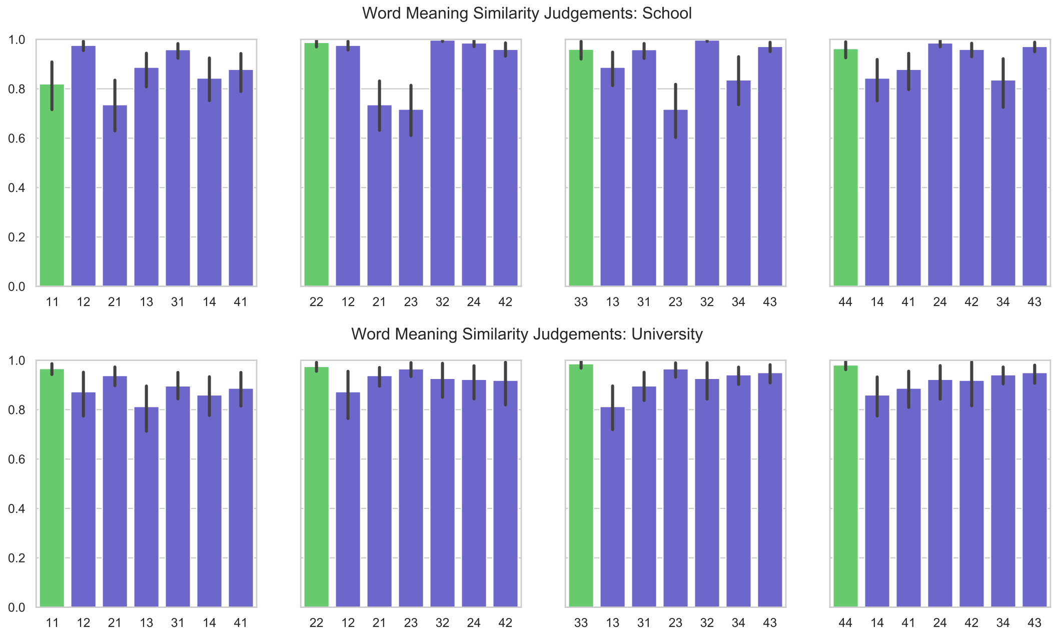


Figure 5.11: Mean word sense similarity ratings given to same-sense (green) and cross-sense (blue) sentence pairings eliciting different interpretations of targets *school*, and *university*. Senses: 1-building, 2-administration, 3-institution, 4-students.

similarity judgement of 0.55 in sample `21` for target *dinner* (see Figure 5.6). The two readings here thus seem to be perceived as distinct, but more closely related than homonymic meaning alternations.

For *door* and *window*, our targets for the *physical/aperture* alternation, we observed no significant differences in ratings for same-sense and cross-sense comparisons (Figure 5.18). This indicates that the two readings here are not perceived as distinct, and could potentially be represented as different facets in the same underspecified entry of the mental lexicon. Interestingly, the *physical/aperture* alternation of *door* has famously been used by Cruse (1995) to showcase that polysemic alternations can lead to zeugmatic co-predication.[7] We will investigate the co-predication ratings for this sample in Section 5.3.2.

Targets *CD* and *DVD* yielded mixed results, with similarity ratings for same-sense and cross-sense comparisons of *CD* showing only little differences, while *DVD* shows some significant drops of cross-sense ratings compared to the same-sense ratings - partially due to near perfect similarity scores for same-sense comparisons `11` and `33` (see Figure 5.19). The related *physical/information* alternations of *book* and *record* displayed similar effects, with *record*'s same-sense ratings referring to the *medium* receiving comparatively low similarity ratings between 0.8 and 0.9 only,

[7] As exemplified by ‘They took the door of its hinges and walked through it.’

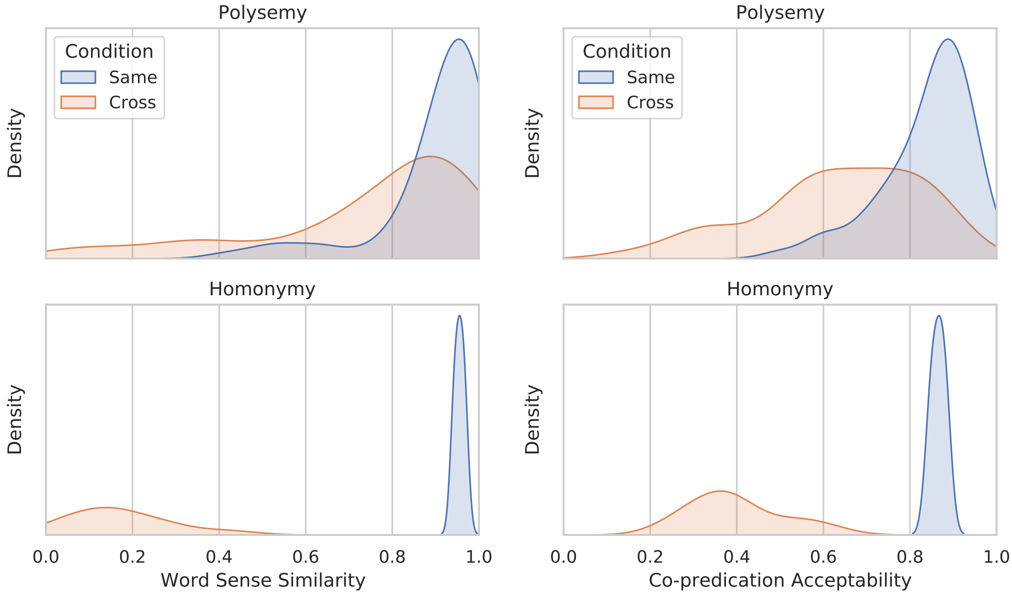


Figure 5.12: Normalised distributions of explicit sense similarity ratings (left) and co-predication acceptability judgements (right) given to same-sense (blue) and cross-sense (orange) samples with polysemic and homonymic alternations.

while same-sense comparisons of the homonymic interpretations relating to a *file* (sense 4) or *achievement* (sense 5, see Figure 5.20). Still, most polysemic cross-sense pairings are overall rated significantly higher than homonymic cross-sense pairings.

Lastly, investigating similarity judgements collected for polysemic targets *school* and *university*, taken to refer to a *building* (sense 1), an *administration* (sense 2), an *institution* (sense 3) or a collection of *students* (sense 4), we find that for *school*, the cross-sense comparisons' similarity scores fluctuate between less similar, equally as similar, and sometimes even more similar than those of same-sense comparisons. For *university*, same-sense scores are consistently very high, and cross-sense samples consequently more often significantly lower, even though here the numerical disparity is not as large as for school (see Figure 5.11).

We will return to a more detailed analysis of patterns in the similarity ratings obtained for different types of alternations in Section 5.3.4

5.3.2 Co-Predication Acceptability Ratings

Besides explicit similarity ratings, we collected an additional 8,640 judgements from 192 participants rating the acceptability of co-predication structures created from our sample sentences. After adding data collected during the pilot and filtering out noisy annotations, we retained a total of 7,379 judgements, for an average of 16.75 annotations per item word (minimum 12).

To filter low-quality annotations in the co-predication study, we again used two different control items. One of the control items in each questionnaire contained a sentence of similar length to the test items, but exhibited a regular conjunctive clause instead of an actual co-predication structure to prevent any accidental infelicitous co-predication. An example of such a `test-same` item is

(64) A group of boys were playing Frisbee in the park and a girl tried to balance on a slack line.

`Test-random` items on the other hand started as a regular conjunctive sentence mentioning a (potentially homonymic) target, but the conjunctive clause would be a scrambled filler sentence:

(65) The match ended without a clear winner and the off the managed bass hook get to.

Submissions were excluded from analysis if *both* the `test-same` item was rated below 0.7 acceptability and the `test-random` item was rated above 0.2 acceptability. Per-questionnaire inter-annotator agreement here only reached a Krippendorff's alpha rating of 0.34, indicating stronger individual differences in the participants' use of the continuous rating scale.

In co-predication structures, predication order can have a major effect on the acceptability of the resulting sentence (e.g. Murphy, 2021, also see Chapter 2.5.1). Since in this study we aimed to investigate co-predication acceptability as an indicator of word sense similarity, considerable effort was spent on reducing these ordering effects by stipulating a strict sample template (see Chapter 4.3.2). To investigate whether our samples still showed signs of predicate ordering effects, we again compared all pairs of items that contained the same predications but in a different order, as exemplified in Example (66) for *newspaper*'s item `12`:

(66) 12a The newspaper fired its editor in chief and got wet from the rain.
12b The newspaper got wet from the rain and fired its editor in chief.

Since this comparison was comprised of a total of 229 of order-pairs, we applied Bonferroni Correction to determine a more conservative significance threshold accounting for the large number of tests. Only 1 of the 229 pairwise comparisons passed this Bonferroni corrected significance level of 0.00021. We therefore argue that our samples are mostly free from predication order effects, and indeed primarily test for the acceptability of invoking different senses of the target words in the same sentence. Based on this observation, we again combined results before further analysis.

The right column of Figure 5.12 shows the distributions of collected co-predication acceptability ratings split by sample condition and ambiguity type. The average acceptability rating for co-predication structures invoking the same sense in both predications is 0.83, the mean acceptability for homonymic cross-sense samples is 0.41, and the mean acceptability for polysemic alternations is 0.64-significantly lower than the same-sense mean but significantly higher than the homonym mean (see Table 5.1, row 2). These results support previous observations of co-predication acceptability also appearing to be a graded measure rather than a binary signal, and challenge co-predication as a linguistic test to distinguish polysemy from homonymy. Same-sense and homonymic samples were rated quite consistently, with only 10.34% and 5.88% of pairwise comparisons passing Bonferroni correction, respectively. Polyseme samples again show some degree of inconsistency, with 21.66% of comparisons among polysemic cross-sense samples passing the corrected significance threshold of 0.00015. These results parallel the observations made when investigating the explicit similarity ratings, and provide additional evidence for the non-uniformity in interpreting polysemic samples.

Qualitative Analysis

Figure 5.13 shows the mean acceptability ratings for co-predication structures eliciting different combinations of interpretations of the ambiguous targets *newspaper* and *magazine*. Like in all of the following figures of this kind, ratings for co-predication structures eliciting the same sense are shown first, in green colour, ratings for polysemic cross-sense samples are shown next, in blue, and additional homonymic cross-sense pairings are shown last, in red colour. The x-axis again shows the combination of sample contexts indicated by their context identifiers (first position indicates variant a, and second position variant b), and does not indicate predicate ordering, as results from both orderings were combined for this analysis. The y-axis displays mean acceptability on a scale from 0 (totally unacceptable) to 1 (perfectly acceptable). Ratings again are grouped into different sub-plots by sense, with some being repeated for clarity.

Newspaper and Magazine The acceptability ratings for *newspaper* show a very similar behaviour to the similarity scores for the same target as displayed in Figure 5.13, albeit with consistently lower acceptability scores than similarity ratings. All cross-sense structures are rated significantly lower than the same-sense combinations, which are rated above 0.9 acceptability. The differences between the *organisation* reading (sense 1) and the *physical* reading in sense 2 here however become even

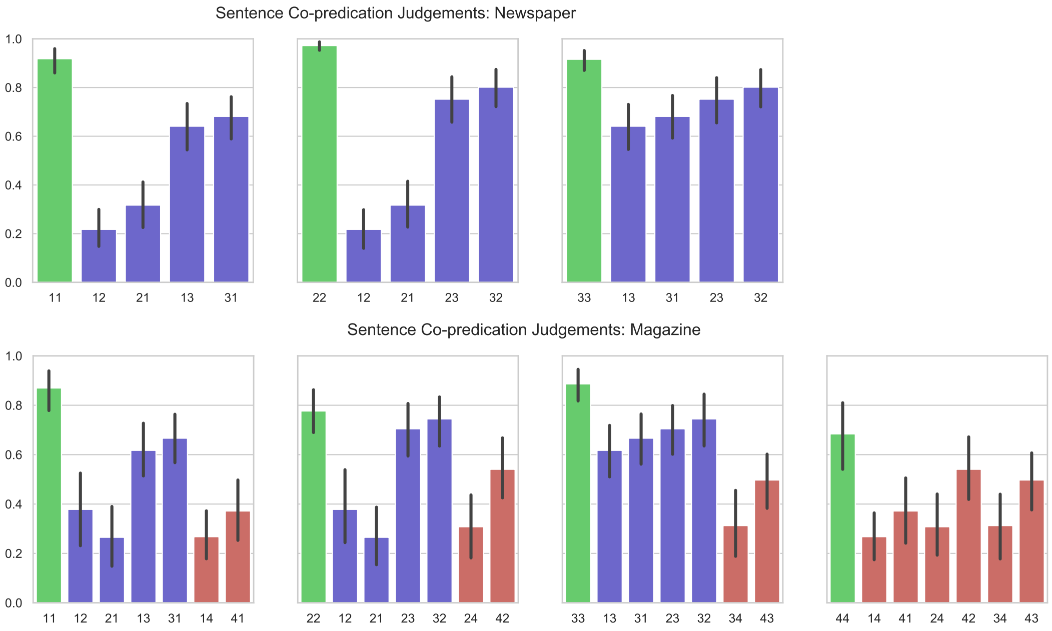


Figure 5.13: Mean acceptability ratings given to same-sense (green) and cross-sense (blue) co-predication structures eliciting different interpretations of targets *newspaper* and *magazine*. Additional homonymic comparisons are shown in red. Senses: 1-organisation, 2-physical, 3-content, 4-storage type.

more apparent, with co-predication acceptability scores for combinations 12 and 21 rated with an acceptability between 0.2 and 0.3-significantly lower than the ratings for other cross-sense combinations rated at acceptability levels between 0.6 and 0.8. Related target *magazine* overall duplicates this behaviour, showing that *organisation/physical* structures are rated at similar acceptability levels as the cross-sense structures with a homonymic alternation. This indicates that at least for this sense alternation, co-predication tests would not be able to distinguish between the polysemic *organisation/physical* and homonymic *organisation/storage* combinations. For *newspaper*, two of the same-sense pairings were annotated with acceptability scores below 0.8: The *physical/physical* combination 22 and the homonymic same-sense item 44:

(67) 22a The magazine just kept falling off the small living room table and was covered in paw prints after a cat sat on it.

22b The magazine was covered in paw prints after a cat sat on it and just kept falling off the small living room table.

44a The magazine contained all kinds of defunct WW2 weaponry and was originally designed for storing ballistic missiles.

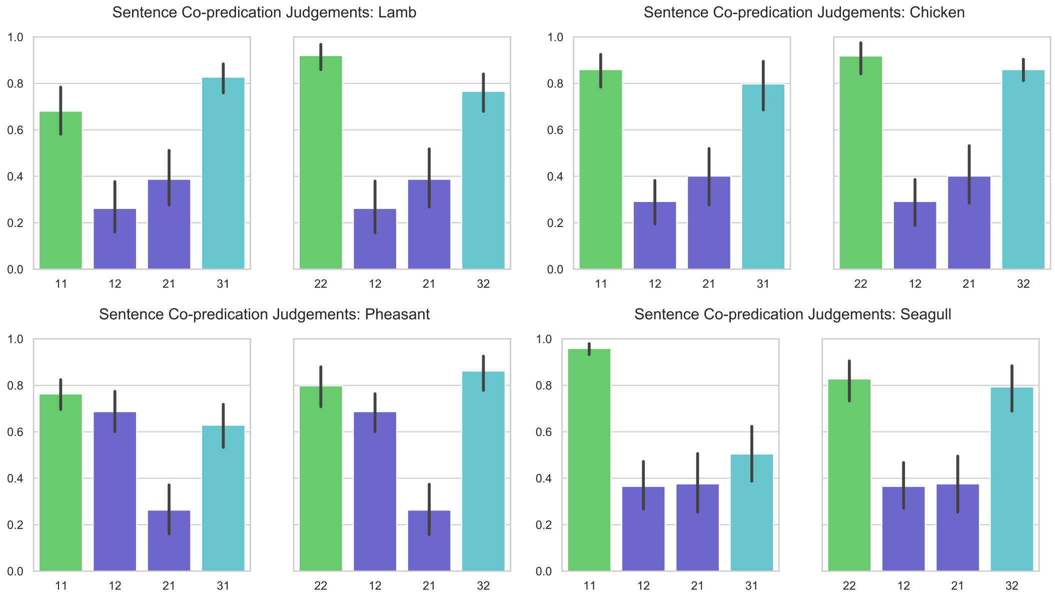


Figure 5.14: Mean acceptability ratings given to same-sense (green) and cross-sense (blue) co-predication structures eliciting different interpretations of targets *lamb*, *chicken*, *pheasant* and *seagull*. Experimental context modifications in teal. Senses: 1-animal, 2-meat, 3-animal⊕.

44b The magazine was originally designed for storing ballistic missiles and contained all kinds of defunct WW2 weaponry.

While in the 22 samples the complexity of the predications might be to blame for the lower acceptability scores - especially considering that the sample pairings received relatively high similarity scores - we again cannot pinpoint a clear explanation of the drop in acceptability judgements for the 44 samples other than the possible interpretation bias and potential ambiguity mentioned in the analyses of the *magazine* similarity judgements in Section 5.3.1.

Animal/Meat: Lamb, Chicken, Pheasant and Seagull The acceptability judgements given to different combinations of the *animal/meat* interpretations of seminal targets *lamb* and *chicken* and the less frequent, more experimental options *pheasant* and *seagull* overall show a very similar behaviour. In contrast to the similarity ratings where the difference between same-sense and cross-sense ratings was bigger for the seminal targets than for the experimental ones (see Figure 5.14), here the cross-sense acceptability ratings are consistently significantly lower than the same-sense ratings - with the noteworthy exception of sample 12 for *pheasant*, which received an acceptability rating not significantly lower than the same-sense sample 11, and much higher than for example the same combination of senses elicited by different predications in item 21.

(68) 12a The pheasant definitely gave the hunters a run for their money and tasted much better than anything he'd ever eaten.

12b The pheasant tasted much better than anything he'd ever eaten and definitely gave the hunters a run for their money.

21a The pheasant was marinated in milk over night to make it tender and was foraging for some seeds on a small clearing.

21b The pheasant was foraging for some seeds on a small clearing and was marinated in milk over night to make it tender.

Investigating these materials, the predication 'gave the hunters a run for their money' in sample 12 - by virtue of mentioning hunters who contribute to turning the animal into a foodstuff - could facilitate sense shifting towards a *food* reading, and therefore not as unequivocally require an *animal* reading as the 'was foraging for some seeds on a small clearing' alternative in 21. This hypothesis gains some support from the experimental *animal*⊕ samples, which were specifically designed to facilitate sense shifting by introducing the animal as something reared to be processed into a food. In most cases, co-predication structures including an experimental *animal*⊕ reading received acceptability scores similar to - or even higher than - the original same-sense structures, independent of whether they produced same- or cross-sense predications. This effect here is much more pronounced than for the explicit word similarity judgements, and underlines the sensitivity of co-predication tests towards what exactly is predicated of the targets rather than which exact sense interpretation that predication evokes. One exception to this observation is the *seagull* sample 31 where the predication 'the only thing they were able to catch' might put a logical restriction on the availability of the target for the other predication, and consequently lead to lower acceptability scores than comparable structures:

(69) 31a The seagull was the only thing they were able to catch that day and stole a sandwich from an unsuspecting beachgoer.

31b The seagull stole a sandwich from an unsuspecting beachgoer and was the only thing they were able to catch that day.

Content-For-Container: Wine, Beer, Milk and Juice The acceptability ratings for the *content-for-container* items included in our materials again seem to depend strongly on what is being predicated in the samples contexts: For *beer*, *milk* and *juice*, cross-sense co-predication structures are in some cases rated to be

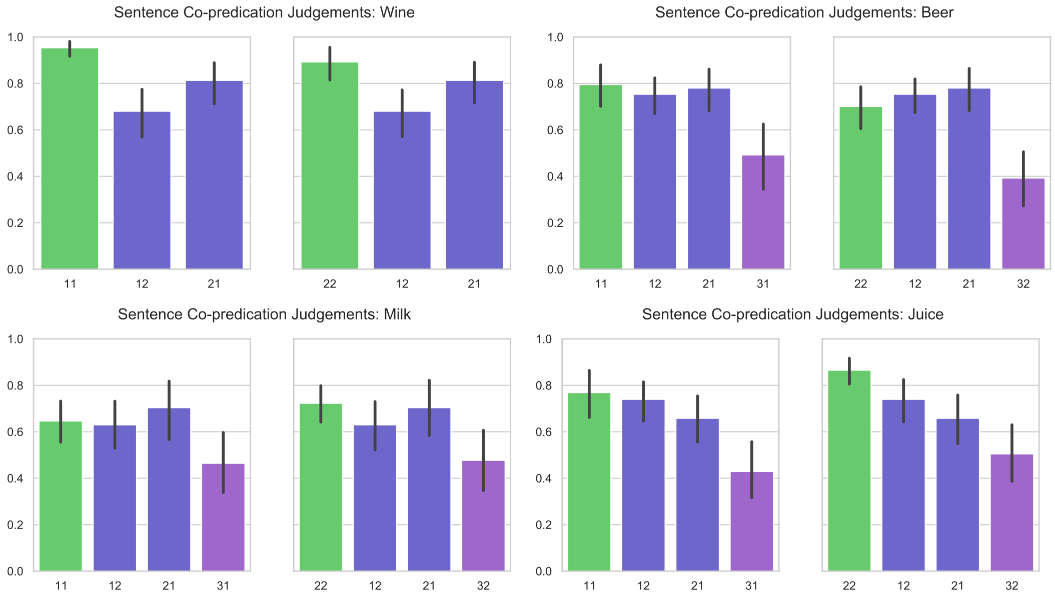


Figure 5.15: Mean acceptability ratings given to same-sense (green) and cross-sense (blue) co-predication structures eliciting different interpretations of targets *wine*, *beer*, *milk* and *juice*. Experimental context modifications in purple. Senses: 1-container, 2-content, 3-container⊖.

more acceptable than the respective same-sense sample (see for example *beer*'s items `12` and `21` compared to `22`), in some cases they received comparable acceptability scores (*beer*'s `12` and `21` versus `11`) - and in some cases, ratings are significantly lower (*juice*'s `12` and `21` versus `22`). Only for seminal target *wine* are the cross-sense constructions rated consistently less acceptable than the same-sense alternatives. A central aspect of this observation are the low acceptability scores for the same-sense comparisons, going as low as 0.65 for *milk*'s *container/container* item `11` while the container/container construction for *wine* received an acceptability score of 0.95:

(70) 11a The milk was just too large to fit into the fridge's door compartment and had a beautiful drawing of grazing cows on the front.

11b The milk had a beautiful drawing of grazing cows on the front and was just too large to fit into the fridge's door compartment.

(71) 11a The wine lay in a padded wooden box and was a little dusty from storage.

11b The wine was a little dusty from storage and lay in a padded wooden box.

One immediately apparent difference here is the length and consequently the potential complexity of the sentences. While our sample template called for contexts that were 'as short as possible,' we found that the prototypicality of the target has an effect on how short a context sample could be for it to still be considered to 'invoke a certain sense as clearly as possible' - a second and not less crucial requirement for our items. While *wine* is typically associated with being stored in glass bottles - as it has been for centuries - we found that less prototypical targets like *milk* and *juice* do not have as strong an association with a certain container type. As a result, the *container* reading of targets like *wine* and *beer* are easier to invoke, and specifying a type of container for *milk* and *juice* requires more descriptive contexts. And while this is likely to have less of an effect on explicit similarity ratings, specifying two longer clauses can reduce the perceived acceptability of a co-predication structure combining these samples. In the case of the low acceptability reading for *beer*'s *content/content* sample 22, we however have to admit an oversight in our materials: the two sample contexts here mention different characters, which does not impact the similarity of the target use, but does decrease the acceptability of the resulting co-predication structure.

(72) 22a The beer tasted exactly like Sue had remembered it and thoroughly refreshed Ben after his 10k evening run.

22b The beer thoroughly refreshed Ben after his 10k evening run and tasted exactly like Sue had remembered it.

One last observation concerning the content-for-container alternations relates to the acceptability ratings for experimental samples containing a container⊖ predication, which consistently received significantly lower acceptability scores than the original same-sense and cross-sense structures. Focusing on the physical aspects of the coerced *container* reading here thus apparently further impedes a felicitous co-predication.

Container-for-Content: Glass, Bottle, Cup The co-predication acceptability ratings for the *container-for-content* targets *glass*, *bottle* and *cup* in Figure 5.16 display a very similar behaviour to their explicit sense similarity judgements shown in Figure 5.4. Cross-sense samples are usually rated less acceptable than the same-sense references - except for any of the *content/content* combinations (items 22), which for all three targets received relatively low acceptability scores. The *content/content* samples included in our materials were

(73) **Glass**

22 The glass tasted like an apple juice blended with forest fruits and thoroughly refreshed Ben after his 10k morning run.

Bottle

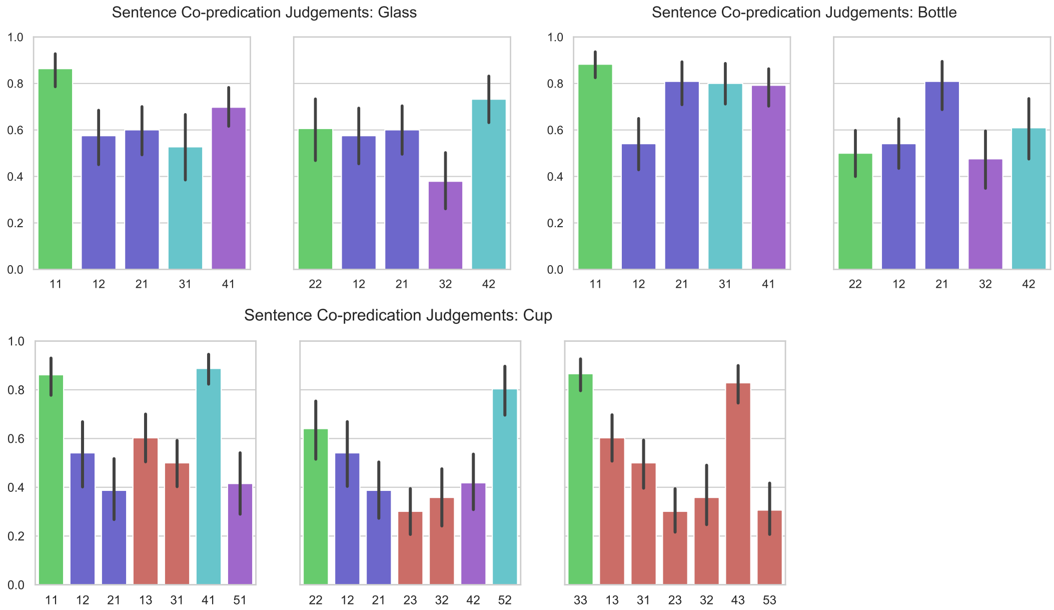


Figure 5.16: Mean acceptability ratings given to same-sense (green) and cross-sense (blue) co-predication structures eliciting different interpretations of targets *glass*, *bottle*, and *cup*. Homonymic interpretations in red, experimental context modifications in teal and purple. Senses for *glass* and *bottle*: 1-container, 2-content, 3-container⊖, 4 content⊖. Senses for *cup*: 1-container, 2-content, 3-trophy, 4-container⊖, 5- content⊖.

22 The bottle tasted exactly like Sue had always imagined and made them talk a lot louder than they normally did.

Cup

22 The cup tasted much sweeter than Jon remembered and was sweetened by naturally occurring fructose.

Individually, all of these pairings received high similarity scores, but combining them in a co-predication structure appears to reduce the acceptability of the *container-for-content* reading of the targets.

Inspecting the *container*⊖ and *content*⊖ samples included in our experiments, the effect of impeding the sense shifting is strongest for the less prototypical target *cup*, with same-sense pairings 41 and 52 receiving similar - or even higher - ratings than the original same-sense pairings, and cross-sense structures 52 and 51 rated among the lowest. The experimental *context*⊖ samples for *glass* and *beer* have a similar but less pronounced effect on the collected acceptability judgements.

Other alternations The co-predication acceptability ratings for the *event/food* alternations of *lunch* and *dinner* show less clear differences between same-sense and

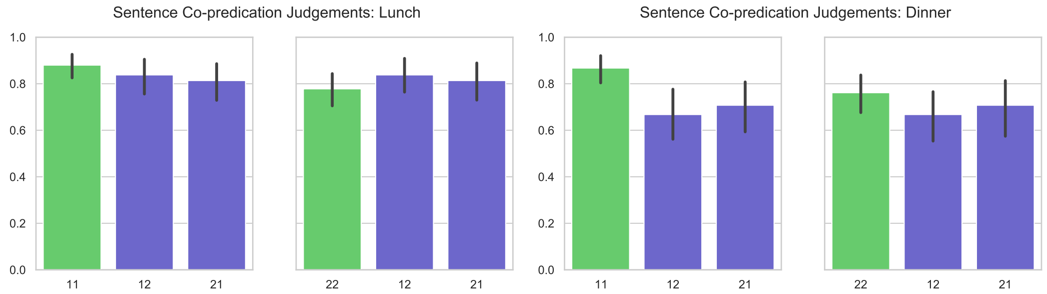


Figure 5.17: Mean acceptability ratings given to same-sense (green) and cross-sense (blue) co-predication structures eliciting different interpretations of targets *lunch*, and *dinner*. Senses: 1-food, 2-event.

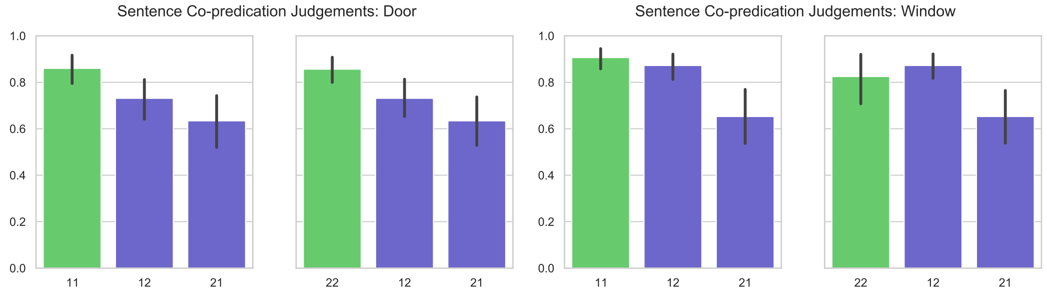


Figure 5.18: Mean acceptability ratings given to same-sense (green) and cross-sense (blue) co-predication structures eliciting different interpretations of targets *door*, and *window*. Senses: 1-opening, 2-physical.

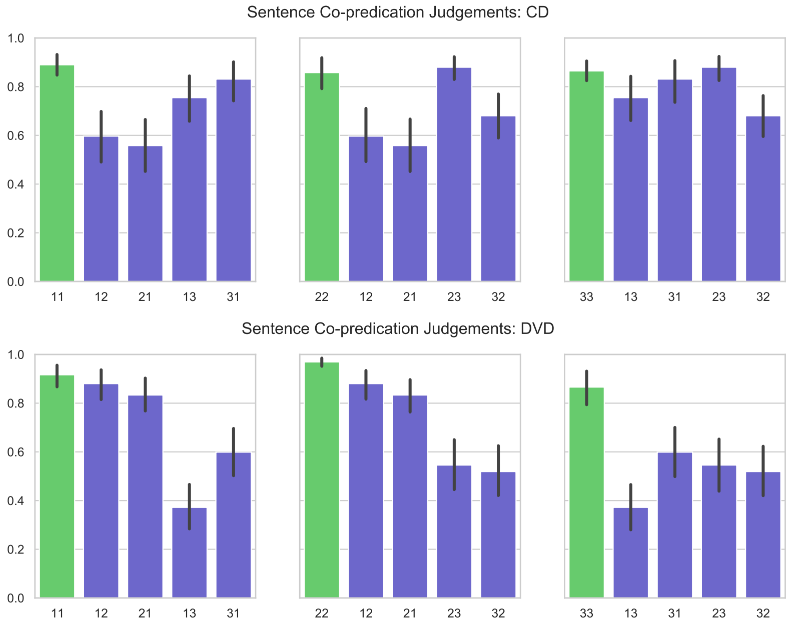


Figure 5.19: Mean acceptability ratings given to same-sense (green) and cross-sense (blue) co-predication structures eliciting different interpretations of targets *CD*, and *DVD*. Senses: 1-physical, 2-medium, 3-content.

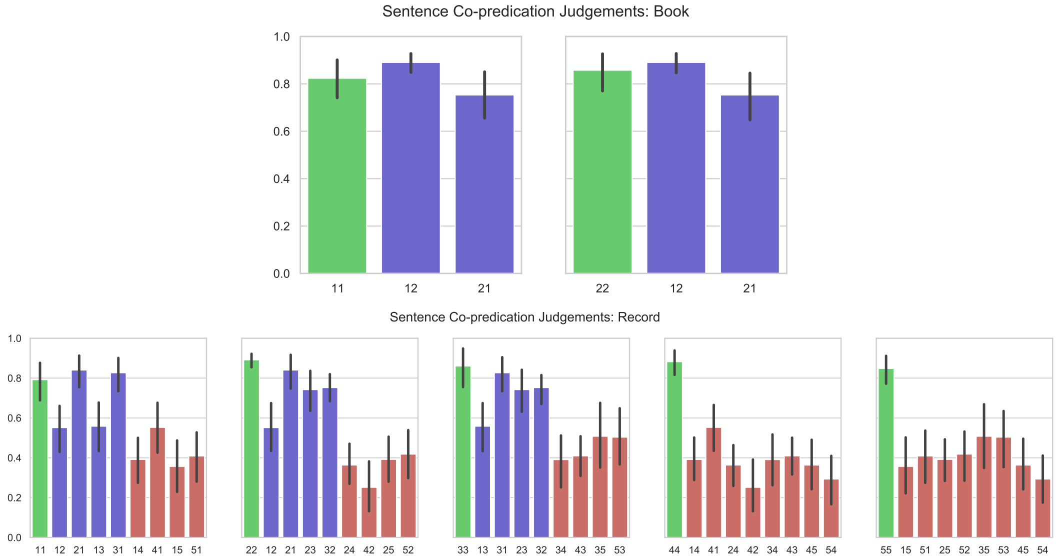


Figure 5.20: Mean acceptability ratings given to same-sense (green) and cross-sense (blue) co-predication structures eliciting different interpretations of targets *book*, and *record*. Homonymic interpretations in red. Senses: 1-physical, 2-content, 3-album, 4-paperwork, 5-achievement.

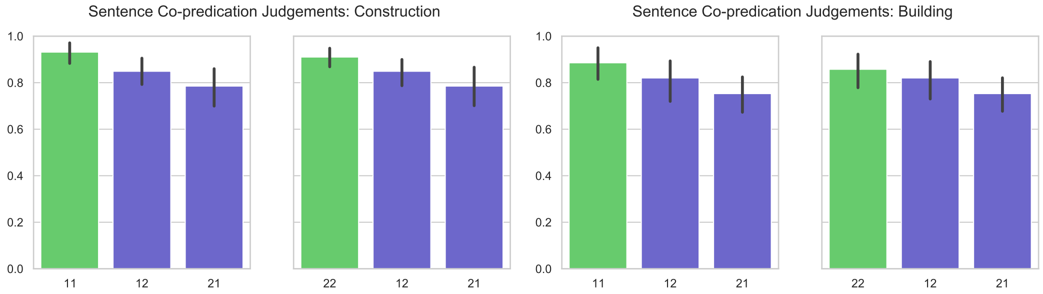


Figure 5.21: Mean acceptability ratings given to same-sense (green) and cross-sense (blue) sentence pairings eliciting different interpretations of targets *construction*, and *building*. Senses: 1-process, 2-result.

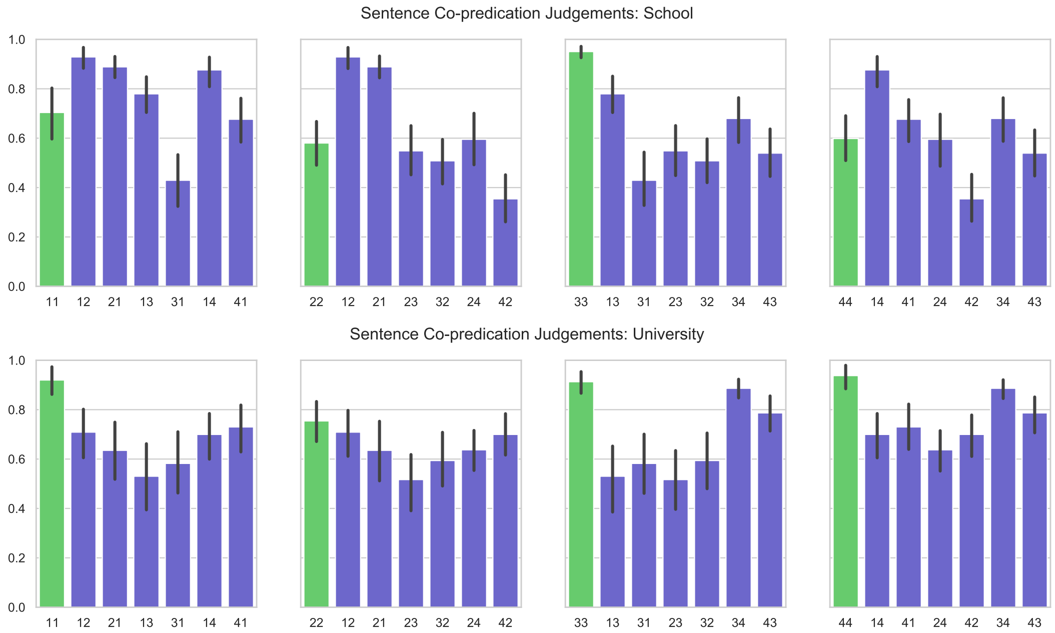


Figure 5.22: Mean acceptability ratings given to same-sense (green) and cross-sense (blue) co-predication structures eliciting different interpretations of targets *school*, and *university*. Senses: 1-building, 2-administration, 3-institution, 4-students.

cross-sense pairings than their explicit similarity ratings, with only *dinner* showing significant differences between the cross-sense readings and the *event* same-sense reference. This might indicate that participants were so accustomed to an underspecified reading of targets like *lunch* and *dinner* that in the co-predication setting they did not notice the different readings as much as in the explicit similarity comparison.

Door and *window* again show relatively clear differences between same-sense and cross-sense samples - with the exception of the ratings window's cross-sense sample 12, which received an acceptability score of 0.87, on par with the same-sense references. Sample 12 combines the scores for the following two items:

(74) 12a The window offered a great view of the nearby town centre and was made out of four equally large rectangular panes.

12b The window was made out of four equally large rectangular panes and offered a great view of the nearby town centre.

While the similarity scores for *CD* and *DVD* cross-sense samples were all relatively high, the co-predication acceptability scores for these targets show a clearer difference between same-sense and cross-sense means. Noteworthy here are the significantly lower acceptability ratings assigned to *physical*/*medium* alternations 12 and 21 for target *CD*, and the drop in acceptability for any sample including *DVD*'s

	Same-Sense			**Cross-Sense**		
Measure	**Pol.**	**Hom.**	**p**	**Pol.**	**Hom.**	**p**
Similarity	0.89	0.96	0.03	0.73	0.17	<0.05
Acceptability	0.83	0.86	0.10	0.64	0.41	<0.05
Word2Vec	0.60	0.65	0.12	0.55	0.58	0.06
ELMo	0.90	0.87	0.14	0.87	0.82	<0.05
BERT Base	0.91	0.93	0.22	0.88	0.78	<0.05
BERT Base (L4)	0.93	0.95	0.27	0.91	0.82	<0.05
BERT Large	0.79	0.85	0.15	0.72	0.44	<0.05
BERT Large (L4)	0.88	0.91	0.18	0.84	0.64	<0.05

Table 5.1: Word sense similarity distribution means for the different measures investigated in this study. Significance levels calculated by comparing the two underlying distributions through Mann-Whitney *U*.

content reading 3, which were not visible in the sense similarity judgements.

Book and *record* samples have been rated very similar in their co-predication acceptability as in their explicit sense similarity, with here the homonymic cross-sense samples invoking the *trophy* reading being assigned significantly higher acceptability than similarity ratings.

The acceptability scores for *school* and *university* finally don't fully match their similarity ratings, with a large part of *school*'s same sense samples receiving relatively low acceptability scores. Significant drops in similarity ratings are not necessarily mirrored by notable drops in acceptability, and *school* and *university* both show a different pattern in their assigned acceptability scores.

5.3.3 Computational Predictions

We extracted contextualised embeddings of target words using the models described in Section 5.2.2 and determined pairwise similarity scores by calculating the embeddings' cosine similarity (1-cosine). As samples were encoded individually, there are no potential order effects here. Figure 5.23 visualises the distribution of target embedding similarity scores, and the second part of Table 5.1 details their distribution means.

It is instantly noticeable that all contextualised models assign a much narrower range of similarity scores to the encoded ambiguous samples, suggesting that vector representations created by these language models only occupy a fraction of the embedding space (Ethayarajh, 2019). In the static baseline approach, the distribution of similarity scores assigned to homonymic and polysemic cross-sense samples does

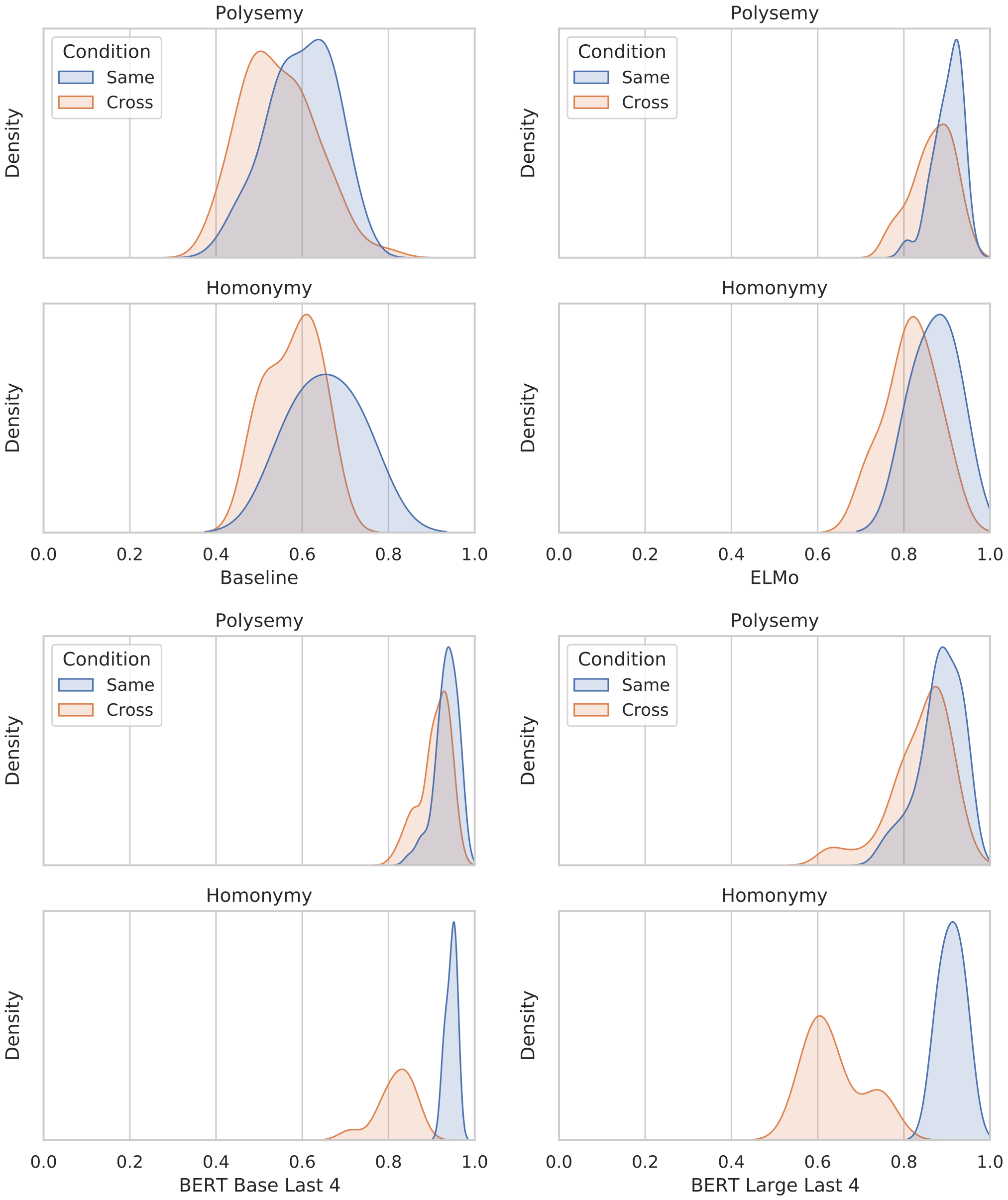


Figure 5.23: Distributions of embedding similarity scores obtained for same-sense (blue) and cross-sense (orange) samples with polysemic and homonymic alternations. BERT results for summing over the last four hidden states.

Combination		**Correlation**		**OLS Regression Analysis**			
First Measure	**Second Measure**	**r**	**p**	**Coef.**	**R^2**	**F-stat.**	**Prob.**
Similarity	Acceptability	0.698	1.09E-25	0.484	0.487	156.571	1.09E-25
Acceptability	Similarity	0.698	1.09E-25	1.005	0.487	156.571	1.09E-25
ELMo	Similarity	0.515	1.11E-12	2.863	0.265	59.475	1.11E-12
ELMo	Acceptability	0.523	4.39E-13	2.018	0.273	61.973	4.39E-13
BERT Base	Similarity	0.641	1.02E-20	4.070	0.411	115.185	1.02E-20
BERT Base	Acceptability	0.560	3.43E-15	2.469	0.314	75.521	3.43E-15
BERT Large	Similarity	0.687	1.22E-24	2.181	0.472	147.361	1.22E-24
BERT Large	Acceptability	0.550	1.40E-14	1.212	0.302	71.520	1.40E-14
Word2Vec	Similarity	0.206	0.008	0.675	0.042	7.309	0.008
Word2Vec	Acceptability	0.311	4.39E-05	0.707	0.097	17.625	4.39E-05

Table 5.2: Correlations between measures of contextualised word sense similarity. The first set of columns displays pairwise correlation based on Pearson's r, the second set shows the key statistics obtained from an OLS regression analysis. BERT results for summing over the last four hidden states.

not differ significantly (p-value = 0.06) - nor do the distribution of homonymic same-sense and cross-sense samples (p-value = 0.09). This indicates that the baseline does not capture any differences between homonymic and polysemic sense extensions - nor does it appear to be able to tell if a homonym is used in the same interpretation or referring to a completely unrelated concept. ELMo surprisingly also struggles with the same distinction (p-value = 0.09), but all BERT models produce clearly distinct distributions for cross-sense samples invoking different polysemic or homonymic sense extensions, as well as same-sense samples (all p-values <0.05). BERT Base and BERT Large finally produce very similar distributions, with BERT Large showing an overall larger difference between cosine similarity scores assigned to cross-sense and same-sense samples.

In order to establish a measure of correlation between the similarity scores predicted by the contextualised models and the collected human judgements, we next calculated their pairwise correlation scores (Pearson's r), and performed an ordinary least squares (OLS) regression for each combination of contextualised language model and human sense similarity measure. The results of these calculations are displayed in Table 4.4, and a selection of the pairwise comparisons is visualised in Figure 5.24. Since correlation and regression results for summing over the last four layers significantly outperformed those of using the last hidden layer state only, we here report only those scores, and refer to to the summed embedding when mentioning BERT Base or BERT Large encodings.

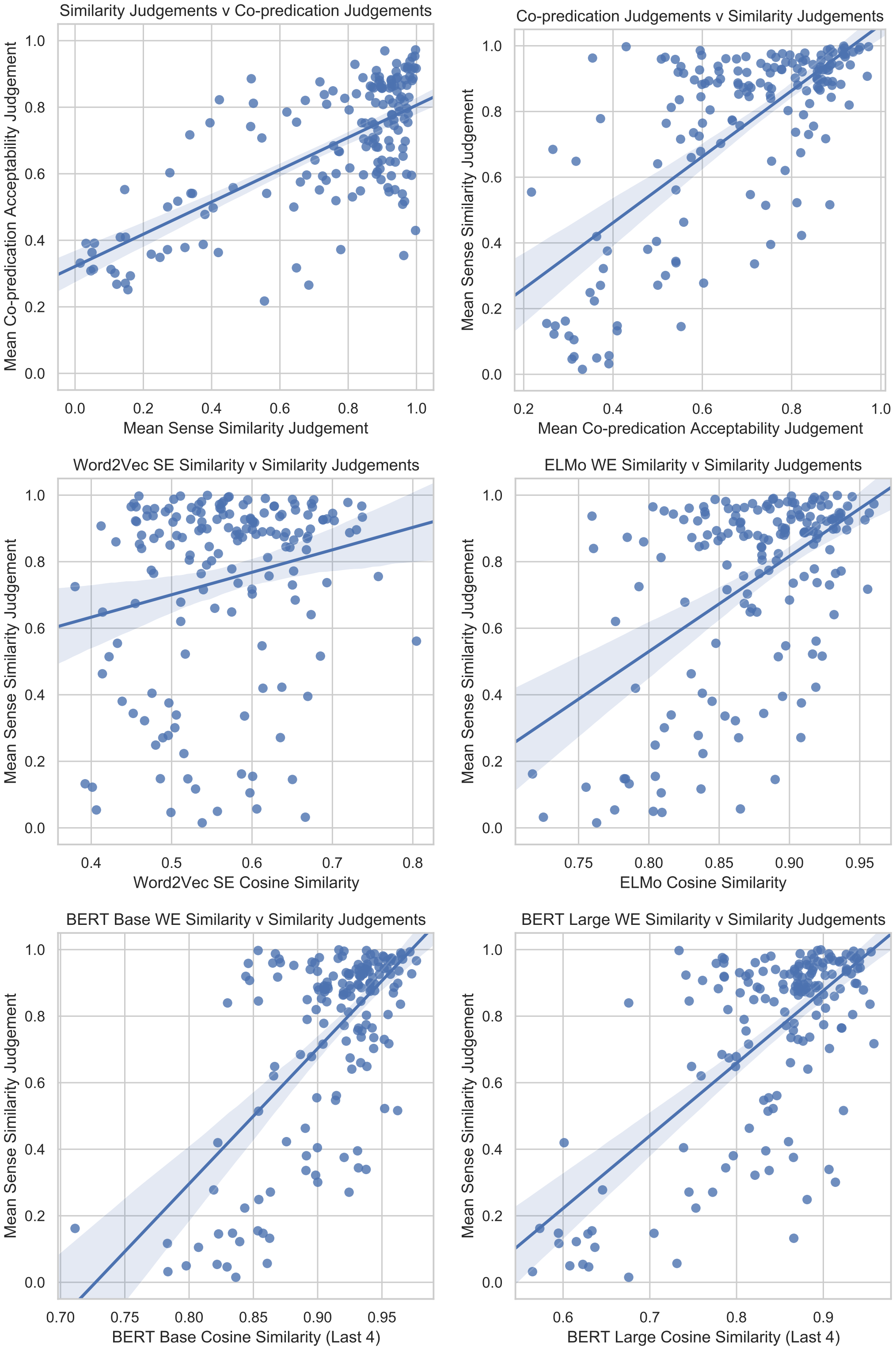


Figure 5.24: Correlations of the two human measures of word sense similarity (top row), and correlations of the different computational models with the explicit sense similarity ratings together with the best linear fit. Scaling of x-axis adjusted for clarity. BERT results for summing over the last four hidden states.

The static Word2Vec baseline displays a low but significant correlation with both human similarity measures (Person's r of 0.21 with similarity ratings and 0.31 with co-predication ratings), and shows an overall low goodness-of-fit, with R^2 values of the OLS regression at 4% and 10%, respectively. ELMo clearly outperforms this baseline, both in terms of correlation with the human measures, as well as in its goodness-of-fit in the OLS regression analysis. Its correlation with the explicit sense similarity judgements and co-predication acceptability ratings both calculate at 0.52. The BERT models finally perform on a similar level as ELMo when predicting co-predication acceptability (0.56 for BERT Base and 0.55 for BERT Large), but BERT Large is clearly the best-performing model when predicting explicit similarity scores, with a correlation of 0.69 to the human annotation, and an R^2 goodness-of-fit of 47%.

Qualitative Analysis

Figure 5.24 visualises the correlations between a selection of collected similarity measures. The top row shows a comparison between our annotators' explicit word sense similarity judgements and the co-predication acceptability ratings, as well as vice versa. With one measure on each axis, the data points' locations indicate the mean rating of each ambiguous test item with respect to those two measures. What is immediately visible is the clustering of samples towards the upper end of either rating scale, gathering same-sense and near-identity cross-sense samples. When comparing this cluster between the two metrics, it seems that co-predication acceptability spreads these ratings wider than the explicit sense similarity scores do, given that the samples that fall into the similarity score's 0.8-1 range have obtained acceptability scores ranging between as low as 0.4 and 1. Explicit similarity ratings on the other hand use the entire scale, while no co-predication acceptability mean came in lower than 0.2. The next four figures each compare the explicit word sense similarity judgements with one of the computational methods. The comparison with the Word2Vec baseline here clearly shows the low correlation between the two measures, while the other three figures roughly exhibit the same patterns as observed in comparing the two human annotations - when scaling the x-axis to make up for the fact that contextualised embedding similarities all are highly similar to each other. Like acceptability scores, ELMo's embedding similarity spreads samples that received high similarity ratings item relatively wide, while the BERT models keep those samples very closely grouped. BERT Base here exhibits an even tighter grouping than BERT Large - which however makes it more difficult to find a good best fit, leading to a higher goodness-of-fit for BERT Large.

Rather than discussing the details of the similarity scores given to each target

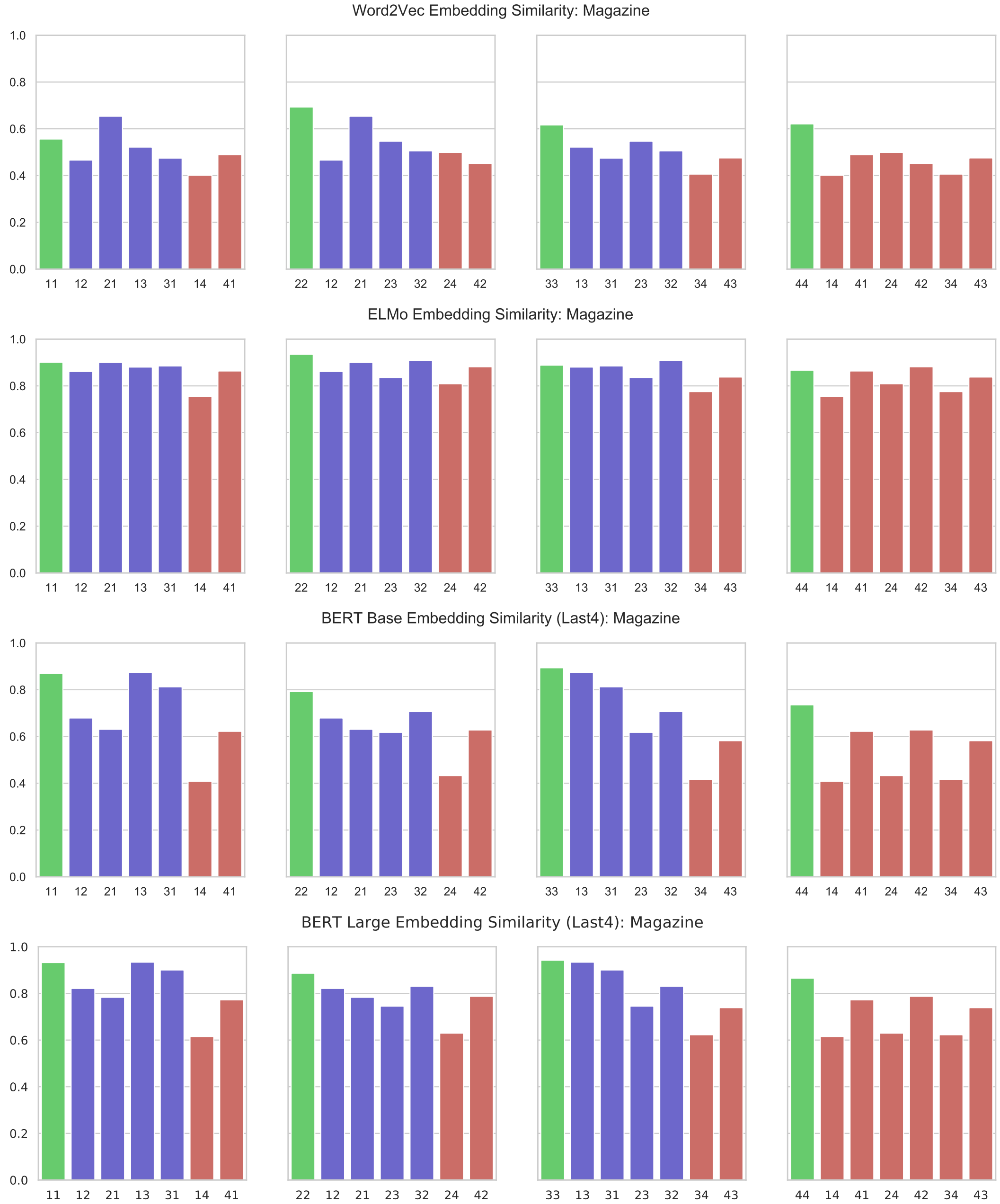


Figure 5.25: Contextualised embedding similarities of the different sample combinations for *magazine* as predicted by the four different computational models. From top to bottom: Word2Vec Baseline, ELMo, BERT Base (last 4 layers) and BERT Large (last 4 layers). Senses: 1-organisation, 2-physical, 3-content, 4-storage type.

word by each model, we will here only briefly mention some observations based on target *magazine* before moving on to an investigation of similarity patterns in each of the similarity measures in the next section. Figure 5.25 shows the calculated embedding similarities for different sample combinations for *magazine* as predicted by the four different computational models. The similarity scores based on the Word2Vec baseline correctly show relatively high marks for same-sense samples, and are clearly lower for at least the homonymic cross-sense alternations. In this case, most of the polysemic cross-sense samples also receive lower similarity ratings than the same-sense references, but do not show the drop in scores for *organisation/physical* samples `12` and `21` that we observed in both human ratings. ELMo assigns very similar scores to all *magazine* samples, making it difficult to tell apart same-sense samples from either polysemic or homonymic cross-sense samples. This is in line with the previous observation that ELMo's overall distributions of same-sense and cross-sense sample ratings do not significantly differ from each other. BERT Base and BERT Large finally again show very similar patterns in their ratings. While BERT Base spreads ratings on a wider part of the scale, the relative differences between ratings are slightly larger for BERT Large. Both rate polysemic cross-sense samples lower (or in one case as high as) the same-sense references, and homonymic cross-sense samples receive clearly lower ratings than any of the other comparisons. BERT Base and BERT Large rate *organisation/physical* samples `12` and `21` lower than the *organisation/content* alternations `13` and `31` - much like the annotators did - but also predict that the comparison of *physical/information* combinations elicits a clearly lower similarity score than the *organisation/content* combination.

5.3.4 Similarity Patterns

The full annotated dataset allows for an investigation not previously possible with the pilot data: that of similarity patterns within and across different types of polysemic alternations. Now the collected data contains annotations for at least two target words for the ten different types of regular metonymic alternations tested in this study, we can inspect whether there are regularities between them - and how well each of the different similarity measures captures these patterns. To inspect these potential patterns, we established a set of similarity maps containing the mean similarity ratings for each combination of senses a given target word can take on, and compared these between targets of the same type. As an example, Figure 5.26 displays the similarity maps for target words *newspaper* and *magazine*. This way of interpreting the data simplifies the previously shown bar charts into a single graph, reducing nuance but improving the intuitiveness of the plots. The brighter a field

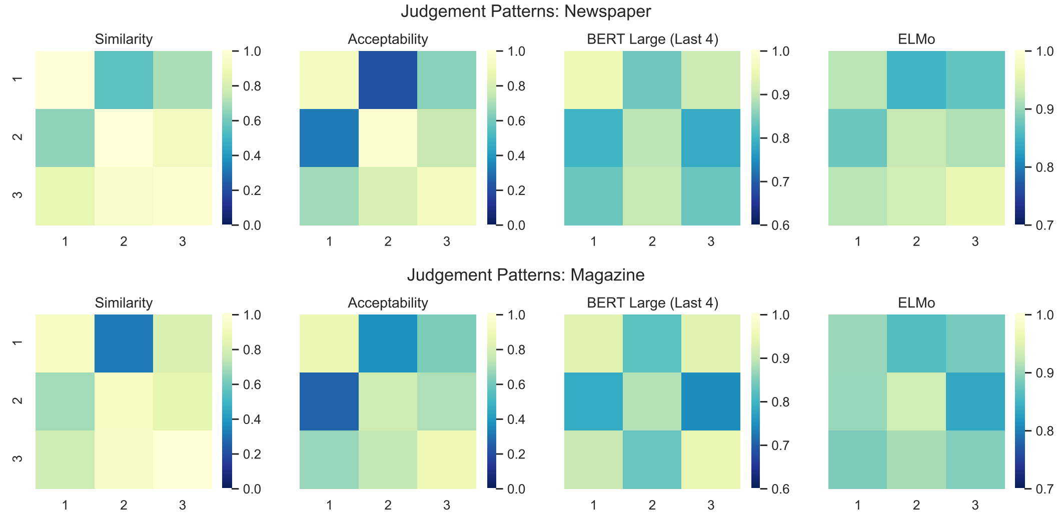


Figure 5.26: Similarity patterns in the sense similarity ratings for polysemes *newspaper* and *magazine*. Senses: 1-physical, 2-information, 3-organisation. Colour scales adjusted for computational measures.

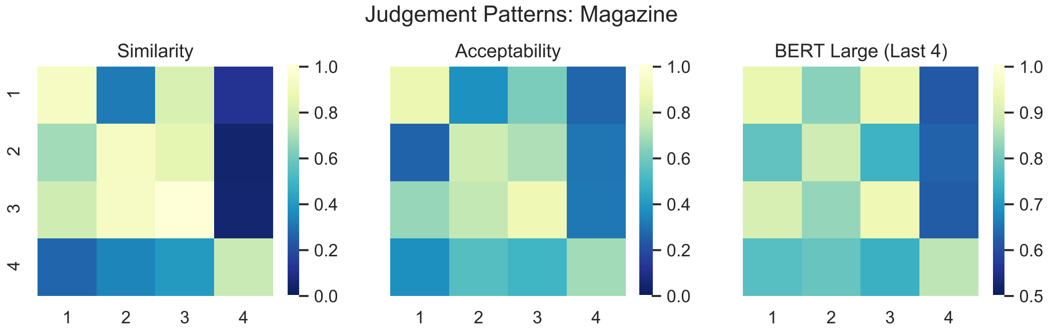


Figure 5.27: Similarity heat map for all tested interpretations of *magazine*, including homonymic alternation 4-storage type. Senses: 1-physical, 2-information, 3-organisation. Colour scales adjusted for computational measures.

in the similarity map, the higher is the similarity or acceptability score given to the indicated combination of senses. On a first glance, this visualisation clearly shows the drop in ratings for *organisation/physical* samples 12 and 21 observed earlier, both showing up in dark blue. On the other hand, the downwards diagonal should always be relatively bright, as it represents the same-sense samples that should receive ratings of perfect similarity/acceptability. Note that these similarity maps only include senses common to all targets, in this case common to *newspaper* and *magazine*. Figure 5.27 contains the full similarity map calculated for *magazine*, which now also includes the homonymic interpretation as a type of storage. These plots clearly show the low ratings given to its homonymic cross-sense samples in each of the different measures.

	Pairwise		**Overall**	
Measure	***r***	**p <0.05**	***r***	**p**
Similarity	0.44	3/24 (12.5%)	0.53	8.260e-10
Acceptability	0.44	4/24 (16.7%)	0.62	5.306e-14
ELMo	0.14	0/24 (0%)	0.21	0.025
BERT Large	0.28	1/24 (4.2%)	0.27	0.003

Table 5.3: Mean Pearson correlation of polysemic word sense similarity patterns across different target words allowing the same alternation of senses, number of significant comparisons, and overall pattern correlation.

Returning to the similarity maps with senses common to both *newspaper* and *magazine*, the correlation between these similarity maps reaches 0.89 (p-value $<$ 0.05) when comparing the human annotations of explicit sense similarity, and even 0.95 (p-value $<$ 0.05) based on the co-predication acceptability ratings - which we take to indicate a clear pattern in the target words' similarity ratings, and thus a likely regularity in all targets allowing for this alternation. In the similarity maps based on the cosines between BERT Large embeddings, the correlation between *newspaper* and *magazine* reaches only 0.65 (p-value = 0.06), and just 0.34 (p-value = 0.37) in the ELMo similarity maps.

The overall similarity pattern correlations across target words of the same type of polysemic alternation can be found in Table 5.3. The first set of scores are based on the correlations of all pairwise comparisons of polysemes that allow for the same alternation. Both human annotations show a correlation of 0.44 (Pearson's r) between similarity patterns of targets of the same type, with 3 pairings reaching significance based on explicit sense similarity judgements, and 4 based on co-predication acceptability judgements. The average pairwise pattern correlation for ELMo was calculated at only 0.14, with no significant matches, and BERT Large cosine similarity produced an average pairwise correlation of 0.28 with one significant pairing. The number of significant comparisons reported here however is very likely to be misrepresent the actual data due to the small number of senses tested. We therefore also calculated a second score by appending all pairwise comparisons into two separate lists and determining the correlation between these two lists. This approach should return a better estimate of the overall pattern consistency for a given measure, but might under-represent inconsistent patterns. Based on this alternative calculation, co-predication acceptability shows the highest pattern consistency with an overall correlation of 0.62 between the scores assigned to different targets of the same type, followed by explicit similarity ratings with an overall correlation of 0.53.

ELMo's correlation score slightly increased to 0.21, while BERT Large's correlation score remained largely unchanged at 0.28.

The lower correlation scores for the two computational approaches seem to indicate that sense similarity patterns are not represented as consistently in these models as in the human annotations. To quantify this intuition, we also compared the similarity maps produced by the human annotations with those produced by the contextualised language models. The mean correlation between BERT Large's similarity maps and the explicit word sense similarity maps here is 0.49, with one significantly similar pairing, and 0.52 compared to co-predication similarity maps (with 4 significant pairings) - rates surprisingly comparable to the correlation between the two human annotations (mean $r = 0.54$, 10 comparisons with $p<0.05$).

Qualitative Analysis

Figure 5.28 shows the similarity maps for the tested *animal/meat* alternation targets *chicken*, *lamb*, *pheasant* and *seagull.* Both *chicken* and *lamb* again are considered common variants, while *pheasant* is less frequent and *seagull* would typically not be considered a member of this type. For all targets, co-predication acceptability judgements show a clear distinction between same-sense (`11` and `22`) and cross-sense samples (`12` and `21`), while this distinction is less clearly apparent in the other three measures for the experimental targets *pheasant* and *seagull.* This mirrors previous observations especially concerning the surprisingly consistently high similarity ratings for *seagull*, but now also indicates one of the aspects contributing to BERT Large better correlating with the explicit similarity scores than the co-predication acceptability ratings. Overall, we take these similarity maps as strong evidence for a consistent similarity pattern in at least the more prototypical representatives of the frequently discussed *animal/meat* alternation.

Figure 5.29 displays the same set of similarity maps for the *content-for-container* alternation. Here the distinction between same-sense and cross-sense samples is much less pronounced in all of the measures, with experimental target *juice* showing the clearest sign of their distinction. Overall it thus seems that neither annotators nor language models consider the *content* and *container* reading notably dissimilar, supporting an under-specified representation of this alternation.

The *container-for-content* alternation in Figure 5.30 on the other hand again shows a relatively clear pattern in at least the explicit similarity ratings for its same-sense and cross-sense readings, indicating that both readings are available, but are not considered similar. Co-predication acceptability here seems more sensitive to the use of the *container-for-content* alternation, with *content/content* sample

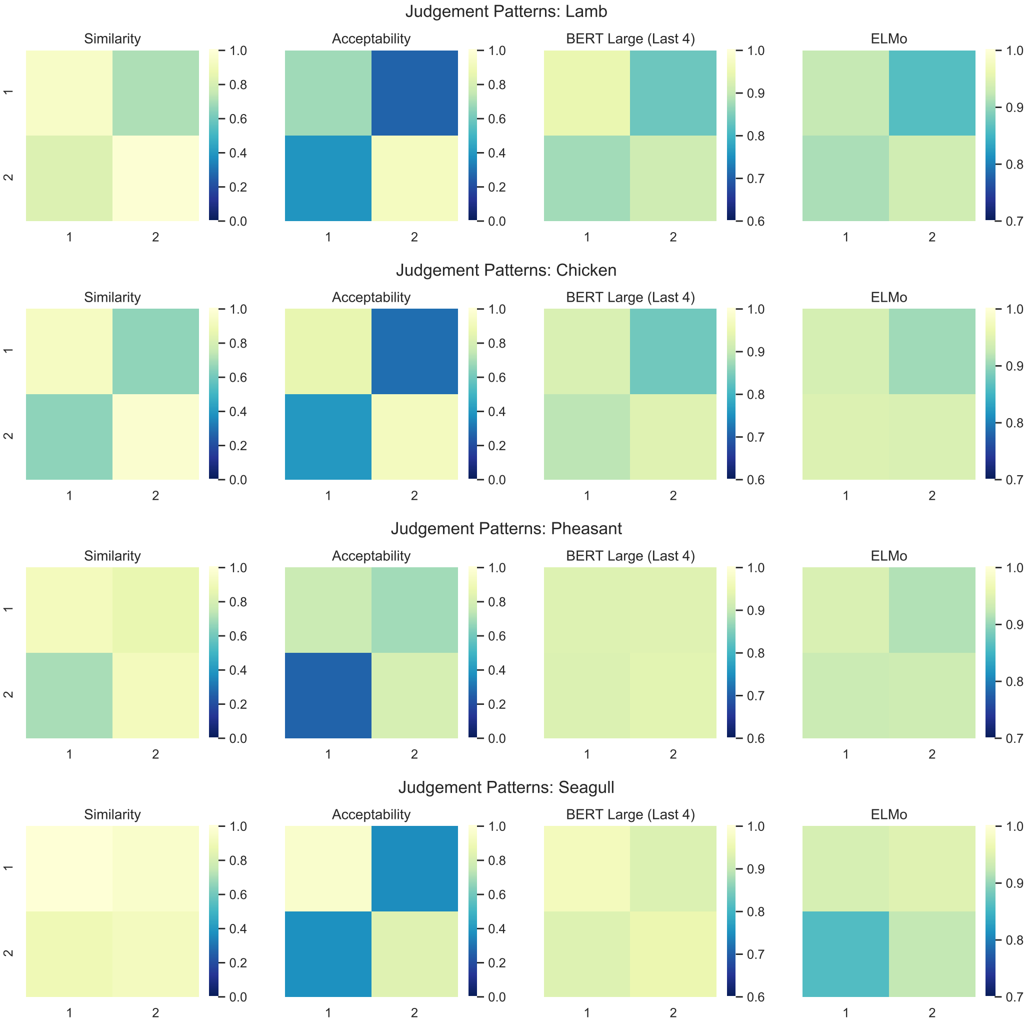


Figure 5.28: Similarity patterns in the sense similarity ratings for *animal/meat* alternation polysemes. Senses: 1-animal, 2-meat. Colour-scales adjusted for computational measures.

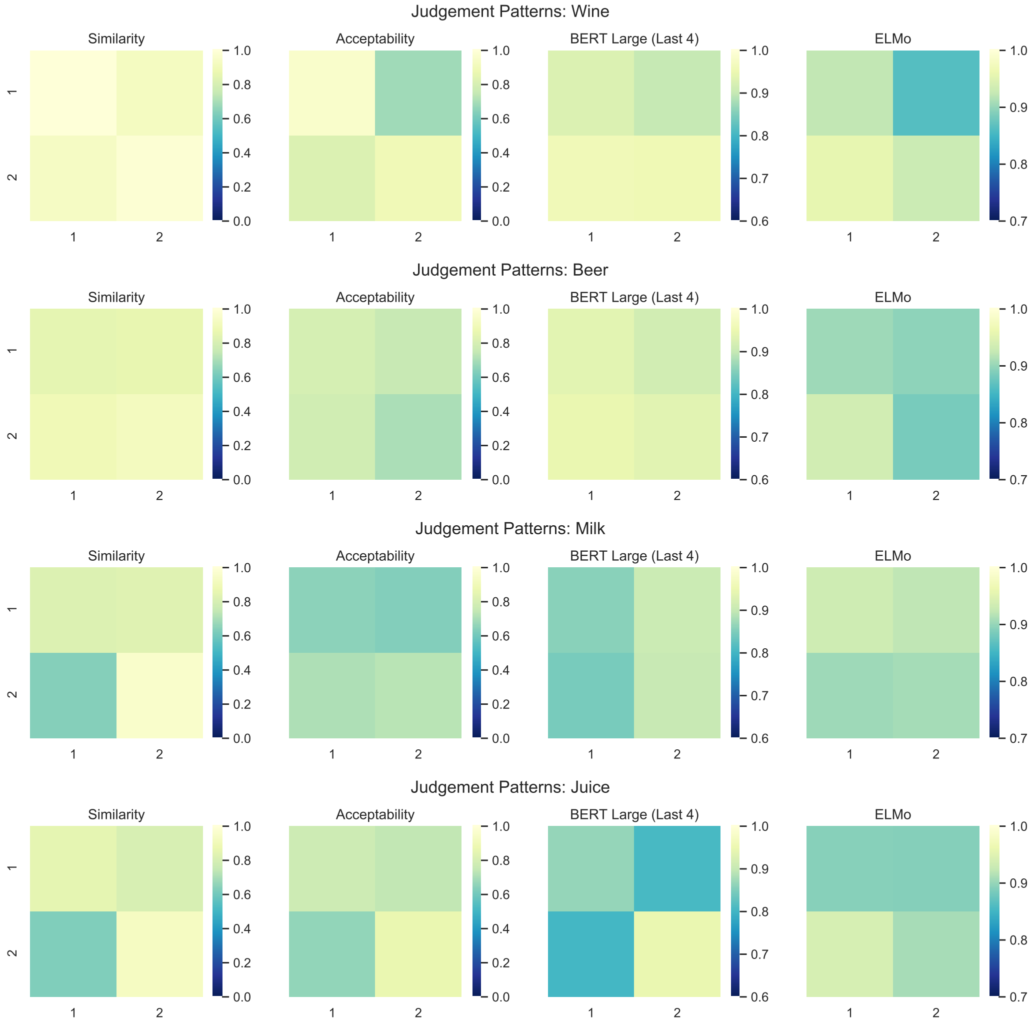


Figure 5.29: Similarity patterns in the sense similarity ratings for *content-for-container* alternation polysemes. Senses: 1-content, 2-container. Colour-scales adjusted for computational measures.

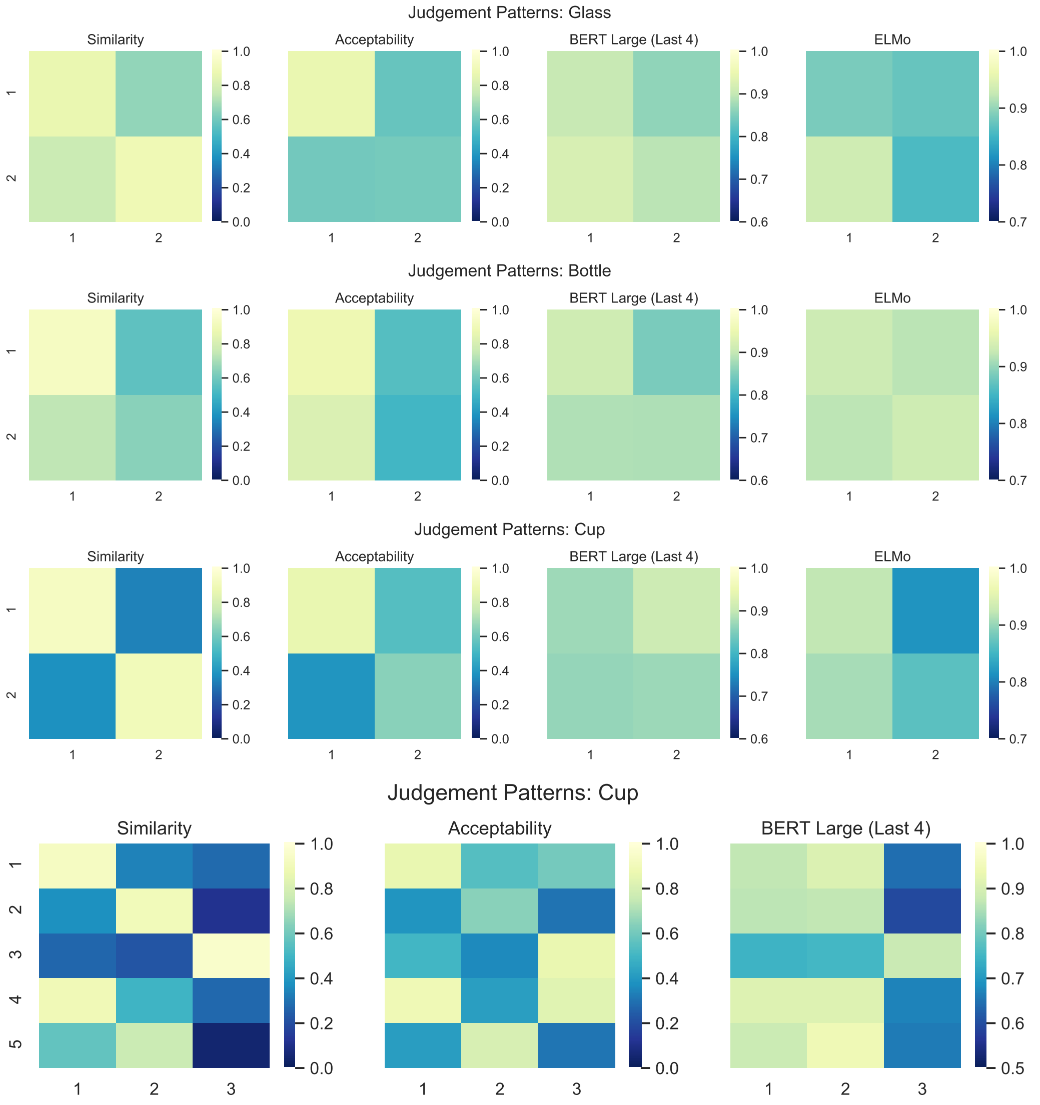


Figure 5.30: Similarity patterns in the sense similarity ratings for *container-for-content* alternation polysemes, with the last row showing all tested alternations of *cup*, including homonymic readings. Polysemic senses: 1-container, 2-content. Colour-scales adjusted for computational measures.

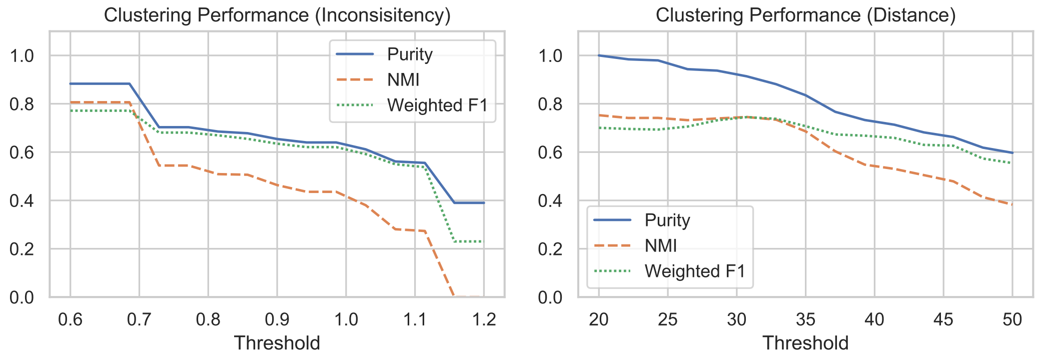


Figure 5.31: Clustering performance for the inconsistency (left) and distance (right) criterion when grouping BERT Large contextualised embeddings with linear Ward clustering based on clustering threshold t.

22 receiving lower acceptability scores than the container reference in 11. Neither ELMo nor BERT seem to produce consistent ratings for the different targets, nor to the patterns show a high correlation with the human annotations. From among the targets, experimental *cup* shows the largest differences in same-sense and cross-sense samples, indicating that for this alternation the difference in sense interpretation might be more pronounced for less prototypical targets than it is in frequently used ones. All remaining similarity maps can be found in Appendix B.2.

5.3.5 Word Sense Clustering

As BERT Large displays a high correlation with the human judgements of word sense similarity and some capability in replicating similarity patterns across target words, we finally wanted to investigate how well BERT's contextualised embeddings could be used to cluster samples of polysemous word uses (also see Del Tredici and Bel, 2015; McCarthy et al., 2016; Garí Soler and Apidianaki, 2021). To provide a tentative analysis, for each target word we clustered BERT Large's contextualised target encodings based on their similarity using the hierarchical Ward clustering method implemented in SciPy.[8] We opted for hierarchical clustering as this method has to determine the optimal number of clusters itself, and does not take this number as an argument like most other clustering approaches do. We experimented with two different clustering criteria based only on a threshold parameter t: using *node inconsistency*, all leaf descendants of a cluster node belong to the same cluster if that node and all these descendants have an inconsistent value less than or equal to a threshold value t. Under the *distance* criterion, clusters are formed so that the

[8] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.fcluster.html`

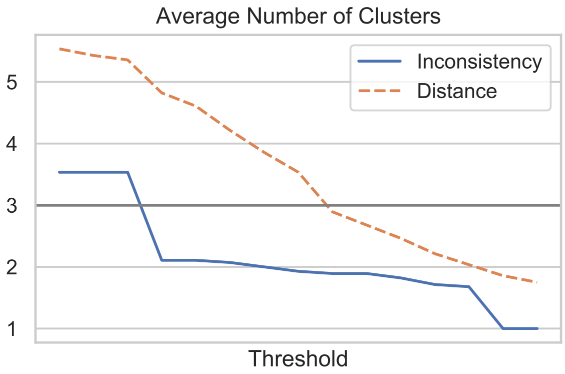


Figure 5.32: Average number of clusters produced by the clustering methods (gold mean: 3.0).

Criterion	***t***	**#C**	**NMI**	**F1**	**P**	**R**
Inconsistency	<0.7	3.54	0.60	0.77	0.86	0.71
Distance	31	4.21	0.75	0.75	0.90	0.64

Table 5.4: Best-performing settings for inconsistency and distance-based hierarchical Ward clustering of target word senses. #C is the average number of clusters produced per target.

observations in each cluster have no greater distance than the set threshold value t.

Figure 5.31 shows the development of cluster purity, Normalised Mutual Information (NMI) and weighted F1 scores for different values of threshold t using the inconsistency criterion (left) and distance criterion (right). Figure 5.32 plots the average number of clusters produced by both measures with increasing threshold t (gold mean: 3.0). The quantitatively best-performing settings are displayed in Table 5.4. Both settings produce more clusters than the traditional grouping of the tested targets would assume, which indicates that especially precision scores might be artificially high - but overall the clustering seems to produce sensible results with F1 scores of 0.77 and 0.75, respectively.

Qualitative Analysis

Figure 5.33 shows dendograms for the relative distances between different sense interpretations of *newspaper* and *magazine* based on the hierarchical Ward clustering applied in the previous section. The grouping of *newspaper* interpretations clearly separates the *organisation* sense 1 from the *physical* interpretation 2, but splits the *information* samples 3 among the two, indicating the similarity in their contextualised embeddings. For *magazine*, the clustering of samples creates four distinct

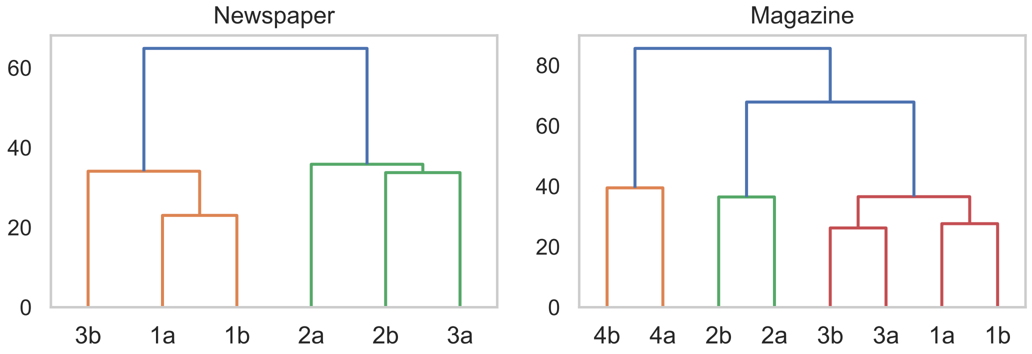


Figure 5.33: Dendograms of BERT Large contextualised embedding similarity for targets *newspaper* and *magazine* based on hierarchical Ward clustering. Numbers indicate expected sense distinctions. Senses: 1-physical, 2-information, 3-organisation, 4-storage.

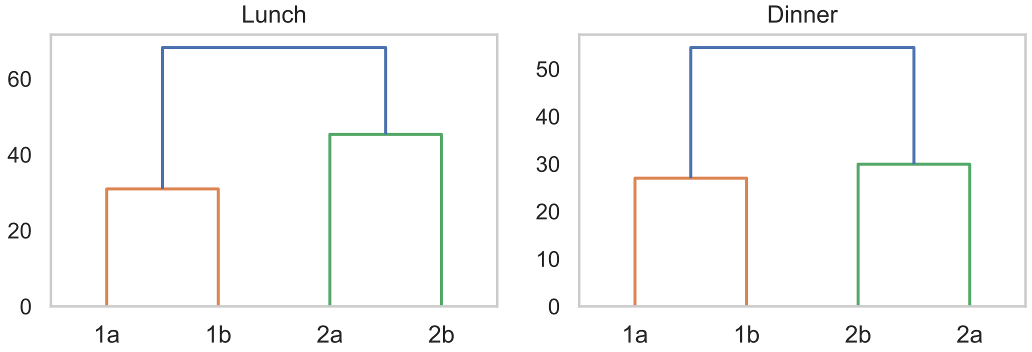


Figure 5.34: Dendograms of BERT Large contextualised embedding similarity for targets *lunch* and *dinner* based on hierarchical Ward clustering. Numbers indicate expected sense distinctions. Senses: 1-event, 2-food.

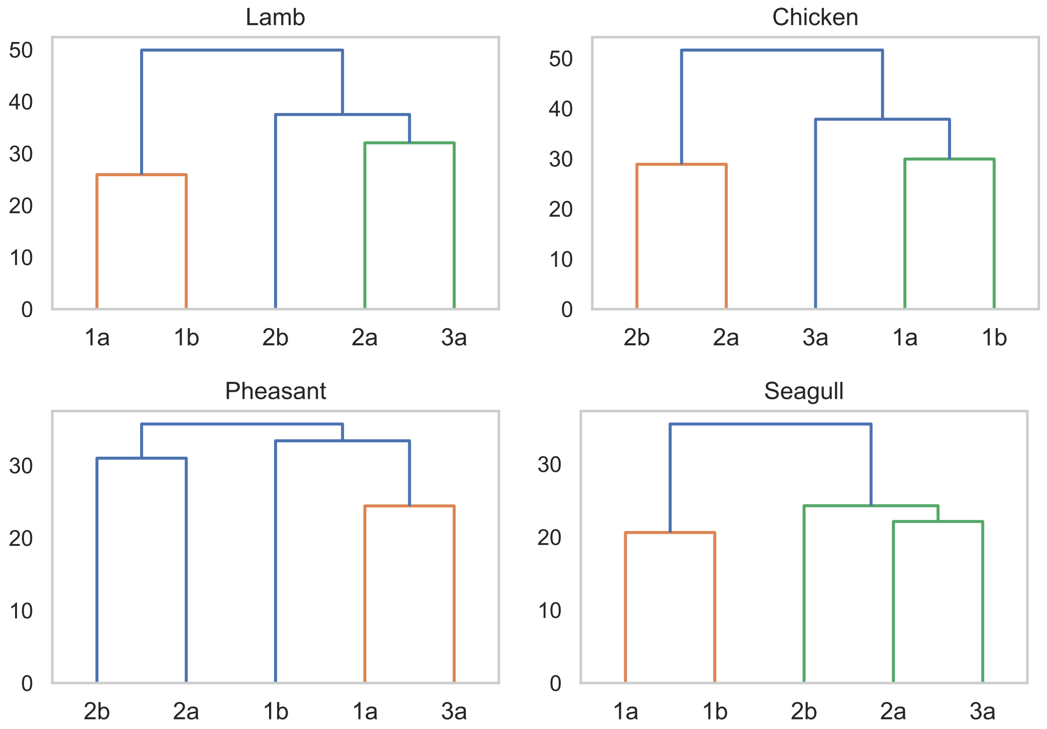


Figure 5.35: Dendograms of BERT Large contextualised embedding similarity for targets *lamb*, *chicken*, *pheasant* and *seagull*. Senses: 1-animal, 2-meat, 3-animal⊕.

groupings and clearly separates the three polysemic senses from the homonymic *storage* reading 4.

The clustering of targets exhibiting an *food/event* alternation (Figure 5.34) or *process/result* alternation seem to work similarly well, with the dendograms of our *animal/meat* targets with experimental *animal*⊕ samples indicating that these samples intended to facilitate sense shifting indeed end up being clustered with either sense interpretation equally often. Other alternations consistently lead to a misclustering of sense extensions, including the *content-for-container* alternation, for which dendograms based on the hierarchical sense clustering are shown in Figure 5.36. All remaining dendogram visualisations can be found in Appendix B.3.

5.4 Discussion

In this chapter we presented the collection and analysis of a novel, human-annotated dataset of word sense similarity. The dataset contains explicit word sense similarity ratings and co-predication acceptability ratings for 28 polysemic targets exhibiting ten different types of alternations. Targets are either seminal examples of a given type of alternation, or represent less prototypical variants that allow an investigation the extend of the regularity of selected type.

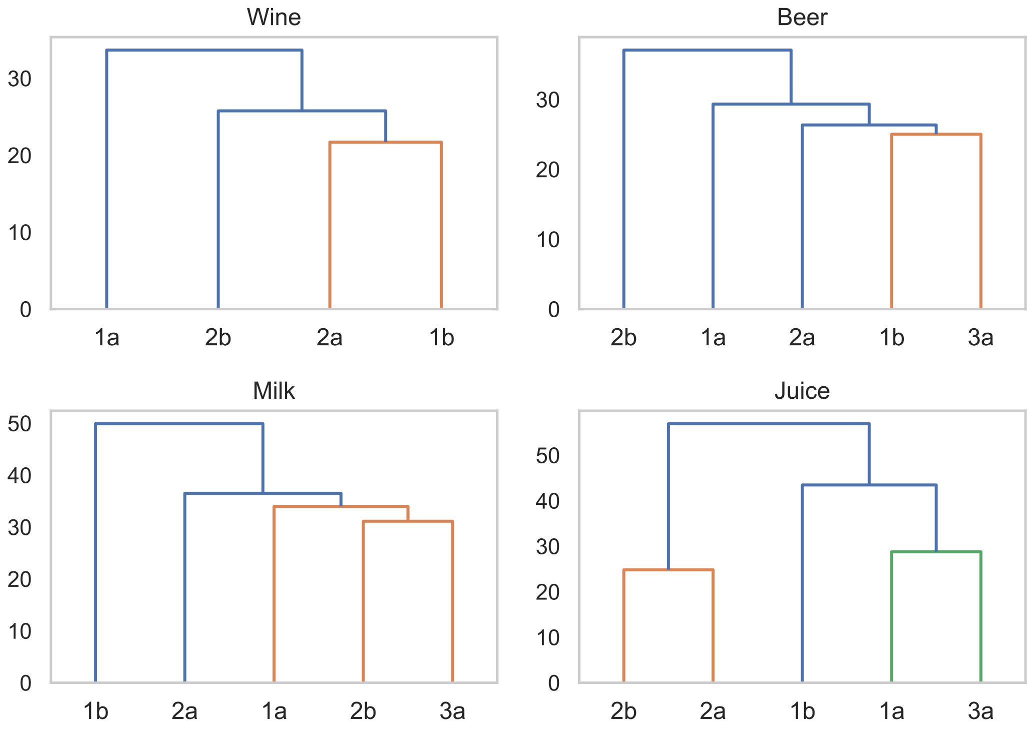


Figure 5.36: Dendograms of BERT Large contextualised embedding similarity for targets *wine*, *beer*, *milk* and *juice* based on hierarchical Ward clustering. Senses: 1-container, 2-content, 3-container⊖.

We collected close to 18,000 annotations through crowd-sourcing on Amazon Mechanical Turk, finding that annotators make use of the provided graded rating scale in judging the explicit similarity as well as the co-predication acceptability of sample pairs. Both ratings show significant differences in the overall means of same-sense and cross-sense judgements, indicating that polysemous extensions are not always perceived as invoking identical interpretations, but also contain cases of perceived identity of sense. Investigating judgements for individual target words, we observed significant drops in the similarity or acceptability ratings for a select number of sample combinations, showing that those samples specifically invoke sense combinations that are clearly distinguished by the annotators. In some cases, these polysemic cross-sense samples can obtain ratings as low as the homonymic test items included in the study.

The collected data provides intriguing empirical evidence challenging traditional models of mental word sense representation. We suggest that observations of significant similarity differences between different polysemic senses are difficult to explain when assuming a fully under-specified mental representation of polysemic sense: if all interpretations were to be stored in the same, unstructured entry, we would not expect participants to clearly distinguish their interpretations. Because all of the senses stored in an under-specified entry should allow for cost-free sense switching and be co-activated, finding evidence of perceived differences in meaning indicates that the mental representations of these senses are likely more structured than assumed by one representation models. On the other hand, some cross-sense polyseme readings do receive similarity ratings and co-predication acceptability ratings close to those or exceeding those of same-sense items. This observation suggests that in the processing of some polysemic senses, no distinction is made in their interpretation - even though the invoked senses are not identical. While this is in line with the assumptions of one representation models, it is a challenging finding for sense enumeration approaches, which in these cases will struggle to specify the necessary contrast and selection criteria to warrant separate entries for the invoked senses.

The collected data however fits in well with recent proposals of a more structured mental representation of polysemic sense (see e.g. Ortega-Andrés and Vicente, 2019), where word sense distance could be an underlying factor in determining the similarity of sense interpretations and their co-activation. Assuming that the collected word sense similarity and co-predication acceptability ratings are a proxy of the senses' distances in their mental representation, our data supports the potential of a distanced-based grouping of senses within an otherwise still under-specified entry. A model like this would allow for the co-activation of just a subset of sense interpretations (those that are grouped closest together), which allows for the observation of near identity ratings in their comparison, as well as significant differences in the interpretation of those senses that are represented in different groups or clusters - leading to observations of similarity ratings as low as those for homonymic controls.

Because the full dataset contains at least two target words for each type of alternation, it also allowed us to investigate potential patterns in the similarity and acceptability ratings collected. We found that for some alternations the differences in sense interpretations are relatively consistent across targets of the same type, while for others these patterns could not be established. We suggest that these observations even more underline the heterogeneity within different phenomena of polysemy, as not even all types of regular metonymic polysemy seem to exhibit consistent patterns in their sense similarity judgements. The data however also suggests that not every polysemic word allows for its own, idiosyncratic set of sense extensions, but that there is some potential for a classification or grouping of polysemic expressions.

A second set of analyses focused on evaluating the performance of the default implementations of contextualised language models such as ELMo and BERT in predicting the human annotations. We extracted contextualised word embeddings for target words within their respective context sentences, and considered the cosine similarity as a measure of their similarity. Especially the similarity scores calculated with BERT Large exhibit a good correlation with the collected human judgements of explicit word sense similarity, but does not consistently predict the same similarity patterns as observed in the annotations. Still, when using BERT Large's contextualised target embeddings to group context samples by the senses they invoke, a hierarchical clustering returns the expected grouping of senses in almost half of the ten polysemic alternations tested in this study. Together, we suggest that these observations indicate a promising potential of BERT Large providing comparatively cheap indications of nuanced word sense similarities in future work, especially so if optimised or fine-tuned for this task. As fine-tuning large contextualised models like BERT however requires a much larger annotated dataset of graded word sense similarity, in the next Chapter we will explore using corpus data to bootstrap the development of such a large-scale expansion of the dataset presented here.

Chapter 6

Polysemy Detection with Contextualised Language Models

With especially BERT Large exhibiting a promising correlation with human judgements on the graded similarity between different uses of an ambiguous expression, we started wondering whether contextualised language models like BERT could be used as an alternative to costly human judgements. Obtaining representative automatic annotations of word sense similarity through these tools could allow for an investigation of lexical ambiguity and specifically polysemy on a much larger scale than would be possible otherwise - and provide resources large enough to benefit the computational linguistics community in training and evaluating models.

This chapter's primary focus is our last research question

Q2a Can contextualised word embeddings be used to classify and identify (new) polysemic targets?

Investigating the potential of BERT Large particularly, in we will present a pilot algorithm providing a largely unsupervised heuristics for identifying words that could potentially allow for the same sense alternations as the targets tested in the studies presented earlier. A method like this could be used to explore the productivity of different types of polysemic alternations, and can provide interesting distributional and corpus-based insights into the real-world usage of polysemous expressions. For a selection of potential targets, a preliminary analysis of the results produced by the pilot algorithm will be presented in Section 6.2.4.

6.1 Target Identification

We propose that establishing a corpus of polysemous target words which allow for the same set of alternations as a given reference set can best be approached in two steps: in a first step corpus-based evidence is used to pinpoint potential target words that could allow for the same alternations as a given reference word, and in a second step an investigation of corpus samples containing the proposed new target word is used as an indication of the new target's polysemic potential with respect to the alternation observed in the referent.

In our pilot, we utilised BERT's masked token prediction pre-training objective to aid us in automatically identifying target words that potentially allow for the same set of (polysemic) alternations as a reference word. To do so, we first collected a substantial amount of corpus samples containing a reference word from a plain, un-annotated corpus of natural language use. For our pilot, we used the Wikipedia dataset[1], which we assumed would provide us with relatively formal, clear and well-formed sample sentences. Where possible, we collected up to 2,048 sentences from the training section of the Wikipedia corpus[2] that contained a given reference word from our original set of 28 targets. Samples were only added if the word type of the reference word in the corpus sentence matched the word type of the referent in our experimental samples, which in all cases were nouns (i.e. the sample target had to be tagged as NN or NNS by TextBlob's[3] PatternTagger), and if the corpus sentence was no longer than 25 words. As a result of this collection procedure, of the 28 reference words only *duck* did not return the full set of 2,048 corpus sample sentences, but 1,618.[4]

With a representative number of corpus samples collected for each reference word, we next blanked out the reference words from the corpus samples, and used BERT's masked token prediction pipeline to predict the top 10 most likely tokens to fill the blank in each corpus sample. Collating predictions across all sample sentences for a given reference word, we then established the top 20 tokens that were included in the individual top 10 predictions most often, and filtered out all non-word and sub-word tokens to retain at least seven potential substitutes per reference word. In almost all cases the top predicted word equalled the actual reference word that was blanked out in the samples, with exception of *lamb* where the top prediction was *chicken*, *pheasant* where the top prediction was *deer*, and *seagull*, where the top

[1] `https://huggingface.co/datasets/wikipedia`

[2] The Wikipedia Training section contains a total of 6,078,422 documents

[3] `https://textblob.readthedocs.io/en/dev/quickstart.html#part-of-speech-tagging`

[4] See Section 6.3 for a discussion on potential improvements to the pilot procedure.

Substitute	**Occurrence**
fish	26.5%
meat	22.4%
pork	21.4%
rice	19.3%
beef	17.0%
pig	11.6%
food	10.8%
duck	10.2%
bird	10.0%

Table 6.1: Most likely substitutes for reference word *chicken* in Wikipedia corpus samples as predicted by BERT's masked token prediction. Percentages indicate the occurrence of the substitute in the top 10 predictions per sample.

prediction was *bird*.

As an example, Table 6.1 contains the most likely substitutes for reference word *chicken* in the collected set of Wikipedia corpus samples as predicted by BERT. Percentages indicate the occurrence of the substitute in the top 10 predictions per sample, so *fish* for example was included in the top 10 predictions of 26.5% of the corpus samples. Upon preliminary inspection, we expect *fish* and *duck*, and potentially *beef* and to some extend *bird* to allow for polysemic alternations in the same dimensions as the original target *chicken*, that is allowing for an *animal* reading as well as a *food* reading. *Meat*, *pork*, *rice* and *food* itself should only allow for a *food* reading and not for an *animal* reading, while *pig* should only allow for an *animal* reading (since *pork* is the *food* extension of *pig*). Our goal now is to develop an unsupervised procedure to assess the potential targets and for each decide whether they indeed allow for the same set of alternations as the reference word *chicken* does.

6.2 Target Evaluation

We considered two different approaches to scoring the potential substitutes predicted by BERT: either by inspecting the perplexity of replacing original targets with the substitutes in the collected corpus samples as a means of assessing the felicity of replacing the target, or by analysing the distribution of corpus samples containing the proposed substitute itself. As we soon realised that the first approach did not yield much informative data, we will focus mostly on the second approach.

6.2.1 Substitution-based Sample Scoring

As a first attempt at evaluating potential substitutes, we considered applying traditional scoring methods such as BLEU, ROGUE or Earth Mover Distance. Given that we had already collected a reference set of corpus sentences containing our original targets, we could directly apply these methods to score corpus samples after replacing the original target word with a given substitute to rate the acceptability of its substitution. The reasoning behind this approach is based on the assumption that if a given substitute indeed allows for the same set of alternations as the original target, most if not all of the corpus samples should still be acceptable after replacing the target word. If a given substitute however allows for only one or none of the original sense alternations, a respective portion of the corpus samples should be rendered infelicitous after replacing the original target with the substitute.

Since traditional methods like BLEU, ROGUE or Earth Mover Distance only consider the amount of words or characters that need to be altered to turn a predicted sentence into the reference sentence, all of these methods would simply flag the substitute word as deviation from the gold reference - without having the capability of assessing the actual acceptability or correctness of the substitute sentence. Any approach comparing substitute and reference sentences on a token or character level rather than including some representation of word meaning therefore would fall short in evaluating potentially polysemic targets and could not be used in our investigation.

To overcome this limitation, we next investigated using BERT or GPT-2 as language models to calculate the perplexity of substitute sentences and establish the change in perplexity when moving from the original sample sentences to the substitute sentences. We quickly realised that using BERT to score its own substitute predictions introduced an issue of circularity, with words that were predicted more often naturally also receiving lower perplexity scores when replacing the original targets.[5] GPT-2 on the other hand - which we had excluded from our original experiments due to its formulation as a left-to-right sequential language model - provided a straightforward and more independent means of evaluation. We used an algorithm suggested by the developers of the Transformer GPT-2 pyTorch implementation[6] to process an entire batch of corpus samples with original or replaced

[5] Due its formulation as a masked language model, BERT itself is not capable of calculating sentence perplexity scores. Salazar et al. (2020) however presented a work-around for this limitation by developing an algorithm that scores sentences via their pseudo-log-likelihood, which is computed by sequentially masking individual words.

[6] `https://huggingface.co/docs/transformers/perplexity`

Substitute	**Occurrence**	**Perplexity**
fish	26.5%	48.8
meat	22.4%	50.5
pork	21.4%	49.7
rice	19.3%	50.7
beef	17.0%	50.1
pig	11.6%	49.7
food	10.8%	53.5
duck	10.2%	48.7
bird	10.0%	50.0

Table 6.2: Most likely substitutes for reference word *chicken* in Wikipedia corpus samples as predicted by BERT's masked token prediction. Percentages indicate the occurrence of the substitute in the top 10 predictions per sample. Perplexity scores are based on a calculation with GPT-2 after replacing *chicken* with the respective substitute in the selection of corpus samples.

target words using a sliding window with stride 512, which, according to the authors, provides an approach closer to the 'true auto-regressive decomposition of a sequence likelihood' than the auto-regressive factorisation of an input sequence presented in the original GPT-2 implementation (Radford et al., 2018).

Returning to the example of *chicken*, the selection of sample sentences from the Wikipedia corpus containing the word *chicken* returned an overall perplexity score of 45.4. We then replaced the target word in each of the sample sentences with a given substitute, like for example *fish*, *meat* and *rice*, and recalculated the selection's perplexity score. Table 6.2 displays the perplexity scores for BERT's top substitute predictions for *chicken* (as presented in Table 6.1). As expected, in all cases the perplexity of the sample selection containing the original target word was lower than the lowest perplexity score calculated for one of its predicted substitutes.

From among the *chicken* substitutes, *bird* and *fish* receive the lowest perplexity scores - much in line with our expectation of these words allowing for a similar polysemy pattern as *chicken* does. This however was not the case for most other targets and their proposed substitutions, with *book* substitutes *novel*, *story* and *series* for example scoring the lowest perplexity - while *story* and *series* do not allow for a physical reading - and other words potentially allowing for a *physical/content* alternation like *album* being assigned the highest perplexity score from the set of substitutes.

Overall, we made two observations limiting the applicability of the perplexity scoring approach. Firstly, the GPT-2 scoring seems to be very sensitive to the proportion of sense extensions represented in the corpus data. In the case of *chicken*, most corpus samples seem to evoke an *animal* reading, causing GPT-2 to score any proposed animal substitute at a lower perplexity than other potential targets. With *book*, the *content* reading was more dominant in the corpus sample, leading to better scores for substitutes with a *content*-only reading. As this means that GPT-2 will fail to correctly indicate polysemy if the data is not properly balanced - and we cannot fully control the balance in the corpus sample in an unsupervised algorithm - perplexity scoring might not be the most reliable tool in evaluating the polysemic potential of a given word.

Secondly, calculating substitute perplexity scores confronts us with the issue of determining a threshold based on which to classify a given substitute as either polysemous or not. When following an unsupervised approach, thresholds should be specified relative to some data-internal reference, which in our experiment can only be the perplexity of the original target. In that case, we could take any substitute that is assigned a perplexity score within a certain range of that of the original target to be considered polysemous in the same way as the reference. Returning to the example of *chicken*'s reveals the limitations of this approach: the perplexity of the original target was calculated at 45.4. Including all substitutes that received a perplexity score only 5% higher (47.5) would not include any, 10% however (49.8) already includes *pork* and *pig* that each do not allow for one of the original readings. For *book* on the other hand, a 10% cutoff would only include the best scoring substitute *novel.* Short of providing annotated data to train a classifier specifically for each target to classify resulting perplexity scores into either indicating polysemy or not, it thus seems unlikely to find an applicable evaluation criterion based on perplexity scores alone.

6.2.2 Distribution-based Sample Scoring

Finding that perplexity scoring wouldn't provide an unsupervised approach to evaluating proposed targets, we investigated an alternative approach to scoring a selection of corpus samples with replaced target words. With this approach, we also collect a selection of corpus sentences that contain a given substitute word, use BERT to obtain contextualised encodings of the potential target word in each of these sample sentences, and evaluate the resulting distribution of vectors in comparison to the original reference sentences to assess its polysemic potential.

Robust Sense Reference Embeddings. To improve the quality of comparisons between BERT embeddings - a crucial aspect of this alternative approach - we first created a new set of more robust 'sense reference' embeddings for the 28 original targets used in our human annotation experiments. To do so, we determined the 20 corpus sample embeddings that were closest to the BERT embeddings of each of the original sample sentences used in the experiments, and, following an approach proposed by Bommasani et al. (2020), averaged their contextualised target word embeddings. As an example, for *chicken* we had used sample sentences

(75) a. The **chicken** sat on the roof of the coop all afternoon long. (animal)
b. The **chicken** was served with steaming hot potato wedges. (food)

to represent its polysemic alternations as *animal* and *food.* According to BERT, the most similar samples from the corpus selection to these two reference sentences were

(76) **Samples most similar to the *animal* reference sentence**

a. Stretch is Garfield 's rubber **chicken** who was given to Garfield on his 6th birthday.

b. Chuck is the clumsiest **chicken** on Rocky Perch Island and the least likely hero you'd ever meet.

c. Billy's pet **chicken** also appears as one of the protagonists of the web-cartoon Biduzidos.

d. The whole squadron almost has Porky crash landing but Porky retaliates and the rescue for the **chicken** becomes a football game.

e. Throughout the episode House and Wilson have a bet to see who can keep a **chicken** in the hospital the longest without security catching on.

Samples most similar to the *food* reference sentence

(77) a. Panuchos feature fried tortillas filled with black beans and topped with turkey or **chicken**, lettuce avocado, and pickled onions.

b. He then presented the main course wilted bok choy with stir-fried **chicken**, chilli, garlic and ginger - plus a squeeze of lemon.

c. Com lam - Glutinous rice cooked in a tube of bamboo of the genus Neohouzeaua and often served with grilled pork or **chicken**.

d. It is normally served with egg or **chicken** fried rice.

e. The main ingredients are grilled meat, **chicken** or pork loin, cured ham, fried green pepper, and sliced tomato.

In order to establish reference embeddings for the *animal* and *food* readings of target

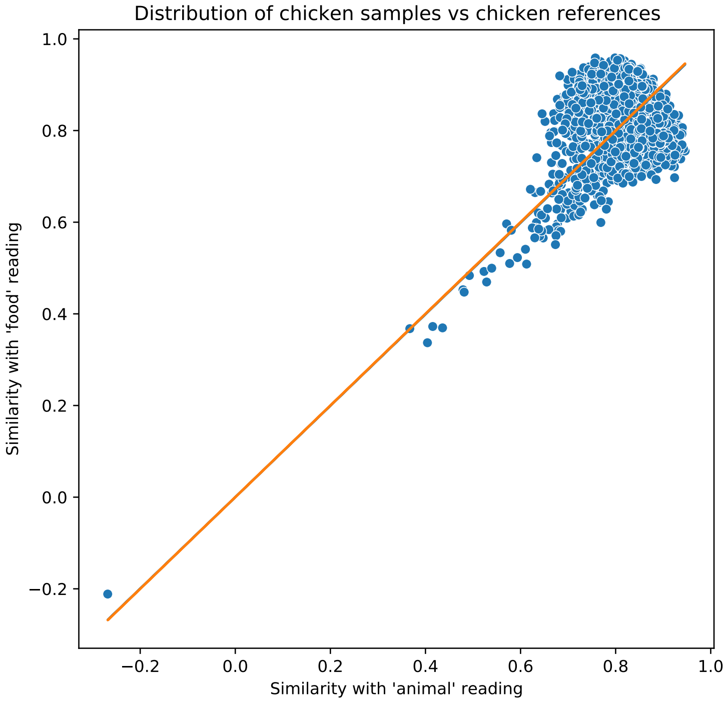


Figure 6.1: Distribution of BERT's Wikipedia *chicken* sample embeddings' similarity scores relative to the *animal* reference embedding on the x-axis, and the *food* reference embedding on the y-axis. True diagonal in orange.

word *chicken*, we averaged the contextualised embeddings for *chicken* per group. This method is meant to establish a more robust representation of the word sense rather than a specific context sentence, which would be the result of encoding just a single sentence (see Bommasani et al., 2020).

These two averaged vectors now were taken to constitute two reference embeddings, representing the target word's two polysemic sense alternations. And since we would like to establish whether a given substitute word allows for polysemic alternations in the same dimensions as the reference word, we continued by investigating whether these reference embeddings already could be regarded as two different dimensions in the interpretation of the target word.

Reference-based Sample Distributions. To test our approach, we first evaluated the full set of *chicken* corpus samples against the two new reference embeddings. We used BERT to derive contextualised word embeddings for a target word within each given corpus sample, again taking the average of the last four hidden layers of BERT Large at the target word's index to represent that word, and averaged encodings of sub-word tokens if tokenisation split the target. We then calculated the cosine similarity (1-cosine) between a sample embedding and each of the two

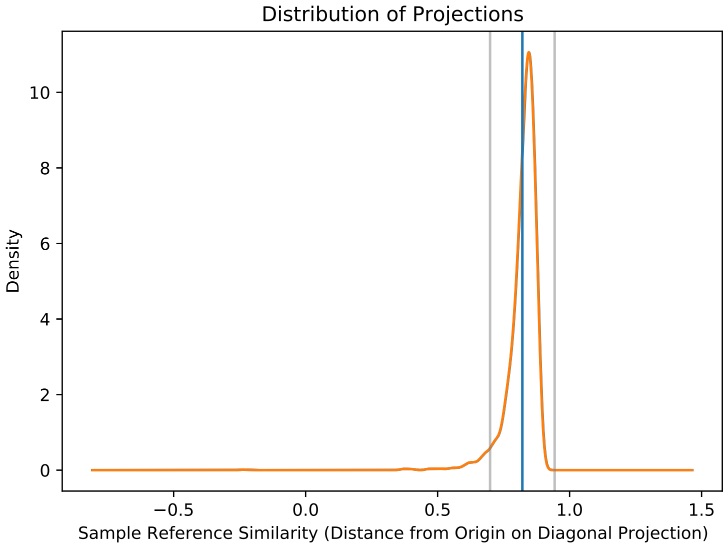


Figure 6.2: Distribution of *chicken* sample similarity scores projected onto the diagonal. Mean of the distribution indicated in blue, 2 standard deviation range in grey.

reference embeddings, and plotted the result as a two-dimensional graph with the two reference embeddings spanning the x- and y-axes, respectively. Consider Figure 6.1 showing the distribution of sample sentences' similarity scores with respect to the *animal* reference embedding on the x-axis, and the *food* reference on the y-axis.

Corpus samples with a low similarity to either reference vector receive low cosine similarity scores on both dimensions, and consequently will be located in the bottom left corner. Samples with high similarities to both references on the other hand will be located in the top right corner of the graph - as are a majority of the samples in this visualisation. Samples that are relatively more similar to one than the other reference vector finally will deviate from the diagonal, with samples that are more closely related to the *animal* reference occupying the space under the diagonal, while samples with a higher similarity to the *food* reference will occupy the top half of the graph.

Sample Filtering - First Pass. A closer inspection of the distribution of *chicken* samples in Figure 6.1 indicates that the deviation from the diagonal is larger in the top right corner of the graph, i.e. for samples that have an overall higher similarity to either of the two reference samples. This spread could provide an interesting signal for the assessment of a target word's polysemic potential, as for polysemes with sense alternations of the same type as the reference words we expect to find some samples that show a higher similarity with one of the reference vectors (indicating

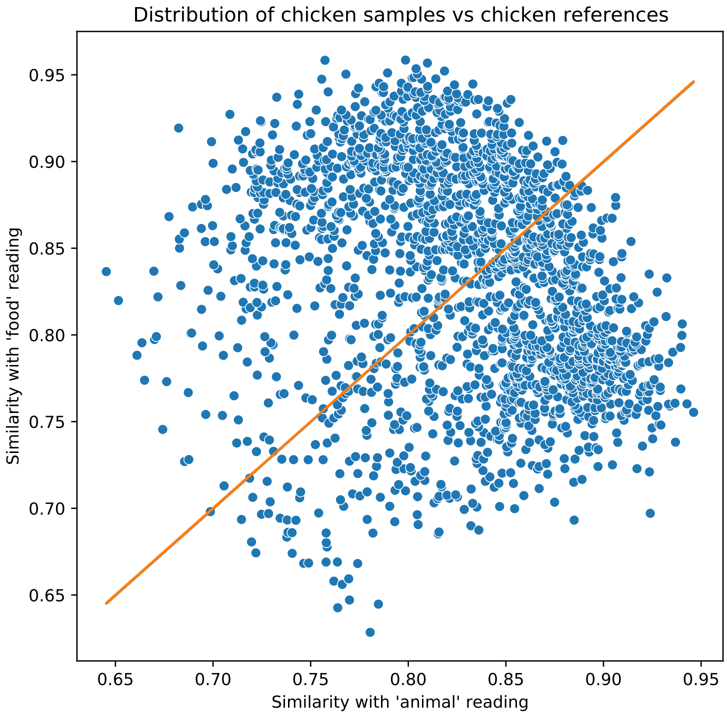


Figure 6.3: Distribution of the filtered *chicken* sample embeddings' similarity scores relative to the *animal* reference embedding on the x-axis, and the *food* reference embedding on the y-axis. True diagonal in orange.

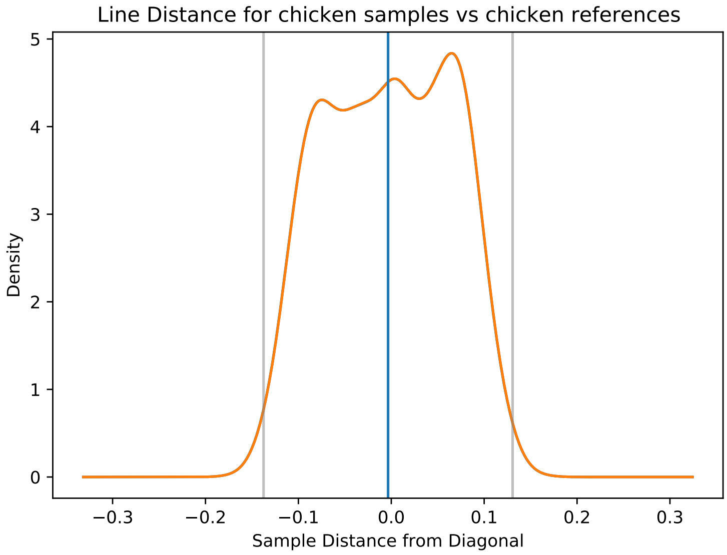


Figure 6.4: Distribution of *chicken* sample similarity scores' distances to the true diagonal. Mean of the distribution indicated in blue, 2 standard deviation range in grey.

that it likely invokes the same sense of the target word as that reference encoding), and some to be more closely aligned with the other.

As we would argue that the necessary and sufficient conditions for determining polysemy with respect to the reference dimensions are met by pinpointing (a minimum number of) corpus sample sentences that allow for either reading, we propose to filter the selection of corpus samples to only retain those that are most likely to represent the target word with respect to the dimensions under consideration. Since we are aiming for an unsupervised approach that could be used to iteratively expand the set of identified target words even without annotated references, this filtering step should be based only on the statistical information provided by the distribution under investigation, and not incorporate any external signals.

One way to do so is to determine the mean of the distribution and to exclude all samples that fall outside a certain threshold of standard deviations from this mean. Since we are primarily interested in the spread of samples relative to the diagonal, we suggest to determine the distribution's mean by projecting all samples onto the diagonal and calculating the mean and standard deviation of the now one-dimensional data. For the *chicken* samples shown in Figure 6.1, this distribution is displayed in Figure 6.2. In this case, filtering out all samples that fall outside of a 2 standard deviation range of the distribution's mean excludes 79 samples, which leaves 1969 for further investigation. Figure 6.3 shows the similarity visualisation of those remaining samples, constituting a 'slice' of the corpus sample.

The remaining slice then can be used to quantify the corpus samples' spread with respect to the two dimensions under investigation, which for example can be done by analysing the sample similarity scores' orthogonal distances to the diagonal. This procedure again reduces the collected data into a single dimension, but this time placing samples on a continuum between the x-axis reference on the left, and the y-axis reference on the right. The resulting distribution of orthogonal line distances of the filtered *chicken* samples to the true diagonal is displayed in Figure 6.4. The mean of the distribution again is indicated in blue, and a range of two standard deviations in grey. This plot shows that the retained slice of *chicken* samples is almost perfectly balanced between the two reference encodings, with the mean of the distribution located at -0.0034. It further also reveals a relatively flat and wide distribution, with a standard deviation of 0.067. This provides us with a reference measure of the samples' spread with respect to the *food* and *animal* dimension of polyseme *chicken*, which we hope could be used to assess the polysemic potential of predicted substitutes in the next step.

Substitute Sample Collection. Having worked out the overall approach, we followed the same steps as when collecting corpus samples for our original experiment targets, but now do so for the full list of predicted substitutes. Combining the top 10 BERT substitute predictions for all of our original target produced a total of 181 potential targets. This increased number of targets slowed down sample collection considerably, and we stopped parsing early after having covered just 25% of the Wikipedia Training set in a matter of a few hours. Still, 139 (77%) of the substitutes were assigned the full set of 2048 corpus samples, and the remaining 42 targets collected an average of 873 samples. Some targets among these however were assigned none or close to no corpus samples, like for example *can* (a substitute for *bottle*), *final* (for *cup*) or *single* (for *record*). As we will discuss in Section 6.3, including a different corpus of less formal and more spontaneous language use might overcome this shortage in future experiments, but for now we will simply exclude these under-represented targets from our investigation.

Once we collected this representative corpus sample for the predicted substitutes, we again used BERT to derive contextualised word embeddings for the different substitutes within their selection of corpus samples, and calculated their cosine similarity to the reference embeddings. This means that even though now we are testing a substitute, like for example *fish* for reference word *chicken*, we are comparing the *fish* contextualised corpus embeddings to the *chicken* reference embeddings encoding its *animal* and *food* reading.

Sample Filtering with Multi-modality Detection. Investigating the similarity scores of *chicken* substitutes, we quickly realised that some substitutes required an additional filtering step to provide us with a uniform sample set. Consider for example the distribution of corpus sample similarity scores of substitute *duck* with respect to the *chicken* reference embeddings for its *animal* and *food* readings in Figure 6.5. While adding only corpus samples that contained substitutes tagged as nouns meant that we did not gather samples using *duck* as a verb, a large part of the Wikipedia samples in the selection appears to use *duck* as a sports term originating in cricketing.[7] Since this use of *duck* constitutes a homonymic reading unrelated to either the reference word's *animal* or *food* reading, these samples cluster in the bottom left corner of the graph. This clustering in turn affects the distribution of projections on the diagonal, artificially lowers the distribution mean, and causes the two standard deviations filtering range to only partially cover the sample 'slice' relevant for to investigation.

[7] In cricket, a *duck* denotes a batsman being dismissed with a score of zero.

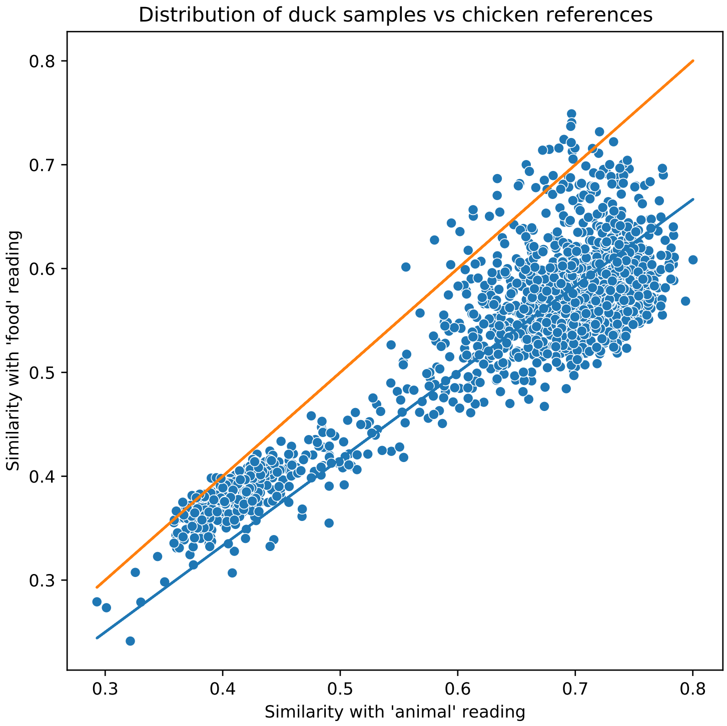


Figure 6.5: Distribution of BERT's Wikipedia *duck* sample embeddings' similarity scores relative to the *chicken animal* reference embedding on the x-axis, and the *food* reference embedding on the y-axis. Best fit through origin in blue, true diagonal in orange.

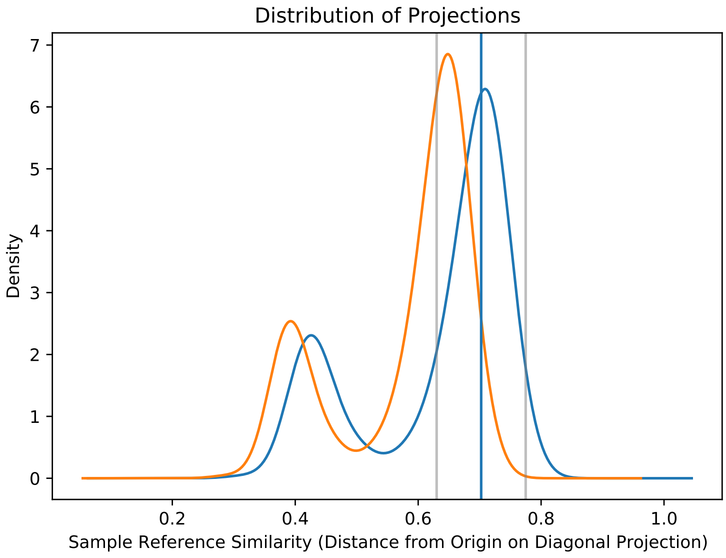


Figure 6.6: Distribution of *duck* sample similarity scores projected onto the true diagonal (orange) and the best fit through the origin (blue). Mean of the best fit distribution (after filtering the low similarity cluster) indicated in blue, 2 standard deviation range in grey.

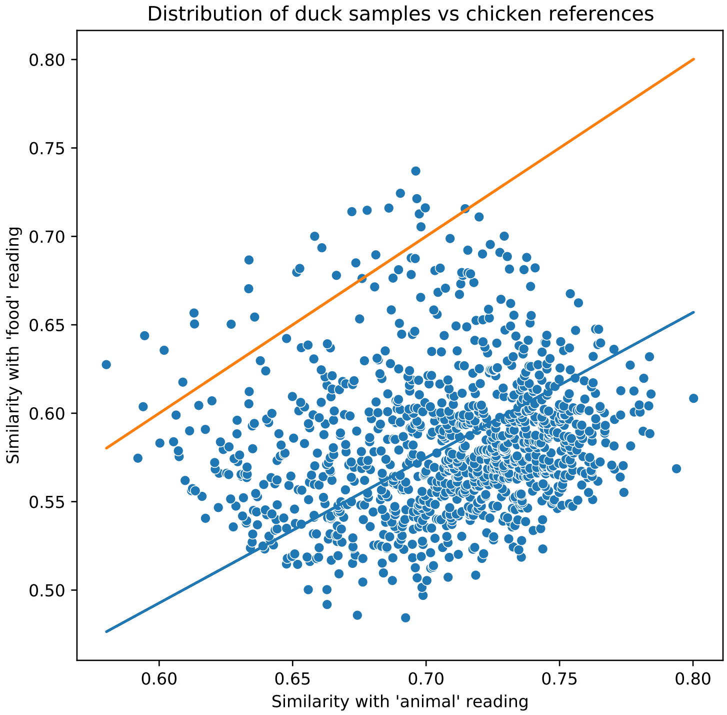


Figure 6.7: Distribution of the filtered *duck* sample embeddings' similarity scores relative to the *chicken animal* reference embedding on the x-axis, and the *food* reference embedding on the y-axis. Best fit through origin in blue, true diagonal in orange.

In order to mitigate this issue, we introduced an additional step before applying sample filtering based on the projection mean: using the sample similarity scores' projections on the diagonal, we first determined whether their distribution was bimodal, i.e. whether there are signs of an extreme clustering of samples splitting the distribution. If that was the case, we only take the clustering of samples with the highest mean to represent the sample slice, and calculate mean and standard deviation scores for filtering only with respect to those samples. To detect multi-modality and automatically determine the intervals of samples belonging to the different clusters we applied the UniDip algorithm (Maurus and Plant, 2016). According to the authors, 'UniDip is a noise robust clustering algorithm for one-dimensional numeric data that recursively extracts peaks of density in the data utilising the Hartigan Dip-test of Uni-modality.'[8] UniDip returns the intervals of samples it deems to belong to different signals given their distribution based on p-value parameter alpha. In our pilot, we used an alpha value of 0.08. If UniDip detected multi-modality and returned multiple intervals, we only used the samples located in the intervals with the highest projection mean to calculate the distribution mean and standard deviations for filtering.

Besides applying this pre-selection step, we also noticed that the corpus samples collected for some substitutes have a clear indication of dominance with respect to one of the two dimensions of polysemic alternations under investigation. Returning to the visualisation of *duck* sample similarity scores in Figure 6.1, we also calculated the best fit through the origin that minimised the mean squared distance to all samples - as indicated in blue. We suggest that the slope of this best fit line could be used as an indicator of sense dominance in the given corpus selection; for *duck* for example this best fit line is located below the true diagonal, indicating that a larger part of samples has a higher similarity to the *animal* than the *food* reference vector. While we first need to filter out the low-similarity samples before drawing any conclusions based on this best fit line, it here already indicated that this sense dominance might skew the distribution of projections on the true diagonal, and we decided to no longer use the true diagonal, but this initial best fit line to calculate the distribution of samples' line projections.

Figure 6.6 showcases these two changes to our sample filtering heuristics, displaying the distribution of best fit projections in blue besides the true diagonal projections in orange, and indicating the filtered best fit distribution's mean and 2 standard deviation intervals. The filtering range now clearly only contains samples that belong to the second distribution, excluding any samples that belong to the

[8] `https://github.com/BenjaminDoran/unidip`

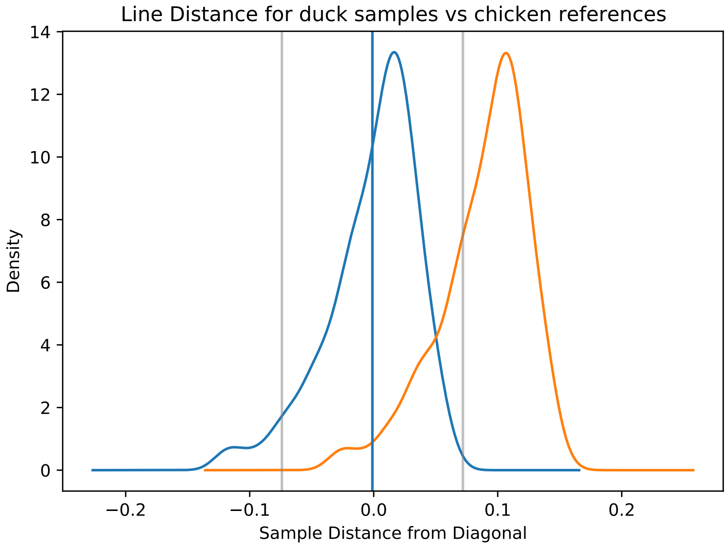


Figure 6.8: Distribution of *duck* sample similarity scores' distances to the true diagonal (orange) and the best fit through the origin (blue). Mean of the distribution indicated in blue, 2 standard deviation range in grey.

homonymic sports term cluster with a best fit projection mean of around 0.42. The result of this filtering is displayed in Figure 6.7, containing only the 'slice' of samples in the 2 standard deviation range. The best fit slope (blue) now is calculated at 0.82, indicating a clear dominance of *animal* samples in the sample slice.

Line Distance Sample Scoring. Based on the filtered sample set, we can again calculate the distribution of line distances to measure a target word's spread of similarity scores with respect to the *animal* and *food* dimensions under investigation. Instead of the true diagonal, for this we now also used the best fit through the origin to centre the resulting distribution around zero and reduce distribution skewness introduced by sense dominance. Figure 6.8 shows the resulting distribution of *duck* samples' line distances to the true diagonal in orange, and the best fit in blue (standard deviation = 0.036.)

As with the initial calculations of substitute perplexities in Section 6.2.1, a central issue with directly comparing the spread of samples' similarity scores is determining an unsupervised, data-driven threshold to decide whether a given distribution indicates polysemic potential with respect to the dimension under investigation or not. In our pilot, we explored two alternatives to making this decision based on the spread of the distribution alone.

Polysemy Detection through Multi-Modality. The first approach utilises the automated detection of multi-modality - this time in the distribution of samples'

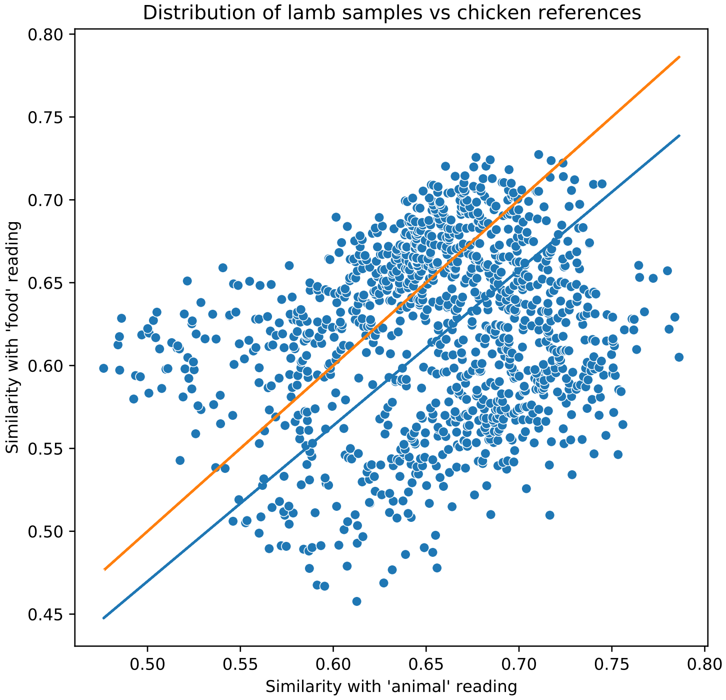


Figure 6.9: Distribution of the filtered *lamb* sample embeddings' similarity scores relative to the *chicken animal* reference embedding on the x-axis, and the *food* reference embedding on the y-axis. True diagonal in orange.

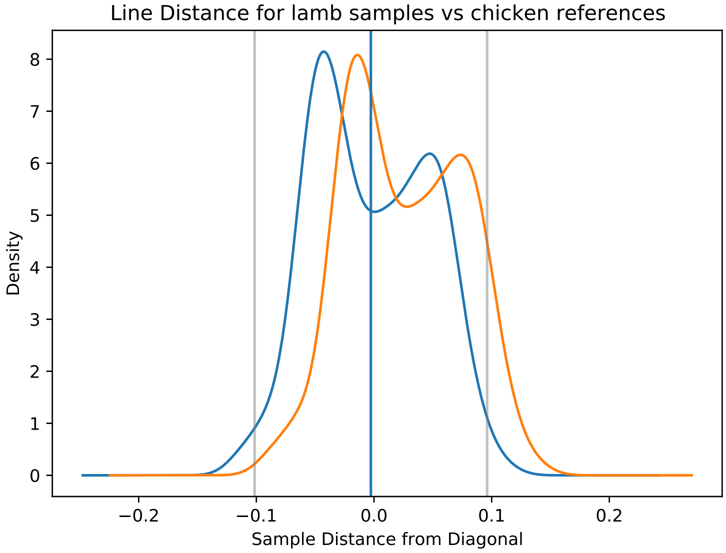


Figure 6.10: Distribution of *lamb* sample similarity scores' distances to the true diagonal (orange) and the best fit through the origin (blue). Mean of the distribution indicated in blue, 2 standard deviation range in grey.

line distances rather than their projections onto that line. Consider for example the distribution of line distances of the filtered slice of *lamb* samples relative to *chicken*'s *animal* and *food* references displayed in Figures 6.9 and 6.10. Through our annotation experiments, we confirmed that *lamb* also allows for both, *animal* and *food* readings. The distribution of *lamb* corpus samples' distances to the best fit (blue) now indicates that the collected *lamb* corpus samples seem to form clusters with respect to their line distance: one below the mean, and one above it. Given that the x-axis captures a continuous indication of sample similarity with the *animal* reference towards the left and the *food* reference towards the right, this bi-modality of the sample distribution might indicate a split of samples into *animal* and *food* reading clusters - corrected for sense dominance by calculating their line distance relative to the best fit rather than the true diagonal.

We again used UniDip to determine whether the one-dimensional distribution of line distances is multi-modal, and interpret multi-modality as a strong indicator of polysemic potential with respect to the senses under investigation. A subsequent evaluation of substitute words however revealed that only a small fraction of sample distributions exhibited clear multi-modality, rendering multi-modality testing a strong but not very sensitive signal for detecting polysemic potential.[9] Given the limited applicability of the multi-modality test, we required a second evaluation criterion that could more reliably quantify any substitute sample's polysemic potential. For our pilot, we developed this more universally applicable criterion based on the previously mentioned assumption that the necessary and sufficient conditions for determining polysemy with respect to the reference dimensions are met by simply pinpointing (a minimum number of) corpus sample sentences that allow for either reading. Applied to the previously established distributions of line distances, determining examples of both sense readings could be implemented by considering only samples of extreme similarity with either sense reference encoding, and determining whether both senses have at least some representation in that selection of corpus samples. Considering an even smaller selection of extreme samples also has the additional benefits of mitigating the impact of little informative mid-field samples, and further reduces the impact of representation biases in the corpus sample.

Sample Filtering - Second Pass. In order to apply a second unsupervised filtering step that would retain only those samples that are closest to either of the given references, we again relied on calculating the means and two standard deviation

[9] See tables C.1 through C.18 in Appendix C for indications of multi-modality in the line distance distributions of the tested potentially polysemic targets

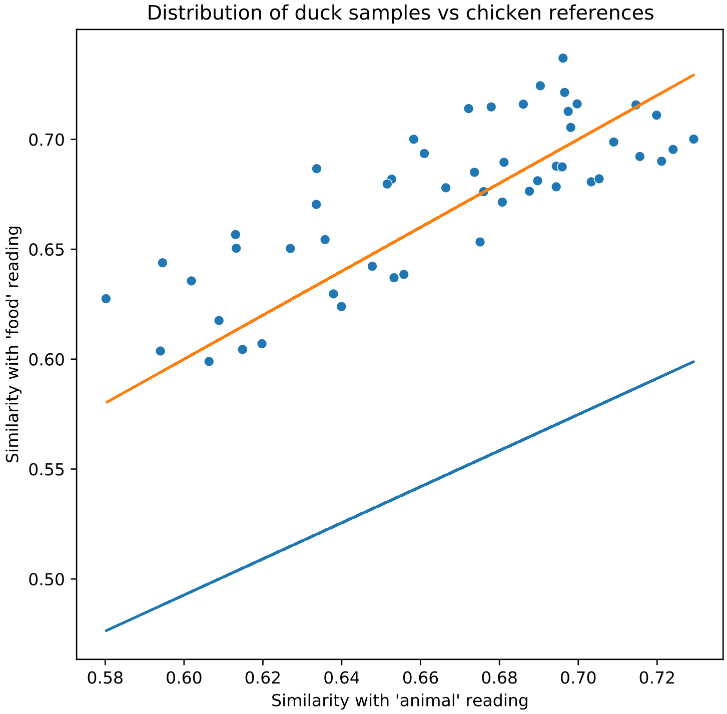


Figure 6.11: Distribution of the double-filtered *duck* sample embeddings' similarity scores relative to the *chicken animal* reference embedding on the x-axis, and the *food* reference embedding on the y-axis. Original sample set's best fit in blue, true diagonal in orange.

ranges of a given distribution - this time considering the distribution of the samples' line distances as used in the previous section, and filtering out samples inside the two standard deviation range rather than those outside of it. Returning to potential target *duck*, figure 6.8 indicates the mean (blue) and two standard deviation range (grey) of target samples' distances to the best fit (as indicated in Figure 6.7). After filtering out all samples contributing to the distribution within the filtering range, the remaining samples are those closest to either of the two reference vectors (see Figure 6.11).

Polysemy Detection through Extreme Samples. Figure 6.11 reveals two interesting observations relevant to our investigation: firstly, the original sample distribution's best fit has a significantly more shallow slope (0.82) than the true diagonal (always 1). This indicates that the target words has a dominant sense in the sample slice, and that that dominance is exhibited by its *animal* reading.[10]. Secondly, after the second pass filtering, all remaining samples are on the *food* side of the best fit, with a majority even on the *food* side of the true diagonal. We propose that this distribution of extreme samples indicates that besides a dominant *animal* reading, the investigated target *duck* also can exhibit a *food* reading - which would render it polysemous with respect to the alternation under investigation.

In order to derive a quantifiable, data-driven method of determining whether a target word's extreme samples indicate polysemy, we developed a heuristics taking into account the unfiltered distributions best fit as an indicator of sense dominance, and the mean of the filtered extreme samples as an indicator of whether the subordinate reading also is available. We assume a target word to indicate sense dominance if the slope of the best fit deviates significantly from the true diagonal, which we take to be the case if it is either below 0.95 or over 1.05. If the slope is significantly lower than the diagonal, the dominant sense is linked to the reference vector on the x-axis, if the slope is higher than the diagonal, the dominant sense is linked to the reading indicated by the reference vector on the y-axis. If the slope of the best fit is not significantly different from the true diagonal, we assume that the selected corpus sample does not exhibit dominance effects with respect to the sense alternations under investigation.

If a target word exhibits sense dominance for one of the sense alternations under investigation, we propose that the subordinate reading also is available - and the target can be considered to exhibit polysemy with respect to the reference interpre-

[10] A slope steeper than the true diagonal on the other hand would indicate a sense dominance of the *food* reading.

Algorithm 1 Heuristics for determining polysemy with respect two two reference readings based on corpus samples' similarity scores' distances to the true diagonal and their best fit through the origin.

```
if Slope of best fit $\leq 0.95$ then
    Dominant reading is sense 1
    if Mean of extreme samples' diagonal distances $< 0$ then
        Subordinate reading also is available. Target is polysemous.
    else
        Subordinate reading not available. Target is not polysemous.
    end if
else if Slope of best fit $\geq 1.05$ then
    Dominant reading is sense 2
    if Mean of extreme samples' diagonal distances $> 0$ then
        Subordinate reading also is available. Target is polysemous.
    else
        Subordinate reading not available. Target is not polysemous.
    end if
else
    No dominant reading
    if Slope of best fit $\leq 1$ then
        if Mean of extreme samples' diagonal distances $\leq$ -0.05 then
            Two readings available. Target is polysemous.
        else
            Target is not polysemous.
        end if
    else
        if Mean of extreme samples' diagonal distances $\geq 0.05$ then
            Two readings available. Target is polysemous.
        else
            Target is not polysemous.
        end if
    end if
end if
```

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
fish	26.46	0.866	1	0.03135		X
meat	22.36	1.079	2	0.03569		
pork	21.44	1.104	2	0.03448		
rice	19.29	0.988	None	0.04358		X
beef	16.99	1.047	None	0.03992		
pig	11.62	0.873	1	0.03449		
food	10.79	1.037	None	0.02391		
duck	10.16	0.821	1	0.03649		X
bird	9.96	0.803	1	0.02377		

Table 6.3: Evaluation of corpus-based substitutes for reference word *chicken*. Senses 1: *animal*, 2: *food*.

tations - if the mean of the filtered extreme samples falls on the other side of the true diagonal. Concretely, if the slope of the best fit is below 0.95, we assume the target word is polysemic if the mean of the filtered extreme samples compared to the true diagonal is below 0. Alternatively, we consider a word polysemic if its slope is above 1.05 and the mean of the remaining samples is greater than zero. In case there is no dominant sense (the slope of the best fit doesn't significantly diverge from the true diagonal), we additionally require the mean of the filtered extreme samples to significantly deviate from the true diagonal to compensate for the higher chance of noise due to the relatively small spread of samples. If the slope of the best fit is below the diagonal but not significantly so (between 1 and 0.95), the mean of the filtered extreme samples relative to the true diagonal has to be below -0.05 to render the target word polysemous, and conversely needs to be larger than 0.05 if the slope is between 1 and 1.05 (not significantly steeper than the true diagonal). Algorithm 1 represents this heuristics in algorithmic form.

6.2.3 Results

Since the pilot heuristics was developed for investigating two sense alternations, we first investigated a selection of reference words with two alternations from our experiments in Chapter 5 for which we had gathered sufficient corpus samples to run our analysis. Out of 15 reference words, 9 were classified as polysemous with respect to the two sense reference embeddings derived from the corpus samples most similar to the sample sentences used in the annotation study. Tables 6.3 through C.16 contain the results of one iteration of our polysemy detection pilot

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
chicken	25.27	0.964	None	0.0407		
beef	23.03	0.982	None	0.0279		
pork	22.12	1.069	2	0.0254		
meat	21.54	1.021	None	0.0249		
goat	20.63	0.912	1	0.0353		X
fish	18.14	0.983	None	0.0260		X
sheep	16.24	0.903	1	0.0374		X
cow	15.41	0.875	1	0.0292		
pig	13.50	0.927	1	0.0366		X
dog	11.68	0.844	1	0.0286		

Table 6.4: Evaluation of corpus-based substitutes for reference word *lamb*. Senses 1: *animal*, 2: *food*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
magazine	62.79	1.066	2	0.0296		X
paper	47.85	1.124	2	0.0288		X
publication	38.38	1.050	None	0.0289		
journal	36.28	1.092	2	0.0330		X
daily	28.56	1.118	2	0.0185		
weekly	26.32	1.140	2	0.0175		
newspapers	24.66	1.046	None	0.0368		
periodical	22.31	1.106	2	0.0201		
news	13.96	0.942	1	0.0268		
press	11.62	0.954	None	0.0316		

Table 6.5: Evaluation of corpus-based substitutes for reference word *newspaper*. Senses 1: *physical*, 2: *organisation*.

using reference words *chicken*, *lamb* and *newspaper*. For each reference word, the tables indicate i) the most often predicted substitute words as derived in Section 6.1 and the percentage of these targets occurring among the top 20 BERT predictions of the reference sample, ii) the slope of the best fit after plotting all target samples against the reference sense embeddings, and whether their distribution indicates a dominant sense according to our heuristics described in Paragraph 6.2.2, iii) the standard deviation of the distribution of a given substitute's similarity scores and whether that distribution is multi-modal (first test for polysemy, see Paragraph 6.2.2 - note that none of the three sets actually contain a word classified as multi-modal), and iv) the binary indicator of whether a given target word is considered polysemous according to the full pilot heuristics detecting primary and secondary uses with respect to the given reference embeddings. The full set of tables can be found in Appendix C.

Overall, 34 out of 159 tested target words (21.38%) were classified as polysemous by the pilot heuristics, with an average of about 3 in each list of BERT's top substitutes. The substitutes for *beer* exhibit the largest number of words classified as polysemous (7 out of 9), while the substitutes for *wine*, *building*, *pheasant*, *lunch* and *window* didn't contain a single word classified to be polysemous with respect to the reference alternations.

6.2.4 Qualitative Analysis

Animal/Food* Alternation: *Chicken*, *Lamb* and *pheasant Three of *chicken*'s substitutes were classified as polysemous: *fish*, *rice* and *duck* (see Table 6.3). Out of *lamb*'s substitutes, *goat*, *fish*, *sheep* and *pig* were classified as polysemous with respect to the reference embeddings derived from the *lamb* corpus samples (Table 6.4). Starting with *lamb*'s substitutes, on a first glance *sheep* and *pig* might be considered false positive results, as both of these animals have a specialised food reading - *pork* for *pig* and *mutton* for sheep. While the heuristics indeed indicate a dominant *animal* reading for both of these targets, it also detects a second, subordinate reading in relative proximity to the *food* reference vector. In order to investigate this labelling decision, we manually inspected those sentences from the *pig* and *sheep* corpus sample that were determined as most similar to the *food* reference.

Pig was rated polysemous with respect to the *lamb* references but not so for the *chicken* references. When compared to the *lamb* reference embeddings, many of the samples closest to the *food* reference vector use the target word *pig* in a compound noun arguably referring to a foodstuff:

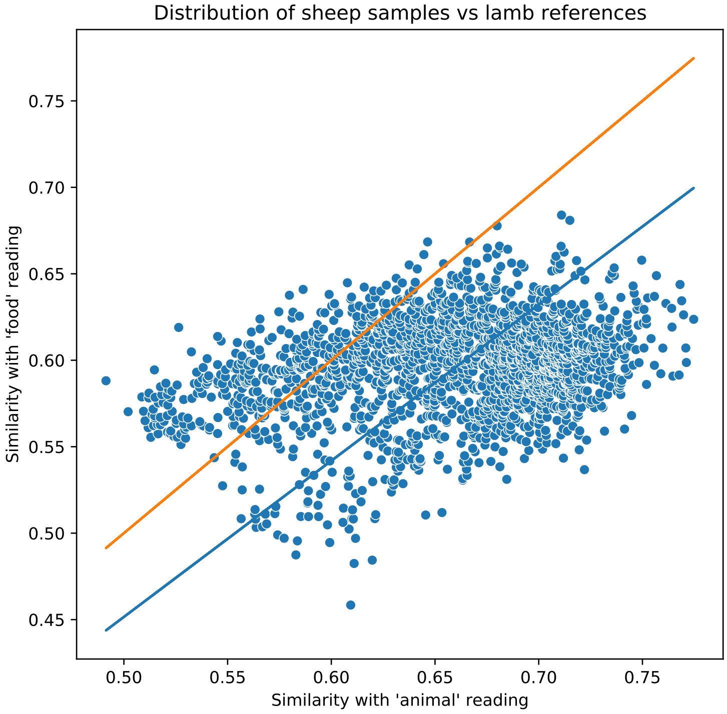


Figure 6.12: Distribution of the filtered *sheep* sample embeddings' similarity scores relative to the *lamb animal* reference embedding on the x-axis, and the *food* reference embedding on the y-axis. True diagonal in orange.

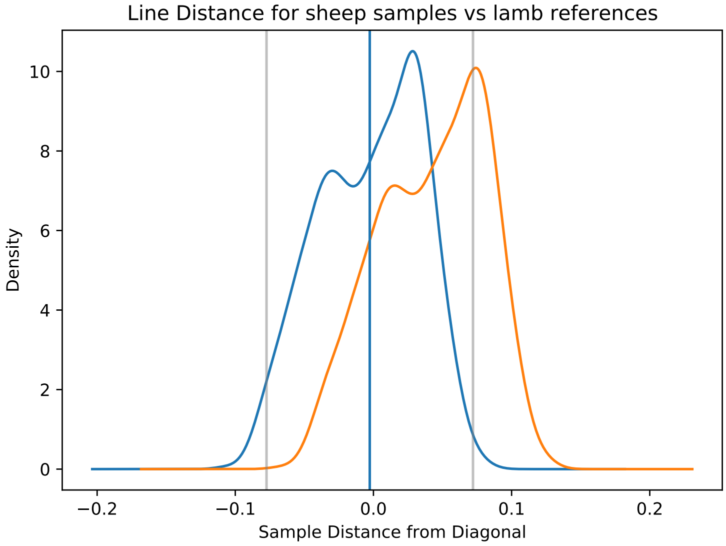


Figure 6.13: Distribution of *sheep* sample similarity scores' distances to the true diagonal (orange) and the best fit through the origin (blue). Mean of the distribution indicated in blue, 2 standard deviation range in grey.

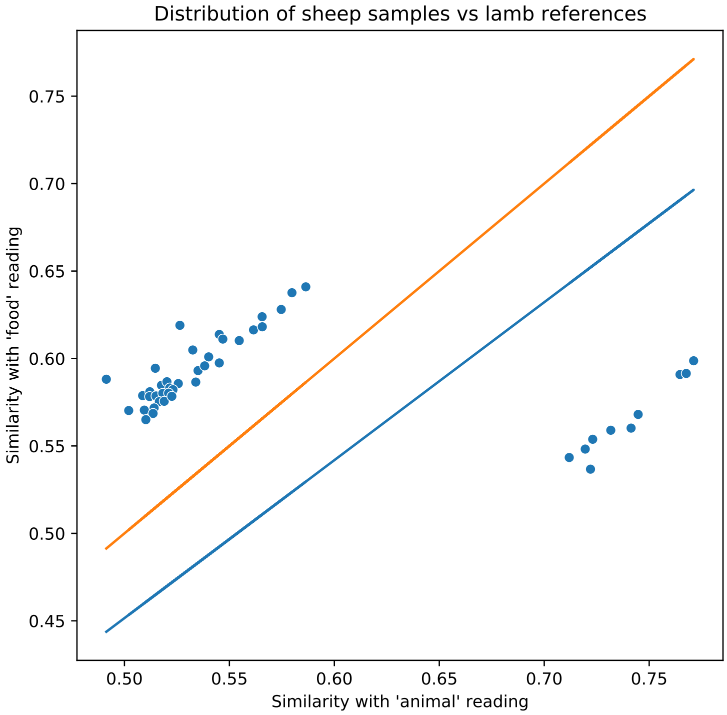


Figure 6.14: Distribution of the double-filtered *sheep* sample embeddings' similarity scores relative to the *lamb animal* reference embedding on the x-axis, and the *food* reference embedding on the y-axis. Original sample set's best fit in blue, true diagonal in orange.

(78) ***Pig* samples most similar to *lamb's food* reference**

a. Corn flour, **pig fat**, lard or butter, cheese, milk and whey are common ingredients.

b. The main ingredients of Chao Tian Guo are a thin pancake, a meat ball, **pig offal**, tofu and soy.

c. Kagoshima **black pig pork** is also produced in Kawanabe.

d. From this nucleus the market grew with stalls for garden produce, **pig meat**, dairy products and fish.

e. Typical local foods include chocho purtumute and **guinea pig** with potatoes among others.

f. Another favorite dish is cook-up or pelau, which combines chicken, **pig tail**, saltfish and vegetables with rice and pigeon peas.

Filtering out compound nouns through for example dependency parsing is a possible option to exclude these false positives in the future. For the evaluation of our pilot, for now we take them as acceptable indications of polysemy given the formulation of our heuristics.

For target *sheep*, the sample sentences flagged as most similar to the *food* reference seem to contain no actual food readings, and instead all appear to refer to livestock and farm animals:

(79) ***Sheep* samples most similar to *lamb's food* reference**

a. Besides agriculture, inhabitants deal with breeding of livestock such as cattle, **sheep**, goats horses, poultry, beekeeping etc.

b. The acorn-eating bovines share the fields with large numbers of horses, **sheep**, goats, cows donkeys, mules and chickens.

c. Cattle, **sheep**, pigs and poultry are all commercially farmed.

d. Amongst the livestock industry, rearing of cattle, **sheep**, pigs and goats is prominent but herds of horses and donkey mules are also found.

e. Livestock includes cattle, mostly dairy, pigs, **sheep**, goats and domestic fowl.

Figures 6.12, 6.13 and 6.14 show the distribution of samples relative to the *lamb* reference embeddings, the distribution of their orthogonal distances to the best fit and true diagonal, and the double-filtered set of extreme samples. Figure 6.12 reveals a very wide distribution, indicating that the two standard deviation 'slicing' of the sample set did not remove many samples - an issue our pilot heuristics can be susceptible to if the original sample set contains a number of extremely low-scoring samples, but not enough to trigger UniDip's multi-modality criterion. The resulting distribution of line distances has two extreme points - while also not passing UniDip's multi-modality threshold, which in this case stretches the distribution relative to a clear uni-modal peak, and leads to a higher number of samples falling outside the two standard deviation threshold for selecting extreme samples. As a result, the ultimately remaining 'extreme' samples - while located on the *food* side of the true diagonal - have relatively low similarity scores compared to the *food* reference embeddings (similarity scores here are ranging between 0.5 and 0.65 while samples most similar to the *animal* reading can obtain similarity scores of close to 0.8). Given that *sheep* is classified as polysemous based on these samples, this calls for either an adjustment in the filtering process, or an additional minimum similarity criterion for the selected 'extreme' samples.

Continuing with the substitutes for *chicken*, *rice* seems to be a surprising prediction for exhibiting both a *food* and an *animal* reading. Investigating the *rice* samples that achieved the highest similarity scores with the *animal* reference however reveals that some refer to rice farming - much like some of *chicken*'s animal samples did refer to chicken breeding:

(80) ***Rice* samples most similar to *chicken's animal* reference**

a. In the very last episode, she is seen taking over her husband 's spot in the **rice fields**.

b. It is believed that the cult was created by **rice farmers** in need of land and water and at its peak was extremely popular.

c. General Surayud had Sitthichai 's device mass-produced, and **rice mills** and markets were forced to buy it.

With samples exhibiting a similar type of interpretation as the reference set, there is an argument to be made here that the sense represented by the averaged reference vector might not specifically be an *animal* reading, but rather something more abstract like 'pre-processing stage for a consumable.' Since we manually assigned the vector descriptions based on the alternations tested in the human annotation study, in this case the heuristics might have actually detected the same pattern of alternations, but the labelling of the two different interpretations might have been chosen to be too specific to properly capture the actual nature of the alternation.

The production of false negatives for detecting the *animal/food* alternation in substitutes for *chicken* and *lamb* seems to be limited to *chicken* itself, which according to the heuristics neither is classified as polysemous with respect to its own nor *lamb*'s reference embeddings, and *duck* in the substitutes for *pheasant.* Figures 6.15 and 6.16 show the distributions of line distances of the *chicken* corpus samples

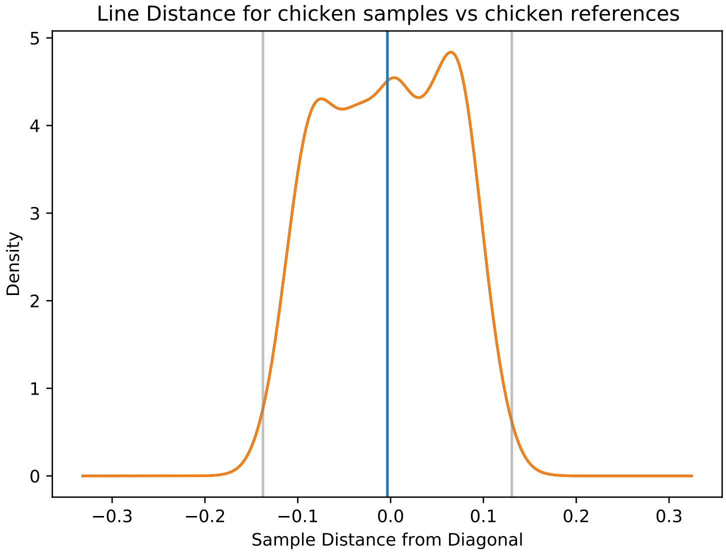


Figure 6.15: Distribution of *chicken* sample similarity scores' distances to the true diagonal (orange) and the best fit through the origin (blue) with respect to *chicken* reference vectors. Mean of the distribution indicated in blue, 2 standard deviation range in grey.

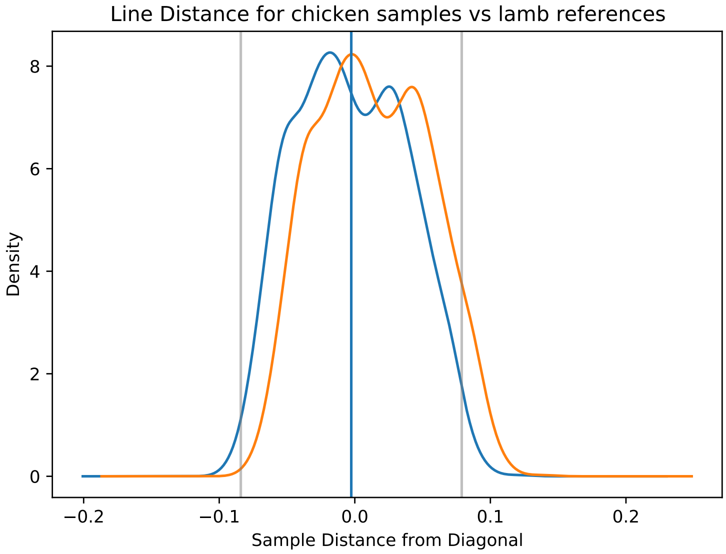


Figure 6.16: Distribution of *chicken* sample similarity scores' distances to the true diagonal (orange) and the best fit through the origin (blue) with respect to *lamb* reference vectors. Mean of the distribution indicated in blue, 2 standard deviation range in grey.

relative to the *chicken* and *lamb* reference embeddings, respectively. Both distributions have multiple extrema, but neither pass the noise threshold implemented in UniDip, which could indicate that we set too strict a multi-modality criterion in our application of the UniDip algorithm.

Inspecting the classification of substitutes for *pheasant* revealed another weak point of the heuristics, as it seems to strongly depend on the initial selection of reference samples. Remember that in order to establish more robust representations of a certain sense encoding, we select and average 20 corpus samples that are closest to each of the reference samples used in the human annotation study. To do so, we usually simply use the first two reference samples from our materials to create these reference vectors. For *pheasant*, this procedure however resulted in a classification flagging all substitutes as *food*, while the list of substitutes only includes animals - none of which except *duck* would traditionally be considered to allow for a *food* reading (Table C.9). Investigating the sample sentences used to create the *pheasant food* reference vector, only one sample arguably exhibits a food reading, indicating that either there are no *food* samples for *pheasant* in the corpus sample selection, or that the reference sample used to identify *food* samples is not suitable. To test this, we also compared the *pheasant* samples to the previously established *chicken* references and investigated the sample sentences that are closest to *chicken*'s *food* reference. The list now appears to almost exclusively contain *food* readings of *pheasant*, indicating that they are indeed present in the corpus sample and that the issue seems to lie with the embedding of the reference sample sentence:

(81) ***Pheasant* samples most similar to *chicken*'s *food* reference**

a. It contained sow's udder, **pheasant**, wild boar and ham in pastry.

b. Rattlesnake and **pheasant** were on the menu as well as elk and venison.

c. Whether it was pork, venison, **pheasant** or beef - it was all eaten up.

d. Audrey Smith recalled having to learn to eat the **pheasant**, lamb and veal that her mother Georgia Anderson brought home from Maymont.

e. Juniper berry sauce is often a popular flavoring choice for quail, **pheasant**, veal, rabbit, venison and other game dishes.

f. The Dutch also enjoy more traditional Christmas-dinners especially meats and game like roast beef, duck, rabbit and **pheasant**.

g. The restaurant became known for serving locally sourced **pheasant** as well as a pâté known as Pheasant Farm Pate.

Since our materials contain two sample sentences for each sense, we repeated the evaluation of *pheasant*'s substitutes when using the second *food* reference sample to create the *food* reference embedding, which now returned *bird*, *pigeon* and *owl* as polysemous. Besides the fact that these three targets could be considered false positives (most of their *food* samples again are unrelated compound nouns like *bird eggs* and *pigeon peas*), the change in classification indicates that the heuristics requires stable and representative sense embeddings to work well, and using just a single reference sentence might be too unstable to provide these in some cases. This is less of an issue in later iterations of the algorithm as reference vectors will have been established using reference vectors based on corpus samples, but it strongly affects the first iteration, and indicates that it might be beneficial to manually select a set of samples representing a certain interpretation to use for the establishment of reference vectors rather than a single reference sentence to kickstart the sample collection.

Content-for-Container **Alternation:** ***Beer*** **and** ***Wine*** *Beer* produced seven substitutes that were classified as polysemous with respect to the *container/content* reference vectors: *wine*, *liquor*, *beers*, *ale*, *coffee*, *drink* and *food* - all of which except *drink* were attested a dominant *container* reading in the corpus sample (Table C.3). *Wine* on the other hand did not produce any substitutes classified as *polysemous*, even though the list of substitutes contains *beer*, *food* and *liquor* (Table C.4). Investigating the samples used to create *wine*'s reference embeddings, it appears that not many of the *container* samples actually seem to refer to wine bottles (some explicitly mention *wine barrels*) and *content* samples mainly refer to vineyards, grapes and producers of wines. This indicates that here again the single reference sentence used to establish the reference vectors might have been insufficient to determine relevant corpus samples.

Container-for-Content **Alternation:** ***Bottle*** **and** ***Glass*** *Bottle* produced two substitutes classified as polysemous with respect to the *container-for-content* alternation under investigation: *box* and *bag* - with *box*'s distribution of line distances exhibiting multi-modality (Table C.5). Figures 6.17 6.18 show the distribution of *box* corpus samples relative to the *bottle* reference vectors, and the resulting distribution of line distances, respectively. Especially Figure 6.17 indicates that the sample distribution again is very wide, causing the heuristics to struggle in selecting a relevant slice of samples for further investigation, including a large number of unrelated and irrelevant samples in the analysis and decision processes. As a result, none of the *box* samples most similar to *bottle*'s *content* reading actually refer to the contents of a box. *Bag* parallels these observations.

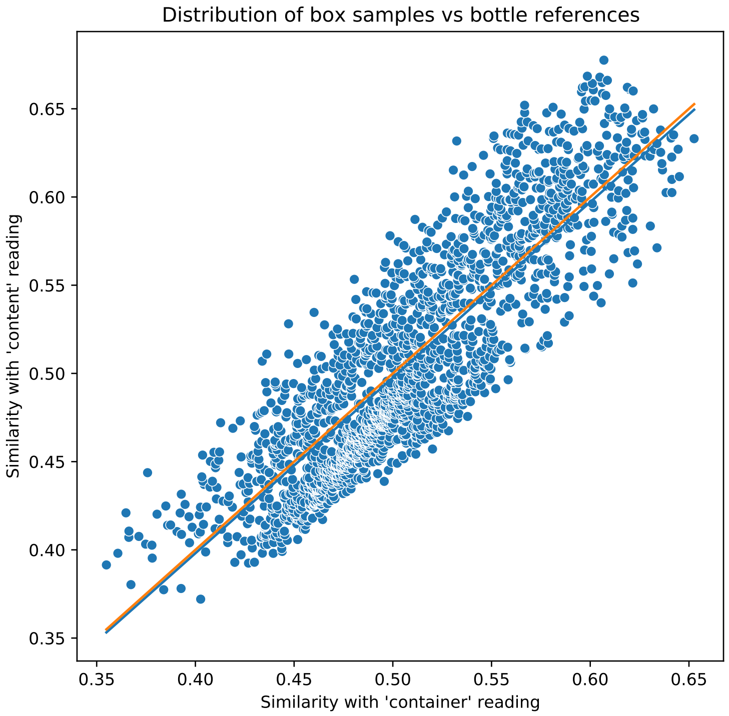


Figure 6.17: Distribution of the filtered *box* sample embeddings' similarity scores relative to the *bottle container* reference embedding on the x-axis, and the *content* reference embedding on the y-axis. True diagonal in orange.

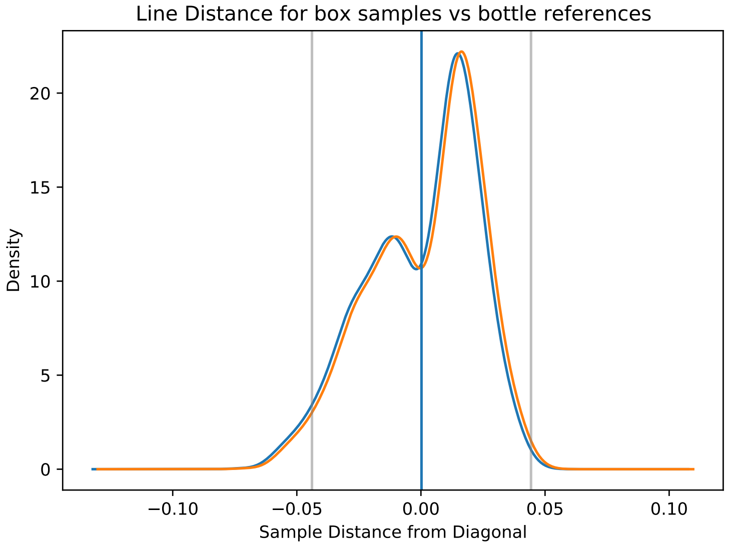


Figure 6.18: Distribution of *box* sample similarity scores' distances to the true diagonal (orange) and the best fit through the origin (blue). Mean of the distribution indicated in blue, 2 standard deviation range in grey.

Glass produced one substitute labelled as polysemous: *steel* - which highlights an issue with our target selection heuristics (Table C.6). As the only restrictions on selecting corpus samples for a given reference word were that they i) had to include the reference word, ii) the reference word had to have the same word type in the sample sentence as in the original experiment sample, and iii) that the corpus sample should not be longer than 25 words, the collected slice of up to 2048 sentences will represent the corpus' statistics with respect to meaning and sense dominance. This means that for words whose polysemic alternations are far less common than a more dominant homonymic interpretation, the corpus sample will largely represent that interpretation. As a result, BERT's top predictions for *glass* - which in our experiments was used to refer to either a container of liquids or its contents - based on the given Wikipedia sample resulted in *stone* (24%), *metal* (21%), *steel* (19%), *crystal* (17%), *wood* (17%), *water* (13%), *window* (11%), and *plastic* (10%) - all of which arguably more connected to *glass* as a building material and none allowing for the *container/content* alternation we would like to investigate. Given the skewed representation of *glass* in the corpus sample, the samples used in creating its reference vectors also are less likely to correctly represent the intended sense reading:

(82) ***Glass* samples included in the *container* reference embedding**

a. Gibbs Abby and Vance escape the explosion unscathed while McGee is hospitalized after being impaled in the stomach by a **shard of glass**.

b. He received a number of wounds to his face from the **shattered glass** and bullet fragments, and his shirt was badly blood-stained.

c. In one incident the homeowner hit the patient with **a piece of glass** before police could arrive to arrest the patient.

d. His uniform was covered in **broken glass**.

e. One passenger suffered minor injuries to their hand from **broken glass**.

(83) ***Glass* samples included in the *content* reference embedding**

a. When the time comes to remember all who died, Cotto **raises his glass** to the memory of Londo Mollari.

b. In 1961 he gave **a glass of champagne** to every member of the audience who had watched Simple Spymen.

c. She became thirsty along the way and stopped at a house where she asked a black woman named Mrs McCarthy for **a glass of water**.

d. The town then assumed the name Buttermilk Fort because travelers passing through were encouraged to stop for **a glass of cold buttermilk** while they rested.

e. Thus it is commonly sung as a toast typically for the first **glass of spirit** at a seated dinner.

One way to improve the corpus-based collection of target words that might allow for the same polysemic alternations as a given reference word, could be including a second filtering step when parsing the corpus for samples: Instead of directly adding a sample sentence containing the target word, we could compare the contextualised BERT encoding of the reference word in that sample sentence with the encodings of the target in our original experiment sample sentences. Given that BERT seems sensitive enough to at least clearly tell apart homonymic usages of the same word, a threshold on the cosine similarity between the sample and reference embeddings could be used to only retain samples that invoke the same *meaning* as the reference samples - independent of the specific *sense* invoked in the sample.

Process/Result* Alternation: *Construction* and *Building With *completion*, only one of *construction*'s substitutes was flagged as polysemic by the pilot heuristics, while for example *building*, *development*, *creation*, *reconstruction* or *expansion* - which would be expected to exhibit a *process/result* alternation - were not (Table C.7). Inspecting the corpus samples that went into *construction*'s reference revealed an expected selection of *process* samples - most of which referring to constructions beginning or ending at certain dates - while some *result* samples specifically referred to sentence and grammatical construction. This linguistic focus affects the selection of target samples, causing *result* samples of substitutes like *expansion* to also partially contain linguistic references, indicating that the two dimensions under investigation do not correctly represent the alternation we aimed to investigate. *Building* on the other hand did not produce any polysemic substitutes - which here can be considered expected behaviour since all substitutes are types of buildings, none of which allow for a process reading (all targets were correctly assigned a dominant *result* reading, see Table C.8).

Event/Food* Alternation: *Lunch* and *Dinner The pilot heuristics did not flag any of *lunch*'s substitutes as polysemous - even though the list of BERT's predictions includes *dinner* and *breakfast* (Table C.10). Using *dinner*'s reference samples, *lunch* and *breakfast* however are classified as polysemous - the latter through the multimodality criterion - together with *tea* (Table C.11). The issue with *lunch* seems to be a bad representation of the *food* sense in the reference vector, as a large part of the samples constituting this vector do appear to exhibit an *event* rather *than* a food reading. This causes target words' *event* samples to also obtain high similarity scores with the *food* reference vectors, and skews the entire distribution into the *food* direction. The reference samples for dinner seem to create more representative sense embeddings, correctly identifying the polysemy in *lunch* and *breakfast.* While *tea* arguably also allows for an *event/food* alternation, the majority of *food* samples here seem to contain compound nouns referring to a variety of concepts, indicating that the heuristics does detect a alternation in meaning, but not necessarily that which we intended the reference vectors to represent:

(84) ***Tea* samples most similar to the *dinner event* reference embedding**

a. The site had a staff restaurant, **tea bar**, games room, and licensed bar.

b. They are pale brown, **weak tea colour** above and whitish below with buff flanks.

c. The small sub-divisional town has scenic beauty and is surrounded by hills, **tea gardens**, forests and rivers.

d. There is a shop, **tea room**, car park and disabled access.

e. Preferably a clump of long grass, **tea tree branches** or pile of loose herbage should be provided.

One might also expect *supper* to exhibit an *event/food* alternation, and upon inspection a good portion of the identified ‘extreme’ samples indeed do invoke either of these readings. The corpus sample size for *supper* however was relatively small, and contained only a total of 277 sentences. This small sample size lead to a relatively wide standard deviation, and as a result did not provide sufficient data to our pilot heuristics to correctly label this target word. Including additional or alternative corpora of natural language use in the corpus sample selection stage will very likely mitigate this issue.

Physical/Aperture* Alternation: *Door* and *Window From the substitutes predicted for *door*, only *gate* was labelled polysemous with respect tot the *physical/aperture* alternation investigated in our annotation study, and *window* did not yield a single polysemic target (Tables C.12 and C.13. Since the substitute lists however contain both *window* and *door*, the heuristics seems to produce false negatives for these seminal examples. While both are classified as having a dominant sense in their *physical* reading, samples most similar to the *aperture* reference vectors also do contain (a majority of) *aperture* readings - these samples however are outweighed by the *physical* samples after our double-filtering approach, leading the heuristics to postulate that the subordinate reading is not available.

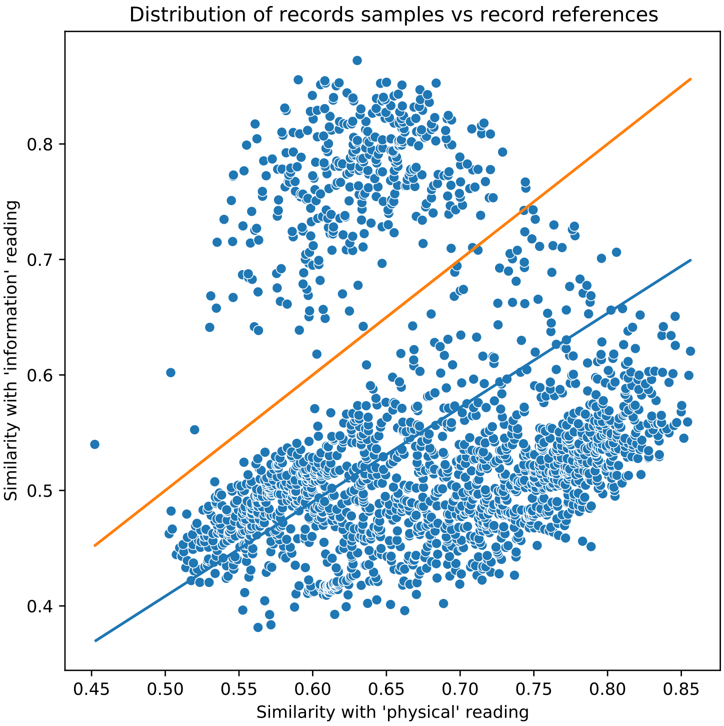


Figure 6.19: Distribution of the filtered *records* sample embeddings' similarity scores relative to the *record physical* reference embedding on the x-axis, and the *information* reference embedding on the y-axis. True diagonal in orange.

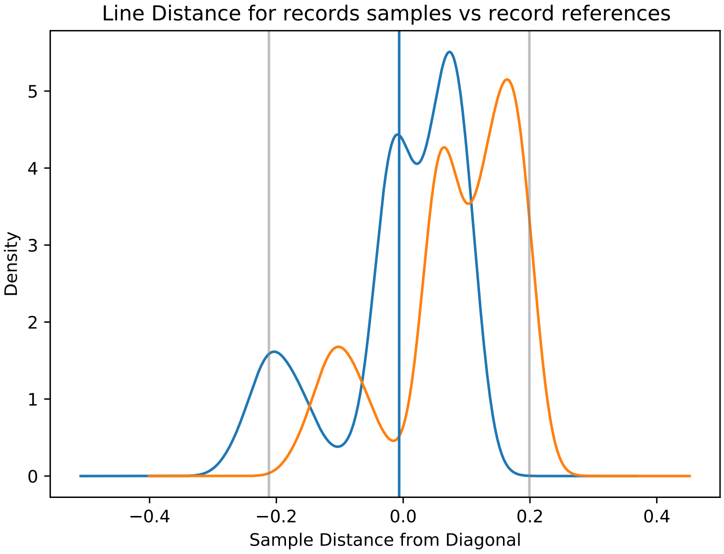


Figure 6.20: Distribution of *box* sample similarity scores' distances to the true diagonal (orange) and the best fit through the origin (blue). Mean of the distribution indicated in blue, 2 standard deviation range in grey.

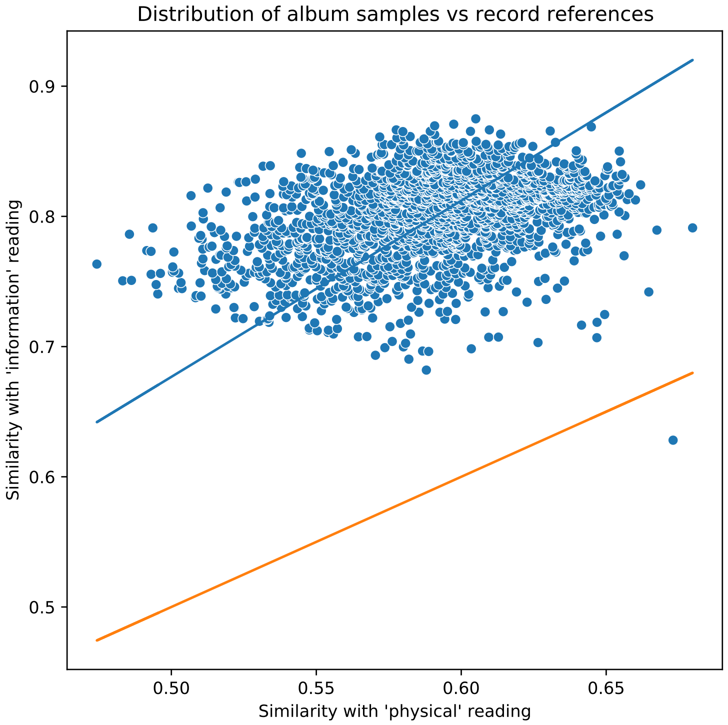


Figure 6.21: Distribution of the filtered *album* sample embeddings' similarity scores relative to the *record physical* reference embedding on the x-axis, and the *information* reference embedding on the y-axis. True diagonal in orange.

Physical/Information* Alternation: *Book* and *Record Out of *book*'s substitutes, only *film* was classified as having both, a *physical* and *information* reading, with *novel*, *album* and potentially *volume* constituting false positives (Table C.14). *Record* produced two substitutes classified as polysemous: *records* and *recordings* - the former of which through the multi-modality criterion (Table C.15 and Figures 6.19 and 6.20). In this case, the classifier missed labelling *album* to be polysemous as well - very like due to the extreme over-representation of *information* readings in the corpus data, as indicated by the scatter plot in Figure 6.21 (best fit slope = 1.35).

On Alternations with more than Two Readings: *Newspaper* The pilot heuristics presented here is formulated to compare sample sentences against two reference vectors, i.e. testing for polysemy with respect to two alternations or sense dimensions. As we saw in the case of *duck*, where a given target words also allows for other, unrelated interpretations, these are expected to receive low similarity scores in comparison with both reference embeddings, and filtered out in the first filtering step. Some of the words investigated in the human annotation experiments however are supposed to allow for three or more polysemic alternations. While the pilot

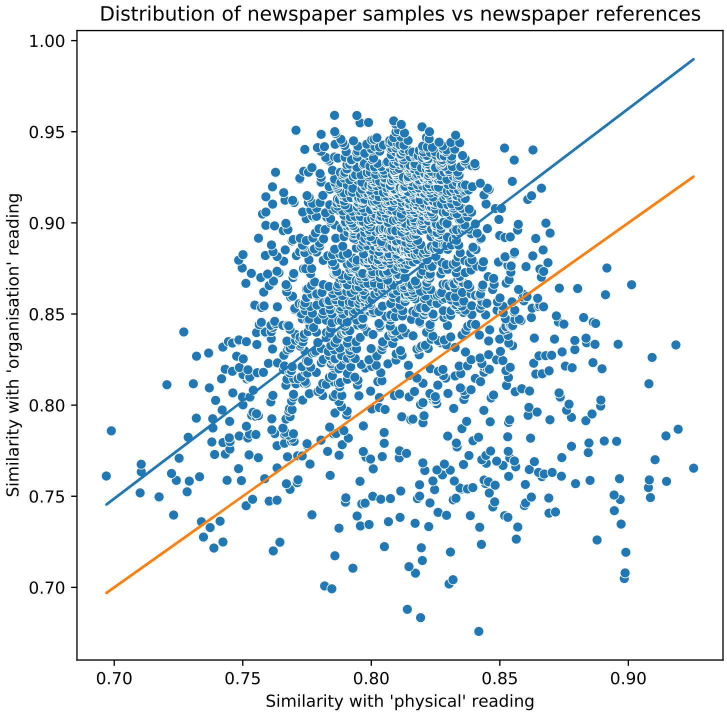


Figure 6.22: Distribution of the filtered *newspaper* sample embeddings' similarity scores relative to the *physical* reference embedding on the x-axis, and the *organisation* reference embedding on the y-axis. True diagonal in orange.

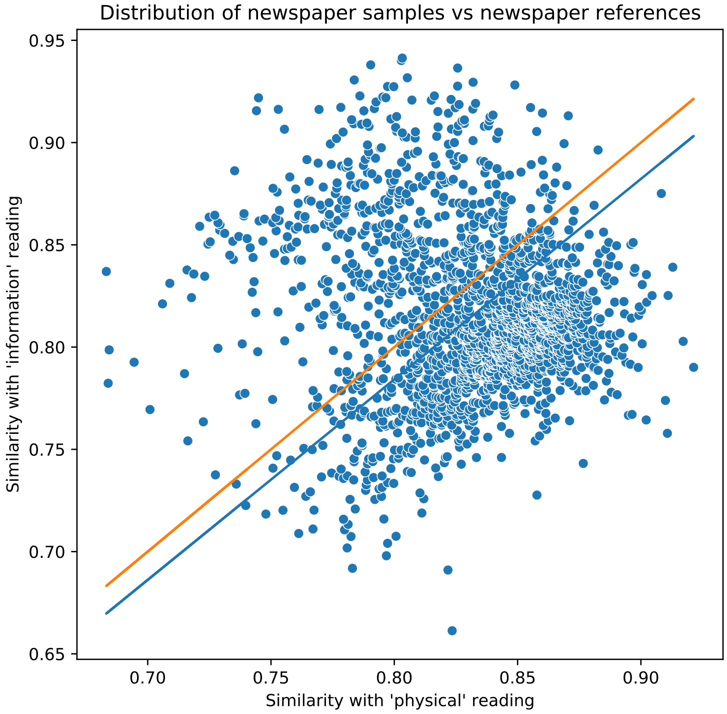


Figure 6.23: Distribution of the filtered *newspaper* sample embeddings' similarity scores relative to the *physical* reference embedding on the x-axis, and the *information* reference embedding on the y-axis. True diagonal in orange.

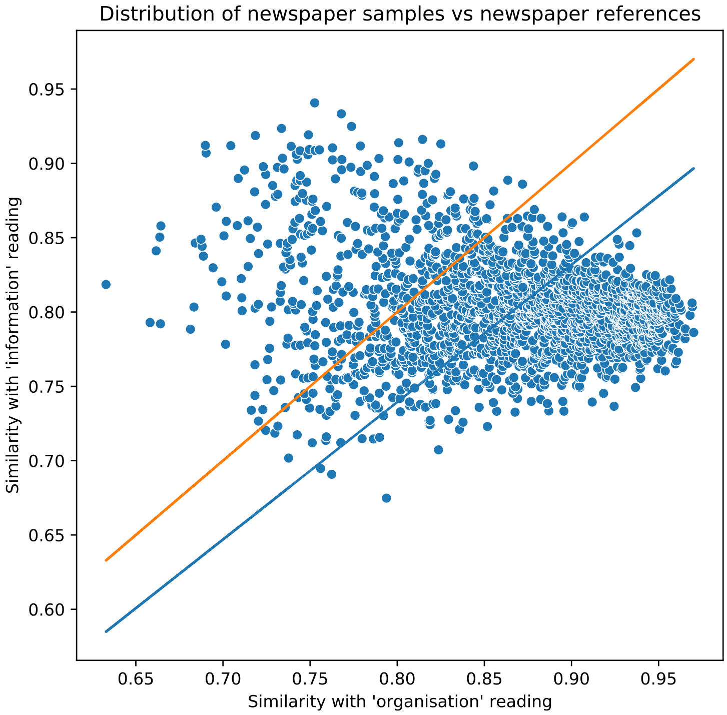


Figure 6.24: Distribution of the filtered *newspaper* sample embeddings' similarity scores relative to the *organisation* reference embedding on the x-axis, and the *information* reference embedding on the y-axis. True diagonal in orange.

heuristics currently cannot process these multiple dimensions simultaneously (and we do not see a principled reason why this should not be possible), it can investigate the different pairings of sense alternations sequentially. Figures 6.22 through 6.24 show the distribution of *newspaper* corpus samples given the three different sets of pairwise sense combinations: *physical/organisation*, *physical/information* and *organisation/information*. According to the heuristics, *newspaper* displays polysemy with respect to each of these alternations, with the *organisation* reading being more dominant than the *physical* and *information* readings, and no clear dominance in the comparison of *physical* and *information* samples.

Tables C.16 through C.18 show the target classification reports for *newspaper* substitutes given the three different sense combinations. When considering the *physical/organisation* alternation, *magazine*, *paper* and *journal* are flagged as polysemous. In the *organisation* samples of *paper*, the word is predominantly used as a short form for *newspaper*, rendering it a valid candidate for polysemy. *Publication* seems to fail the polysemy test due to missing *physical* readings in the corpus samples, similarly so for *daily* and *weekly*, which usually appear as parts of newspaper names in the corpus samples (*local daily*, *business daily*, *Ghanaian daily*). With respect to the *physical/information* alternation, again *magazine*, *paper* and *journal* are classified as polysemous - this time together with *news*. Considering finally the *organisation/information*, *magazine*, *paper* and *journal* are joined by *press*, which gains an information reading when used in compound nouns such as *press reports*, *press accounts* or *press interview*.

6.2.5 Further Iterations

The results presented so far have been derived from a single iteration of our polysemy detection algorithm, based on the annotated sample sentences from our initial set of experiments. Together, 30 new words were identified as potentially polysemous, including *fish*, *duck* and *goat* for the *animal/food* alternation, *liquor*, *ale*, *coffee* and *drink* for the *content-for-container* alternation, *breakfast* and *tea* for the *event/food* alternation and *journal* and *paper* for the three-way *physical, information, organisation* alternation exhibited by *newspaper* and *magazine*. These and all other targets flagged for polysemy now can be used to bootstrap a second iteration of the algorithm by using their corpus samples to generate BERT substitute predictions, and classify these according to the reference embeddings generated from a selection of the previously determined 'extreme' corpus samples best representing the different sense interpretations. Any new target words identified as potentially polysemous again can be added to the dataset and kickstart the next iteration of the algorithm by following the same steps. This process can be run fully automated and does not require any additional input or analysis, except for possibly a final validation of the proposed extended list of polysemes identified for the different alternations under investigation.

6.3 Discussion

In this chapter we presented a pilot for an iterative algorithm to automatically detect polysemy in corpus samples based on an initial set of reference words. The algorithm uses BERT's masked token prediction to generate a list of potential substitutes for a known polyseme, collects corpus samples containing those substitutes, and utilises an unsupervised heuristics to analyse if these target samples exhibit a reference word's polysemic alternation by comparing the similarity of its contextualised word embeddings to reference sense embeddings derived from the reference's corpus samples. In a trial run, a first iteration of the algorithm produced a list of 30 additional potentially polysemic target words given an initial set of 15 reference words exhibiting 8 different types of alternations. Our evaluation shows that most of the proposed targets indeed seem to exhibit the same type of alternation as the reference they were derived from, providing a proof of concept for the approach presented here. The heuristics however also appears to produce a number of false positives and false negatives, and requires for multiple smaller and larger adjustments and modifications to be implemented before the algorithm should be applied to bootstrap the collection of a corpus of polysemic word use.

As our analysis already covered a range of issues identified with the formulation of the algorithm and especially classification heuristics as presented here, we will here only briefly mention central issues and limitations of the pilot algorithm, and how they could be addressed by future work.

6.3.1 Limitations of the Proposed Target Identification Method

A first, principled limitation of the presented polysemy detection algorithm is that it can only detect polysemy with respect to the alternation exhibited by a given reference word, and thus is limited to extending the set of words allowing for a given sense extension rather than identifying new alternations. The algorithm is formulated to utilise the regularity of the ten types of polysemic alternation investigated in Chapters 4 and 5 to identify other words that allow for the same alternation by comparing them to the reference words' sense embeddings. Since the algorithm however is iterative, there is a chance of these reference embeddings deviating from representing their initial sense alternation as they will be re-formulated after each iteration based on a new target word's corpus sample. First evidence of this shift in representation can be observed in the case of *rice* discussed in the previous section, where the majority of samples flagged as most similar to *chicken*'s *animal* reading actually refer to rice as a plant or a precursor to producing a consumable. When in the second iteration of the algorithm now *rice* is used as a reference word, and sense embeddings are derived from its corpus samples, the new reference vectors will effectively no longer represent the initial *food/animal* alternation, but more likely a *crop/food* alternation. While the algorithm will automatically predict substitutes for this alternation and classify targets according to it, the labels applied to the alternation will no longer be accurate[11] and might need to be re-considered in a final evaluation of the algorithm's results. Alternatively, other types of polysemic alternation can be investigated by simply adding a number of reference words exhibiting that alternation to the list of words to be processed by the algorithm.

A second issue with the algorithm's target identification step is linked to its utilisation of corpus samples to propose substitute words. As we mentioned before,

[11] As mentioned earlier, there is an argument to be made that in this case they were not accurate in the first place, either.

the only restrictions on selecting corpus samples for a given reference word were that they i) had to include the reference word, ii) the reference word had to have the same word type in the sample sentence as in the original experiment sample, and iii) that the corpus sample should not be longer than 25 words. As a result, the collected slice of up to 2048 sentences will represent the corpus' statistics with respect to meaning and sense dominance, which means that for words whose polysemic alternations are far less common than a more dominant homonymic interpretation, the corpus sample will largely represent that interpretation. We previously gave the example of *glass*, which in our experiments was used to refer to either a container of liquids or its contents. The corpus samples extracted from Wikipedia to represent *glass* however did so predominantly in its interpretation as a building material, leading BERT to predict only other building materials as potential substitutes. In a similar fashion, all predictions for *cup* referred to its reading as a sports trophy (*championship*, *league*, *tournament*, *competition*, *title*, *trophy*) rather than its reading as a container for liquids (and a potential polysemic alternation referring to its content). One way to improve this corpus-based collection of target words could be including a second filtering step when parsing the corpus for samples: instead of directly adding a sample sentence containing the target word, we could compare the contextualised BERT encoding of the reference word in that sample sentence with the encodings of the target in our reference samples. Given that BERT seems sensitive enough to at least clearly tell apart homonymic usages of the same word, a threshold on the cosine similarity between the sample and reference embeddings could be used to only retain those samples that invoke the same *meaning* as the reference samples - independent of the specific *sense* invoked in the sample.

Another option would be to include other sources of natural language use besides the Wikipedia corpus used in the pilot presented here. While providing well-structured sentences, the texts included in the Wikipedia corpus have a clear objective and focus on providing encyclopedic information. For example, many of the sample sentences containing target *beer* introduce or simply list beer manufacturers rather than mentioning people drinking beer - something we would expect to be the more frequent use of the term in everyday language use. This also contributes to the issue mentioned above, with *glass* samples referring mostly to construction materials, and *cup* almost exclusively invoking a sports reading. Corpora of less formal and more casual language use could help to counter this bias and provide a more balanced selection of samples, facilitating subsequent evaluations and analyses.

6.3.2 Limitations of the Proposed Target Evaluation Method

Since the evaluation of the polysemic potential of a given target word is based on whether its corpus samples show high similarity scores with respect to both sense embeddings derived from the reference corpus samples, the quality of these reference embeddings has a significant impact on the quality of the labelling. As mentioned in the previous section, in some cases a sense reference will be of low quality (i.e. misrepresenting a given sense alternation) because the selected corpus sample does not contain the necessary samples to construct a reference embeddings for a certain sense. This seems to be the case for for example *bottle*, for which the corpus sample does not appear to contain sentences invoking its content reading, and the 20 contextualised embeddings included in the sense reference - even though closest to the *content* sample sentence from our materials - do not actually invoke the content reading of the reference word. In other cases, the reference sample used to identify the 20 samples to be included in a sense embedding does not appear good enough to actually identify relevant sentences in the corpus sample. We observed this for example for *pheasant*, where the sentences closest to the *chicken food* reading used *pheasant* in a food sense, while the sentences included in *pheasant*'s own *food* reference vector largely did not - indicating that the reference sentence used to determine these sentences did not provide high similarity scores with the relevant corpus samples. The first issue can be addressed by using a different or expanded selection of corpora of natural language use to select samples from. We suspect that a more conversational corpus is much more likely to contain sentences using for example *bottle* in its *content* sense, or *glass* in its reading as a container. The second issue can be addressed by providing a larger reference sentence set. While this is less of an issue in later iterations of the algorithm as here a larger number of reference sentences are available already, this is mostly an issue in the first iteration where currently just a single sentence is used to kickstart the construction of a sense reference vector from the corpus. Providing the algorithm with a set of reference sentences instead would provide a more robust signal of the sense that is meant to be represented, and will likely lead to the selection of more relevant and representative samples to be used in the reference sense embedding. At this point we might also want to consider excluding compound nouns, as these seem to capture a slightly different phenomenon but skew the distribution of senses when classified through the pilot heuristics.

Other, less principled improvements and adjustments can be applied to the classification heuristics itself. A wide distribution in the first filtering step for example appears to sometimes hinder the selection of a relevant ‘slice’ of samples for the subsequent steps, especially when their distribution does not meet UniDip’s multi-modality criterion, which allows us to safely discard a large portion of irrelevant samples. A more lenient threshold in the application of UniDip at this point might address this issue - as might using a simpler, less noise-aware approach to determining multi-modality. Similarly, UniDip seems to flag multi-modality in the line distance distributions less often than we might expect, and a lower threshold here might lead to the detection of more polysemic targets through this first criterion.

Chapter 7

Conclusion

Traditionally, different uses of an ambiguous word are either taken to refer to the same or a different *meaning* of a word, or to the same or a different *sense*. Based on accumulating evidence that phenomena of homonymy and polysemy might not be as homogeneous as this traditional view assumes, in this thesis we investigated the notion of distance between the different interpretations of ambiguous word forms. In particular, we explored graded annotations of explicit word sense similarity and implicit co-predication acceptability judgements as measures of interpretation distance, comparing similarity and acceptability scores of polysemic sense alternations and homonymic meaning extensions.

Over the course of two crowd-sourced annotation runs, we collected a total of close to 18,000 similarity and acceptability judgements for custom-made samples invoking different interpretations of ambiguous word forms. The data collected for both measures suggest that the perception of meaning is not an exclusive, binary decision, but might be in fact more gradual in nature, with judgements for polysemic alternations covering the entire spectrum between word sense identity and unrelatedness in meaning. Both ratings show significant differences in the overall means of same-sense and cross-sense judgements, indicating that polysemous extensions are not always perceived as invoking identical interpretations, but also contain cases of perceived identity of sense. Investigating judgements for individual target words, we observed significant drops in the similarity or acceptability ratings for a select number of sample combinations, showing that those samples specifically invoke sense combinations that are clearly distinguished by the annotators. In some cases, these polysemic cross-sense samples can obtain ratings as low as the homonymic test items included in the study.

The collected data provides intriguing empirical evidence challenging traditional models of mental word sense representation. We suggest that observations of significant similarity differences between different polysemic senses are difficult to explain when assuming a fully under-specified mental representation of polysemic sense: if all interpretations were to be stored in the same, unstructured entry, we would not expect participants to clearly distinguish their interpretations. Because all of the senses stored in an under-specified entry should allow for cost-free sense switching and be co-activated, finding evidence of perceived differences in meaning indicates that the mental representations of these senses are likely more structured than assumed by one representation models. On the other hand, some cross-sense polyseme readings do receive similarity ratings and co-predication acceptability ratings close to those or exceeding those of same-sense items. This observation suggests that in the processing of some polysemic senses, no distinction is made in their interpretation - even though the invoked senses are not identical. While this is in line with the assumptions of one representation models, it is a challenging finding for sense enumeration approaches, which in these cases will struggle to specify the necessary contrast and selection criteria to warrant separate entries for the invoked senses.

The collected data however fits in well with recent proposals of a more structured mental representation of polysemic sense (see e.g. Ortega-Andrés and Vicente, 2019), where word sense distance could be an underlying factor in determining the similarity of sense interpretations and their co-activation. Assuming that the collected word sense similarity and co-predication acceptability ratings are a proxy of the senses' distances in their mental representation, our data supports the potential of a distanced-based grouping of senses within an otherwise still under-specified entry. A model like this would allow for the co-activation of just a subset of sense interpretations (those that are grouped closest together), which allows for the observation of near identity ratings in their comparison, as well as significant differences in the interpretation of those senses that are represented in different groups or clusters - leading to observations of similarity ratings as low as those for homonymic controls.

In our study we focused on regular, metonymic polysemes, testing 28 target words exhibiting ten different types of polysemic alternations. Because the full dataset contains at least two target words for each type of alternation, it also allowed us to investigate potential patterns in the similarity and acceptability ratings collected. We found that for some alternations the differences in sense interpretations are relatively consistent across targets of the same type, while for others these patterns could not be established. We suggest that these observations even more underline the heterogeneity within different phenomena of polysemy, as not even all types of regular metonymic polysemy seem to exhibit consistent patterns in their sense similarity judgements. The data however also suggests that not every polysemic word allows for its own, idiosyncratic set of sense extensions, but that there is some potential for a classification or grouping of polysemic expressions.

Besides presenting an analysis of the collected human annotations, this thesis also investigated how well contextualised language models' predictions of word sense similarity correlate with the collected judgements, focusing primarily on the default implementations of contextualised language models such as ELMo and BERT in predicting the human annotations. We extracted contextualised word embeddings for target words within their respective context sentences, and considered the cosine similarity as a measure of their similarity. Especially the similarity scores calculated with BERT Large exhibit a good correlation with the collected human judgements of explicit word sense similarity, but does not consistently predict the same similarity patterns as observed in the annotations. Still, when using BERT Large's contextualised target embeddings to group context samples by the senses they invoke, a hierarchical clustering returns the expected grouping of senses in almost half of the ten polysemic alternations tested in this study. Together, we suggest that these observations indicate a promising potential of BERT Large providing comparatively cheap indications of nuanced word sense similarities in future work, especially so if optimised or fine-tuned for this task.

Finally, this thesis presented a pilot algorithm for automatically extending a list of reference polysemes by detecting additional words that allow for the same set of alternations as the given references. Formulated in an unsupervised fashion, the algorithm is intended to bootstrap the future collection of a large-scale dataset of ambiguous language use with potential applications in traditional linguistics as well as the specialised fine-tuning of computational language models. In its current form, the presented pilot provides a reliable proof of concept, successfully flagging polysemic substitutes for some of the tested reference words, but among its teething problems shows a strong dependence on the quality of the reference sentences specified to kickstart the sample collection procedure.

7.1 Limitations and Future Work

The collected empirical data based on annotator judgements can only say so much about the actual representation of word senses in the mental lexicon. We assume that perceived differences in sense similarity or co-predication acceptability are derived from and therefore indicate differences in their mental representation - but this is a link we have not set out to prove in this thesis. Similarly, what is suggested in a theoretical model of the mental processing of polysemes is unlikely to have a direct implementation in the actual brain activation patterns that underlie our language processor. Nonetheless, probing the way we process and understand language through these proxies can provide us with more and more reliable insights to aid our understanding of the vast and complex phenomenon of language.

Our data collection is based on a number of hand-built, custom samples rated by a fairly large number of annotators. While sound in and by itself as a methodology, it still is a limiting factor to this research: How well do our custom samples cover the phenomenon of polysemy? What aspects do they incorporate - and which ones do they omit or even misrepresent? In this thesis we focused on regular, metonymic polysemic nouns only, and invoked their interpretation in a very specific way. Research like Trott and Bergen (2021) shows an alternate approach to invoking meaning by presenting target words as parts of compound nouns (e.g. 'traffic cone' vs 'ice cream cone') - and do obtain results quite different from ours. Verbs have been shown to be far more productive in the number of senses they can invoke. Do our observations also hold for them? Where exactly is the line between sense alternation and context coercion that is likely to affect those words even more strongly than the nouns tested in our study? Research on polysemy is far from exhaustive, and we hope that future work can help us answer some of these questions. Still, we hope that our highly controlled samples and systematic testing of polysemy patterns is a strong contribution to understanding the complexity of processing ambiguous word forms, and through the extent and amount of tested alternations withstands the risk of cherry-picking potential non-representative examples that might have informed some earlier arguments in the debate on the mental lexicon.

With respect to our evaluation of contextualised language models, we can pinpoint the most direct opportunities for future research: this thesis developed in parallel to the development of transformer models - and the ensuing explosion in optimisation approaches and availability of transformer variants dedicated to certain tasks, issues or linguistic phenomena. Our initial investigation still was based on the - practically ancient and obsolete - ELMo, and there is a likely sizeable potential improvement in our prediction scores in selecting different BERT variants, optimising the extraction of word or even sense representations from its hidden states, final layers and pooled outputs, and even fine-tuning models on similar tasks to leverage their transfer learning and few-shot prediction abilities. We hope that future work on these issues will highlight even more the importance of the data collected in this thesis - as a resource for either training or evaluation - and will see a refinement of the detection heuristics introduced in Chapter 6, yielding an application-ready algorithm useful for a number of research communities.

Finally, this thesis covers literature from linguistics, neuroscience, psycho-linguistics and computational linguistics, disciplines that often appear to work on similar issues without much interaction. We hope that this effort will help to identify untapped potential for increased collaboration between these different approaches, and provide a solid starting point from which to set out for even more impactful joint research on polysemy.

Bibliography

Almuhareb, A. and Poesio, M. (2006). Msda: Wordsense discrimination using context vectors and attributes. In *Proceedings of the 2006 Conference on ECAI 2006: 17th European Conference on Artificial Intelligence August 29 – September 1, 2006, Riva Del Garda, Italy*, page 543–547, NLD. IOS Press.

Amrami, A. and Goldberg, Y. (2018). Word sense induction with neural biLM and symmetric patterns. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4867, Brussels, Belgium. Association for Computational Linguistics.

Amrami, A. and Goldberg, Y. (2019). Towards better substitution-based word sense induction. *ArXiv*, abs/1905.12598.

Anderson, R. C. and Ortony, A. (1975). On putting apples into bottles — a problem of polysemy. *Cognitive Psychology*, 7(2):167–180.

Antunes, S. and Chaves, R. P. (2003). On the Licensing Conditions of Co-Predication. In *Proceedings of the 2nd International Workshop on Generative Approaches to the Lexicon.*

Apresjan, J. D. (1974). Regular polysemy. *Linguistics*, 12:5–32.

Arapinis, A. (2013). Referring to institutional entities: Semantic and ontological perspectives. *Applied Ontology*, 8:31–57. 1.

Arapinis, A. and Vieu, L. (2015). A plea for complex categories in ontologies. *Applied Ontology*, vol. 10(n{\textdegree} 3-4):285–296.

Armendariz, C. S., Purver, M., Ulčar, M., Pollak, S., Ljubešić, N., and Granroth-Wilding, M. (2020). CoSimLex: A resource for evaluating graded word similarity in context. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5878–5886, Marseille, France. European Language Resources Association.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. (2016). Linear algebraic structure of word senses, with applications to polysemy. *CoRR*, abs/1601.03764.

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

Asher, N. (2011). *Lexical Meaning in Context: A Web of Words.* Cambridge University Press.

Asher, N. and Lascarides, A. (2003). *Logics of Conversation.* Cambridge University Press, United States.

Asher, N. and Pustejovsky, J. (2006). A type composition logic for generative lexicon. *Journal of Cognitive Science*, 6(1).

Banerjee, S. and Pedersen, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, page 136–145, Berlin, Heidelberg. Springer-Verlag.

Baroni, M. and Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, page 1183–1193, USA. Association for Computational Linguistics.

Barton, S. B. and Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & Cognition*, 21(4):477–487.

Belinkov, Y. and Glass, J. (2019). Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Bennett, A., Baldwin, T., Lau, J. H., McCarthy, D., and Bond, F. (2016). LexSemTm: A semantic dataset based on all-words unsupervised sense distribution learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1513–1524, Berlin, Germany. Association for Computational Linguistics.

Bentivogli, L., Bernardi, R., Marelli, M., Menini, S., Baroni, M., and Zamparelli, R. (2016). Sick through the semeval glasses. lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Language Resources and Evaluation*, 50(1):95–124.

Beretta, A., Fiorentino, R., and Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: an meg study. *Cognitive Brain Research*, 24(1):57–65.

Blank, A. (1997). *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen.* Max Niemeyer Verlag.

Blevins, T. and Zettlemoyer, L. (2020). Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234.

Boleda, G., Baroni, M., McNally, L., et al. (2013). Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013): long papers; 2013 Mar 20-22; Postdam, Germany. Stroudsburg (USA): Association for Computational Linguistics (ACL); 2013. p. 35-46.* ACL (Association for Computational Linguistics).

Bommasani, R., Davis, K., and Cardie, C. (2020). Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Bowdle, B. F. and Gentner, D. (2005). The career of metaphor. *Psychological Review*, 112(1):193–216.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Breal, M. (1897). *Essai de semantique (Science des significations).* Hachette Paris.

Brown, S. W. (2008). Choosing sense distinctions for WSD: Psycholinguistic evidence. In *Proceedings of ACL-08: HLT, Short Papers*, pages 249–252, Columbus, Ohio. Association for Computational Linguistics.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Brugman, C. (1988). *The story of over: polysemy, semantics, and the structure of the lexicon.* New York: Garland.

Brunner, G., Liu, Y., Pascual, D., Richter, O., Ciaramita, M., and Wattenhofer, R. (2020). On identifiability in transformers. In *International Conference on Learning Representations.*

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., and Pinkal, M. (2006). The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Camacho-Collados, J. and Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *J. Artif. Int. Res.*, 63(1):743–788.

Caramazza, A. and Grober, E. (1976). Polysemy and the structure of the subjective lexicon. *Georgetown University roundtable on languages and linguistics. Semantics: Theory and application*, pages 181–206.

Carston, R. (2002). *Thoughts and Utterances: The Pragmatics of Explicit Communication.* Oxford: Blackwell.

Carston, R. (2013). Word meaning, what is said, and explicature. In Penco, C. and Domaneschi, F., editors, *What is Said and What is Not.* Stanford: Csli Publications.

Chang, T.-Y. and Chen, Y.-N. (2019). What does this word mean? explaining contextualized embeddings with natural language definition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6064–6070, Hong Kong, China. Association for Computational Linguistics.

Chomsky, N. (1957). *Syntactic structures.* Syntactic structures. Mouton, Oxford, England.

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *Quarterly Journal of Experimental Psychology*, 69(5):817–828. PMID: 26785102.

Christianson, K., Hollingworth, A., Halliwell, J. F., and Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42(4):368–407.

Clark, S. (2015). *Vector Space Models of Lexical Meaning*, chapter 16, pages 493–522. John Wiley & Sons, Ltd.

Copestake, A. and Briscoe, T. (1995). Semi-productive polysemy and sense extension. *Journal of Semantics*.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Crain, S. and Steedman, M. (1985). *On not being led up the garden path: the use of context by the psychological syntax processor*, page 320–358. Studies in Natural Language Processing. Cambridge University Press.

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

Cruse, D. A. (1995). Polysemy and related phenomena from a cognitive linguistic viewpoint. In Saint-Dizier, P. and Viegas, E., editors, *Computational Lexical Semantics*, Studies in Natural Language Processing, page 33–49. Cambridge University Press.

Cruse, D. A. (2000). Aspects of the micro-structure of word meanings. In Ravin, Y. and Leacock, C., editors, *Polysemy: Theoretical and Computational Approaches*, pages 30–51. Oxford University Press.

Cruse, D. A. (2004). *Lexical 'facets': between monosemy and polysemy*, pages 25–36. Max Niemeyer Verlag.

Dai, A. M. and Le, Q. V. (2015). Semi-supervised sequence learning. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Dautriche, I. (2015). *Weaving an ambiguous lexicon*. PhD thesis, Université Sorbonne Paris Cité.

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. *Corpora*, 7(2):121–157.

Del Tredici, M. and Bel, N. (2015). A word-embedding-based sense index for regular polysemy representation. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 70–78, Denver, Colorado. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Durkin, K. and Manning, J. (1989). Polysemy and the subjective lexicon: semantic relatedness and the salience of intraword senses. *J Psycholinguist Res*, 18(6):577–612.

Dölling, J. (1995). Ontological domains, semantic sorts and systematic ambiguity. *International Journal of Human-Computer Studies*, 43(5):785–807.

Dölling, J. (2020). *Systematic Polysemy*, pages 1–27. John Wiley & Sons, Ltd.

Erickson, T. D. and Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, 20(5):540–551.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Erk, K. and McCarthy, D. (2009). Graded word sense assignment. In *EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009*.

Erk, K., McCarthy, D., and Gaylord, N. (2009). Investigations on word senses and word usages. In *ACL-IJCNLP 2009 - Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int. Joint Conf. on Natural Language Processing of the AFNLP, Proceedings of the Conf.*

Erk, K., McCarthy, D., and Gaylord, N. (2013). Measuring Word Meaning in Context. *Computational Linguistics*, 39(3):511–554.

Erk, K. and Padó, S. (2010). Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 92–97, Uppsala, Sweden. Association for Computational Linguistics.

Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).*

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, I. L. (2011). *The semantics and pragmatics of polysemy: a relevance--theoretic account.* PhD thesis, UCL (University College London.

Falkum, I. L. (2015). The how and why of polysemy: A pragmatic account. *Lingua.*

Falkum, I. L. and Vicente, A. (2015). Polysemy: Current perspectives and approaches. *Lingua*, 157:1–16. Polysemy: Current Perspectives and Approaches.

Fellbaum, C. and Miller, G. (1998). *Building Semantic Concordances*, pages 197–216.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47(2):164–203.

Ferreira, F., Bailey, K. G., and Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science.*

Ferreira, F., Christianson, K., and Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research*, 30(1):3–20.

Ferreira, F. and Patson, N. D. (2007). The 'Good Enough' Approach to Language Comprehension. *Language and Linguistics Compass.*

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis.*

Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong.* Oxford University Press.

Foraker, S. and Murphy, G. L. (2012). Polysemy in sentence comprehension: Effects of meaning dominance. *Journal of Memory and Language*, 67(4):407–425.

Frazier, L. and Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language.*

Fremann, L. and Lapata, M. (2016). A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Frisson, S. (2009). Semantic underspecification in language processing. *Linguistics and Language Compass.*

Frisson, S. (2015). About bound and scary books: The processing of book polysemies. *Lingua*, 157:17 – 35. Polysemy: Current Perspectives and Approaches.

Frisson, S. and Frazier, L. (2005). Carving up word meaning: Portioning and grinding. *Journal of Memory and Language*, 53(2):277–291.

Frisson, S. and Pickering, M. J. (1999). The Processing of Metonymy: Evidence from Eye Movements. *Journal of Experimental Psychology: Learning Memory and Cognition.*

Frisson, S. and Pickering, M. J. (2001). Obtaining a figurative interpretation of a word: Support for underspecification. *Metaphor and Symbol*, 16(3-4):149–171.

Garí Soler, A. and Apidianaki, M. (2021). Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Geeraerts, D. (1993). Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics*, 4(3):223–272.

Gilliver, P. (2013). Make, put, run: Writing and rewriting three big verbs in the oed. *Dictionaries: Journal of the Dictionary Society of North America*, 34:10–23.

Gillon, B. (1999). *The Lexical Semantics of English Count and Mass Nouns*, volume 10. Springer, Dordrecht.

Gillon, B. S. (1992). Towards a common semantics for english count and mass nouns. *Linguistics and Philosophy*, 15(6):597–639.

Giulianelli, M., Del Tredici, M., and Fernández, R. (2020). Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123(2):178.

Gotham, M. G. H. (2014). *Copredication, quantification and individuation.* PhD thesis, UCL (University College London).

Grefenstette, E. and Sadrzadeh, M. (2011). Experimenting with transitive verbs in a discocat. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, page 62–66, USA. Association for Computational Linguistics.

Haber, J. and Poesio, M. (2020a). Assessing polyseme sense similarity through co-predication acceptability and contextualised embedding distance. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 114–124, Barcelona, Spain (Online). Association for Computational Linguistics.

Haber, J. and Poesio, M. (2020b). Word sense distance in human similarity judgements and contextualised word embeddings. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 128–145, Gothenburg. Association for Computational Linguistics.

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.

Hanks, P. (2000). Do word meanings exist? *Computers and the Humanities*, 34(1/2):205–215.

Hao, Y., Dong, L., Wei, F., and Xu, K. (2020). Investigating learning dynamics of BERT fine-tuning. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 87–92, Suzhou, China. Association for Computational Linguistics.

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

Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Hobbs, J. R., Stickel, M. E., Appelt, D. E., and Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, 63(1):69–142.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780.

Hopper, P. J. (1991). On some principles of grammaticization. In *Approaches to Grammaticalization*. John Benjamins.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Hu, R., Li, S., and Liang, S. (2019). Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2016). Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany. Association for Computational Linguistics.

Jackendoff, R. (1989). What is a concept, that a person may grasp it? *Mind & Language*, 4(1-2):68–102.

Jackendoff, R. S. (1992). *Languages of the mind: Essays on mental representation.* Languages of the mind: Essays on mental representation. The MIT Press, Cambridge, MA, US.

Jain, S. and Wallace, B. C. (2019). Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Ježek, E. and Melloni, C. (2011). Nominals, polysemy, and co-predication. *Journal of Cognitive Science*, 12(1):1–31.

Ježek, E. and Vieu, L. (2014). Distributional analysis of copredication: towards distinguishing systematic polysemy from coercion. In *First Italian Conference* *on Computational Linguistics (CLiC-it)*, volume 1, pages 219–223, Pisa, IT. Pisa University Press.

Just, M. A. and Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Karimi, H. and Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *Quarterly Journal of Experimental Psychology.*

Katz, J. J. (1972). *Semantic Theory.* New York: Harper & Row.

Katz, J. J. and Fodor, J. A. (1963). The structure of a semantic theory. *Language*, 39(2):170–210.

Kempson, R. M. (1977). *Semantic Theory.* Cambridge University Press.

Kilgarriff, A. (1992). *Polysemy.* PhD thesis, University of Sussex.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Kilgarriff, A. (2001). English lexical sample task description. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 17–20, Toulouse, France. Association for Computational Linguistics.

Kitaev, N., Kaiser, L., and Levskaya, A. (2020). Reformer: The efficient transformer. *CoRR*, abs/2001.04451.

Klafka, J. and Ettinger, A. (2020). Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811, Online. Association for Computational Linguistics.

Klein, D. E. and Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language*, 45(2):259–282.

Klein, D. E. and Murphy, G. L. (2002). Paper has been my ruin: conceptual relations of polysemous senses. *Journal of Memory and Language*, 47(4):548–570.

Klepousniotou, E. (2002). The Processing of Lexical Ambiguity: Homonymy and Polysemy in the Mental Lexicon. *Brain and Language*, 81(1):205–223.

Klepousniotou, E., Pike, G. B., Steinhauer, K., and Gracco, V. (2012). Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and Language.*

Klepousniotou, E., Titone, D., and Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(6):1534–1543.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Lau, J. H., Clark, A., and Lappin, S. (2014). Measuring Gradience in Speakers' Grammaticality Judgements. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society (CogSci 2014).*

Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601, Avignon, France. Association for Computational Linguistics.

Laurence, S. and Margolis, E. (1999). Concepts and cognitive science. In Margolis, E. and Laurence, S., editors, *Concepts: Core Readings*, pages 3–81. MIT Press.

Lawrence, C., Kotnis, B., and Niepert, M. (2019). Attending to future tokens for bidirectional sequence generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, 4(1):151–171.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, page 24–26, New York, NY, USA. Association for Computing Machinery.

Levine, Y., Lenz, B., Dagan, O., Ram, O., Padnos, D., Sharir, O., Shalev-Shwartz, S., Shashua, A., and Shoham, Y. (2020). SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online. Association for Computational Linguistics.

Lewis, M., Ghazvininejad, M., Ghosh, G., Aghajanyan, A., Wang, S., and Zettlemoyer, L. (2020a). Pre-training via paraphrasing. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18470–18481. Curran Associates, Inc.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020b). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Lin, Y., Tan, Y. C., and Frank, R. (2019). Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., and Smith, N. A. (2019). Linguistic knowledge and transferability of contextual 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Logačev, P. and Vasishth, S. (2016). Understanding underspecification: A comparison of two computational implementations. *Quarterly Journal of Experimental Psychology*, 69(5):996–1012. PMID: 26960441.

Löhr, G. (2021). Does polysemy support radical contextualism? on the relation between minimalism, contextualism and polysemy. *Inquiry*.

Loureiro, D. and Jorge, A. (2019). Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.

Lyons, J. (1977). *Semantics*, volume 1. Cambridge University Press.

Mann, H. B. and Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2):209–220.

Marslen-Wilson, W. and Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, 8(1):1–71.

Maurus, S. and Plant, C. (2016). Skinny-dip: Clustering in a sea of noise. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1055–1064, New York, NY, USA. Association for Computing Machinery.

McCann, B., Bradbury, J., Xiong, C., and Socher, R. (2017). Learned in translation: Contextualized word vectors. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6297–6308, Red Hook, NY, USA. Curran Associates Inc.

McCarthy, D., Apidianaki, M., and Erk, K. (2016). Word Sense Clustering and Clusterability. *Computational Linguistics*, 42(2):245–275.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2004). Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 279–es, USA. Association for Computational Linguistics.

McCarthy, D. and Navigli, R. (2007). SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

Mickus, T., Paperno, D., Constant, M., and van Deemter, K. (2020). What do you mean, BERT? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290, New York, New York. Association for Computational Linguistics.

Mihalcea, R., Chklovski, T., and Kilgarriff, A. (2004). The senseval-3 English lexical sample task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28, Barcelona, Spain. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Miller, G. A., Chodorow, M., Landes, S., Leacock, C., and Thomas, R. G. (1994). Using a semantic concordance for sense identification. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Miller, G. A. and Johnson-Laird, P. N. (1976). *Language and Perception*. Harvard University Press.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A semantic concordance. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, page 303–308, USA. Association for Computational Linguistics.

Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Murphy, E. (2019). Acceptability properties of abstract senses in copredication. *Perspectives on Abstract Concepts: Cognition, language and communication*, 65:145.

Murphy, E. (2021). *Linguistic Representation and Processing of Copredication*. PhD thesis, UCL (University College London).

Nair, S., Srinivasan, M., and Meylan, S. (2020). Contextualized word embeddings encode aspects of human-like word sense knowledge. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 129–141, Online. Association for Computational Linguistics.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

Navigli, R., Litkowski, K. C., and Hargraves, O. (2007). SemEval-2007 task 07: Coarse-grained English all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35, Prague, Czech Republic. Association for Computational Linguistics.

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, B. and Clarke, D. D. (2003). *Polysemy and flexibility: introduction and overview*, pages 3–30. De Gruyter Mouton.

Norrick, N. R. (1981). *Semiotic Principles in Semantic Theory.* John Benjamins.

Nunberg, G. (1979). The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy*, 3(2):143–184.

Nunberg, G. (1995). Transfers of Meaning. *Journal of Semantics*, 12(2):109–132.

Ortega-Andrés, M. (2021). *Interpretation of Copredicative Sentences: A Rich Underspecification Account of Polysemy*, pages 111–132. Springer International Publishing, Cham.

Ortega-Andrés, M. and Vicente, A. (2019). Polysemy and co-predication. *Glossa: a journal of general linguistics*, 4(1).

Osman, N. (1971). *Kleines Lexikon untergegangener Wörter: Wortuntergang seit dem Ende des 18. Jahrhunderts*, volume 487. Beck.

Ostler, N. and Atkins, B. (1991). Predictable meaning shift: Some linguistic properties of lexical implication rules. In *Lexical Semantics and Knowledge Representation.*

Paradis, C. (2004). Where does metonymy stop? senses, facets, and active zones. *Metaphor and Symbol*, 19(4):245–264.

Paradis, C. (2011). Metonymization: A key mechanism in semantic change. In *Defining Metonymy in Cognitive Linguistics*, pages 61–88. John Benjamins.

Pasini, T., Scozzafava, F., and Scarlini, B. (2020). CluBERT: A cluster-based approach for learning sense distributions in multiple languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4008–4018, Online. Association for Computational Linguistics.

Paul, H. (2002). *Deutsches Wörterbuch*. Max Niemeyer Verlag.

Pedersen, T. and Bruce, R. (1997). Distinguishing word senses in untagged text. In *Second Conference on Empirical Methods in Natural Language Processing*.

Pelletier, F. J. (1975). Non-singular reference: Some preliminaries. *Philosophia*, 5(4):451–465.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

Peters, M. E., Ammar, W., Bhagavatula, C., and Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, Vancouver, Canada. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *CoRR*, abs/1802.05365.

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

Pickering, M. and Frisson, S. (2001). Processing ambiguous verbs: evidence from eye movements. *J Exp Psychol Learn Mem Cogn*, 27(2):556–573.

Pietroski, P. M. (2005). Meaning before truth. In Preyer, G. and Peter, G., editors, *Contextualism in Philosophy: Knowledge, Meaning, and Truth*. Oxford University Press.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Pinkal, M. (1985). *Logik Und Lexikon: Die Semantik des Unbestimmten.* De Gruyter.

Pinkal, M. (1995). *Logic and Lexicon. The semantics of the indefinite.* Kluwer Academic Publishers, Dordrecht.

Poesio, M. (2020). *Ambiguity*, pages 1–38. John Wiley & Sons, Ltd.

Pradhan, S., Moschitti, A., Xue, N., Ng, H. T., Björkelund, A., Uryupina, O., Zhang, Y., and Zhong, Z. (2013). Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2007). Ontonotes: A unified relational semantic representation. *International Journal of Semantic Computing*, 1(04):405–419.

Pustejovsky, J. (1993). Type Coercion and Lexical Selection. In Pustejovsky, J., editor, *Semantics and the Lexicon*, pages 73–94. Springer Netherlands, Dordrecht.

Pustejovsky, J. (1995). *The Generative Lexicon.* MIT Press.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding with unsupervised learning. *OpenAI Blog.*

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog.*

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. *Journal of Machine Learning Research*, 21(140):1–67.

Raganato, A., Camacho-Collados, J., and Navigli, R. (2017). Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ravin, Y. and Leacock, C. (2000). *Polysemy: An Overview*, pages 1–29. Oxford University Press.

Recanati, F. (1998). Truth-conditional pragmatics. In Kasher, A., editor, *Pragmatics: Critical Concepts*, pages 509–511.

Recanati, F. (2017). Contextualism and polysemy. *Dialectica*, 71(3):379–397.

Recasens, M., Hovy, E., and Martí, M. A. (2011). Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Reisinger, J. and Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, page 109–117, USA. Association for Computational Linguistics.

Rodd, J., Gaskell, G., and Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, 46(2):245 – 266.

Rodd, J. M., Gaskell, M. G., and Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, 28(1):89–104.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Salazar, J., Liang, D., Nguyen, T. Q., and Kirchhoff, K. (2020). Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Sandra, D. (1998). What linguists can and can't tell you about the human mind: A reply to croft. 9(4):361–378.

Schlechtweg, D., Schulte im Walde, S., and Eckmann, S. (2018). Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Schneider, N. (2014). *Lexical Semantic Analysis in Natural Language*. PhD thesis, Carnegie Mellon University.

Schumacher, P. (2013). When combinatorial processing results in reconceptualization: toward a new approach of compositionality. *Frontiers in Psychology*, 4:677.

Schuster, T., Ram, O., Barzilay, R., and Globerson, A. (2019). Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Schütze, H. (1998). Automatic word sense discrimination. *Comput. Linguist.*, 24(1):97–123.

Scott, D. W. (1992). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, New York, Chicester.

Serrano, S. and Smith, N. A. (2019). Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Simpson, G. B. (1981). Meaning dominance and semantic context in the processing of lexical ambiguity. *Journal of Verbal Learning and Verbal Behavior*, 20(1):120–136.

Simpson, G. B. (1994). Context and the processing of ambiguous words. *Handbook of psycholinguistics*, 22:359–374.

Snyder, B. and Palmer, M. (2004). The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain. Association for Computational Linguistics.

Srinivasan, M. and Rabagliati, H. (2015). How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, 157:124–152. Polysemy: Current Perspectives and Approaches.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune bert for text classification? In Sun, M., Huang, X., Ji, H., Liu, Z., and Liu, Y., editors, *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

Swets, B., Desmet, T., Clifton, C., and Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory and Cognition*.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, 18(6):645–659.

Tabossi, P., Colombo, L., and Job, R. (1987). Accessing lexical ambiguity: Effects of context and dominance. *Psychological Research*, 49(2):161–167.

Taieb, M. A. H., Zesch, T., and Aouicha, M. B. (2019). A survey of semantic relatedness evaluation datasets and procedures. *Artificial Intelligence Review*, pages 1–42.

Taylor, W. L. (1953). “cloze procedure”: A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.

Tenney, I., Das, D., and Pavlick, E. (2019a). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S. R., Das, D., and Pavlick, E. (2019b). What do you learn from context? probing for sentence structure in contextualized word representations. *ArXiv*, abs/1905.06316.

Traugott, E. C. (2017). Semantic change. *Oxford Research Encyclopedias, Linguistics.*

Travis, C. (1997). Pragmatics. In Hale, B. and Wright, C., editors, *A Companion to the Philosophy of Language*, pages 87–107. Blackwell.

Travis, C. (2008). *Occasion-sensitivity: Selected essays.* Oxford University Press.

Trott, S. and Bergen, B. (2021). RAW-C: Relatedness of ambiguous words in context (a new lexical resource for English). 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 7077–7087, Online. Association for Computational Linguistics.

Tuggy, D. (1993). Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, 4(3):273–290.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Vega Moreno, R. E. (2007). *Creativity and Convention: The pragmatics of everyday figurative speech.* John Benjamins.

Vial, L., Lecouteux, B., and Schwab, D. (2019). Sense vocabulary compression through the semantic knowledge of WordNet for neural word sense disambiguation. In *Proceedings of the 10th Global Wordnet Conference*, pages 108–117, Wroclaw, Poland. Global Wordnet Association.

Vicente, A. (2015). The green leaves and the expert: Polysemy and truth-conditional variability. *Lingua*, 157:54–65. Polysemy: Current Perspectives and Approaches.

Vicente, A. and Falkum, I. L. (2017). Polysemy. *Oxford Research Encyclopedias, Linguistics.*

Wallace, E., Wang, Y., Li, S., Singh, S., and Gardner, M. (2019). Do NLP models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019*

Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Weaver, W. (1955). Translation. *Machine translation of languages*, 14:15–23.

Weinreich, U. (1964). Webster's third: A critique of its semantics. *International Journal of American Linguistics*, 30(4):405–409.

Westera, M. and Boleda, G. (2019). Don't blame distributional semantics if it can't do entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 120–133, Gothenburg, Sweden. Association for Computational Linguistics.

Wiedemann, G., Remus, S., Chawla, A., and Biemann, C. (2019). Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Winkler, S. (2015). *Ambiguity: Language and Communication.* De Gruyter.

Wittgenstein, L. (1953). *Philosophische Untersuchungen - Philosophical investigations.* Macmillan.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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*, pages 156–162, Online. Association for Computational Linguistics.

Zhong, Z. and Ng, H. T. (2010). It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala, Sweden. Association for Computational Linguistics.

Zhuang, L., Wayne, L., Ya, S., and Jun, Z. (2021). A robustly optimized BERT pretraining approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of General Psychology*, 33(2):251–256.

Zwicky, A. M. and Sadock, J. M. (1975). Ambiguity tests and how to fail them. In *Syntax and Semantics volume 4*, pages 1–36. Brill.

Appendix A

Additional Materials for Chapter 4

A.1 Full Sample List

Two sample sentences were created for each interpretation of a target word, and all combinations of sample sentences were turned into test items for annotation. As an example, consider the six sample sentences for polyseme *newspaper*, two each for its three senses (1) *organisation/institution*, (2) *physical object* and (3) *information/content*:

Newspaper Senses: (1) *organisation*, (2) *physical object* and (3) *information/content*

1a The newspaper fired its editor in chief.,
1b The newspaper was sued for defamation.
2a The newspaper lies on the kitchen table.,
2b The newspaper got wet from the rain.
3a The newspaper wasn't very interesting.,
3b The newspaper is rather satirical today.

Combining these sample sentences resulted in the following list of test items included in our sense similarity annotation pilot:

1a1b organisation/organisation
The newspaper fired its editor in chief.,
The newspaper was sued for defamation.
1a2b organisation/physical
The newspaper fired its editor in chief.,
The newspaper got wet from the rain.

1a3b organisation/information

The newspaper fired its editor in chief.,

The newspaper is rather satirical today.

2a1b physical/organisation

The newspaper lies on the kitchen table.,

The newspaper was sued for defamation.

2a2b physical/physical

The newspaper lies on the kitchen table.,

The newspaper got wet from the rain.

2a3b physical/information

The newspaper lies on the kitchen table.,

The newspaper is rather satirical today.

3a1b information/organisation

The newspaper wasn't very interesting.,

The newspaper was sued for defamation.

3a2b information/physical

The newspaper wasn't very interesting.,

The newspaper got wet from the rain.

3a3b information/information

The newspaper wasn't very interesting.,

The newspaper is rather satirical today.

As a general rule, the first number always indicates the first sentence of that interpretation, and the second number the second sentence of that reading. Key 12 therefore is a short way of indicating sentence combination 1a2b. To shorten notation, we will use this key for the remainder of this publication.

Because our sample generation setup can introduce potential order effects, we also created a second validation set where all samples where presented in inverted order, with for example combination 1a2b becoming 2b1a (i.e. the first number now indicating the second sentence of a given sense and vice versa). As reported in Section 4.5.1 and 4.5.2, this yielded significantly different results only in one and 8 of the 67 comparisons, respectively. We therefore concluded that order effects were negligible in our analysis.

Hemingway Senses: (1) *person*, (2) *work*

1a Hemingway was born in Illinois.,

1b Hemingway won a Nobel prize.

2a Hemingway is still widely read today., 2b Hemingway is not suitable for children.

War and Peace Senses: (1) *work*, (2) *content*, (3) *physical copy*

1a War and Peace was finally published in 1869.,

1b War and Peace won a range of international awards.

2a War and Peace chronicles the period of 1805 to 1820.,

2b War and Peace describes a number of historic battles.

3a War and Peace gathers dust on the top shelf.,

3b War and Peace is bound in black embossed leather.

Lunch Senses: (1) *food*, (2) *event*

1a Lunch was exceptionally delicious today.,

1b Lunch got cold while we waited for someone.

2a Lunch took more than an hour yesterday.,

2b Lunch is great for socialising and networking.

Door Senses: (1) *physical*, (2) *aperture*

1a The door was turned into a table top.,

1b The door splintered when they hit it.

2a The door leads to a long hallway.,

2b The door connects the two rooms.

DVD Senses: (1) *physical copy*, (2) *content/information*, (3) *medium*

1a The DVD has some scratches but looks fine.,

1b The DVD got stuck in the player yesterday.

2a The DVD is a low resolution home movie.,

2b The DVD wasn't very entertaining somehow.

3a The DVD will be replaced by BluRay soon.,

3b The DVD has won the battle against VHR.

School Senses: (1) *building*, (2) *administration*, (3) *institution*, (4) *students/faculty*

1a The school was painted during the holidays.,

1b The school needs to be renovated soon.

2a The school requires students to wear a uniform.,

2b The school informed parents about this year's events.

3a The school is well respected among researchers.,

3b The school recently got a more modern website.

4a The school developed an important algebraic proof., 4b The school went on a field trip last summer.

Wine Senses: (1) *container*, (2) *content*

1a The wine lay in a padded wooden box.,

1b The wine is a little dusty from storage.

2a The wine had a beautiful red tint.,

2b The wine tastes great with fish.

Glass Senses: (1) *container*, (2) *content*

1a The glass broke when she dropped it.,

1b The glass fits about 200 ml of liquid.

2a The glass had a thick layer of foam.,

2b The glass was absolutely refreshing.

Construction Senses: (1) *process*, (2) *result*

1a The construction took far longer than expected.,

1b The construction will begin in early September.

2a The construction has a solid steel frame.,

2b The construction is larger than most in the city.

A.1.1 Homonym Samples

Besides these test items, the data collection pilot also included 15 homonym control pairs, one presented in each questionnaire:

0: bat,

The **bat** came in through the open window.,

The **bat** broke when he hit the fence with it.

1: match,

The **match** burned my fingers.,

The **match** ended without a winner.

2: club,

The **club** only admits women older than 50.,

The **club** felt very heavy and unwieldy.

3: bank,

The **bank** was washed out by the current.,

The **bank** increased the interest rate.

4: mole,

The **mole** dug tunnels all throughout the garden.,

The **mole** needs to be removed as it is cancerous.

5: pitcher,

The **pitcher** threw a number of perfect curveballs.,

The **pitcher** broke when the waiter dropped it.

6: rocket,

The **rocket** left the atmosphere at 2AM tonight.,

The **rocket** was bitter taste and ruined the pizza.

7: tank,

The **tank** could easily fit 500 litres of water.,

The **tank** could easily shoot further than 3 miles.

8: watch,

The **watch** slipped off his hand while he was swimming.,

The **watch** reported troop movements on the south border.

9: yard,

The **yard** equals exactly three feet.,

The **yard** is just over 10 feet wide.

10: stall,

The **stall** barely fit the large bull.,

The **stall** didn't have any toilet paper.

11: spring,

The **spring** in the garden feeds the little pond with fresh water.,

The **spring** in the ballpen lets you open it with a simple click.

12: mine,

The **mine** had to close after the accident.,

The **mine** could be defused by an expert.

13: order,

The **order** welcomed the new members.,

The **order** was shipped two weeks late.

14: jumper,

The **jumper** broke a long-standing record.,

The **jumper** didn't really fit her that well.

A.1.2 Synonym Samples

The data collection pilot also included 15 synonym control pairs, one presented in each questionnaire:

0: answer/reply,

The **answer** came after more than a month.,

The **reply** arrived within a couple of minutes.

1: street/road,

The **street** leads to a small town in the mountains.,

The **road** ends at a beautiful hut made from wood.

2: world/planet,

The **world** is heating up because of CO2 emissions.,

The **planet** is heading towards a serious climate crisis.

3: computer/PC,

The **computer** suddenly turned off.,

The **PC** needs to be replaced soon.

4: problem/issue,

The **problem** was solved by replacing a cable.,

The **issue** couldn't be resolved without tools.

5: capability/ability,

The **capability** of modern computers is astonishing.,

The **ability** to read and write is crucially important.

6: area/space,

The **area** was roped off by the police.,

The **space** was littered with rubbish.

7: audience/crowd,

The **audience** was very quiet during the concert.,

The **crowd** was cheering on the football team.

8: note/memo,

The **note** on the fridge read "clean me!".,

The **memo** simply said "Meeting at 1PM".

9: advice/tip,

The **advice** wasn't very good.,

The **tip** helped to fix the TV.

10: photo/image,

The **photo** was of a picturesque lake.,

The **image** shows a red muscle car.

11: building/structure,

The **building** burned down last week.,

The **structure** collapsed years ago.

12: company/organisation,

The **company** had to find a new office building.,

The **organisation** expanded to Eastern Europe.

13: plank/board,

The **plank** was torn out of the floor.,

The **board** covered up a crack in the wall.

14: sea/ocean,

The **sea** was much colder than the beach.,

The **ocean** looked beautiful in the sunset.

A.2 Visualisation of Human Judgements

As each of the target expressions tested in the data collection pilot represents a different type of polysemic alternation, collapsing data across different targets reduces the clarity of results. Figures A.1, A.2 and A.3 therefore show the collected data per target expression, indicating either a sample combination's mean explicit similarity score, mean co-predication acceptability rating or word sense class overlap.

A.3 Similarity Heat Maps

In order to make more intuitively accessible the differences in similarity ratings assigned to samples by human annotators and the different tested computational approaches, we visualised similarity scores per target expression in a min-max scaled heat map, indicating low similarity scores with darker colours and high similarity scores with brighter colours. Figure A.4 through A.6 contain these heat maps for all targets tested in the pilot run.

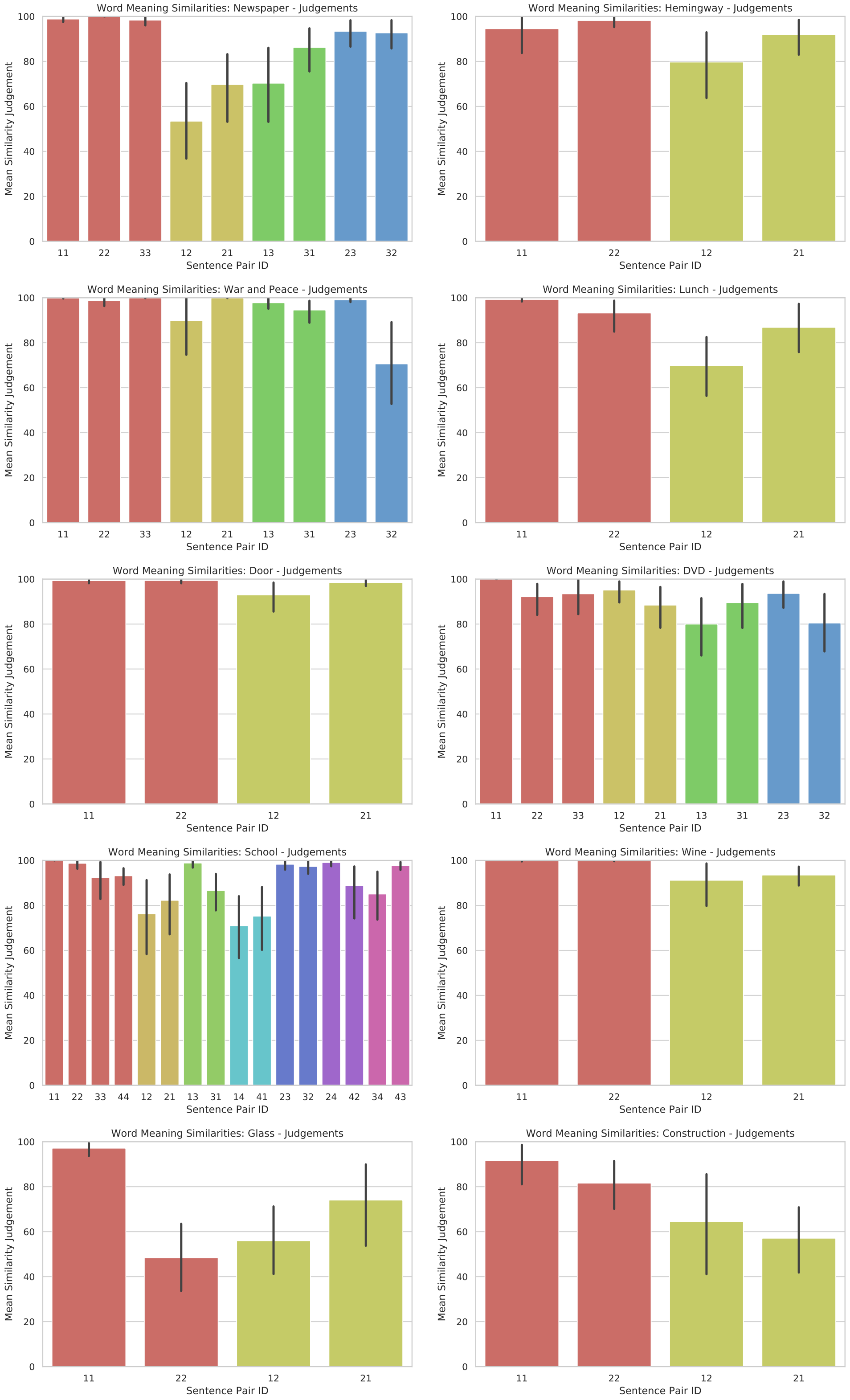


Figure A.1: Mean explicit word sense similarity judgements for the different combinations of same-sense and cross-sense samples of the ten tested types of regular polysemy.

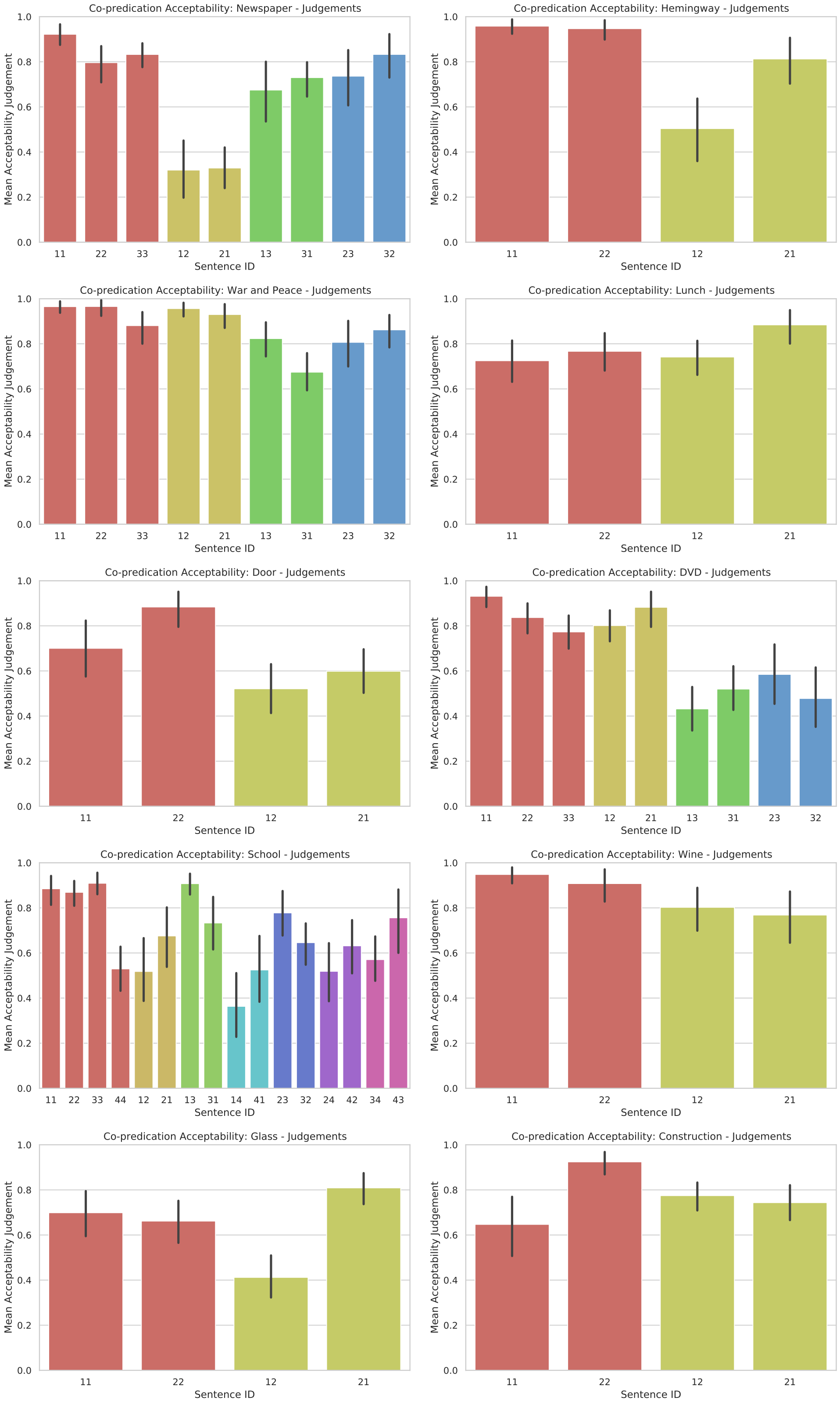


Figure A.2: Mean co-predication acceptability judgements for the different combinations of same-sense and cross-sense samples of the ten tested types of regular polysemy.

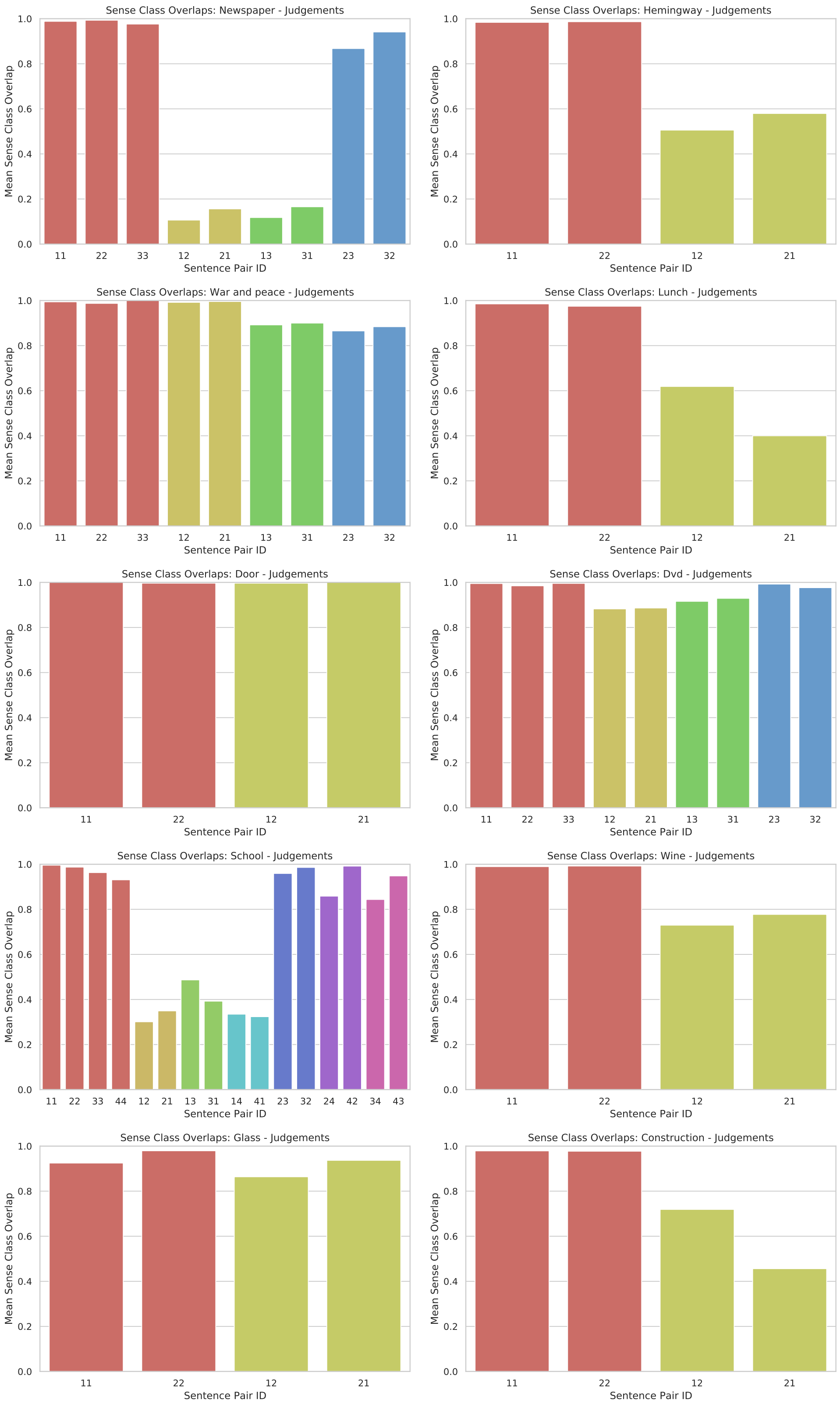


Figure A.3: Sense class overlap scores for the different combinations of same-sense and cross-sense samples of the ten tested types of regular polysemy.

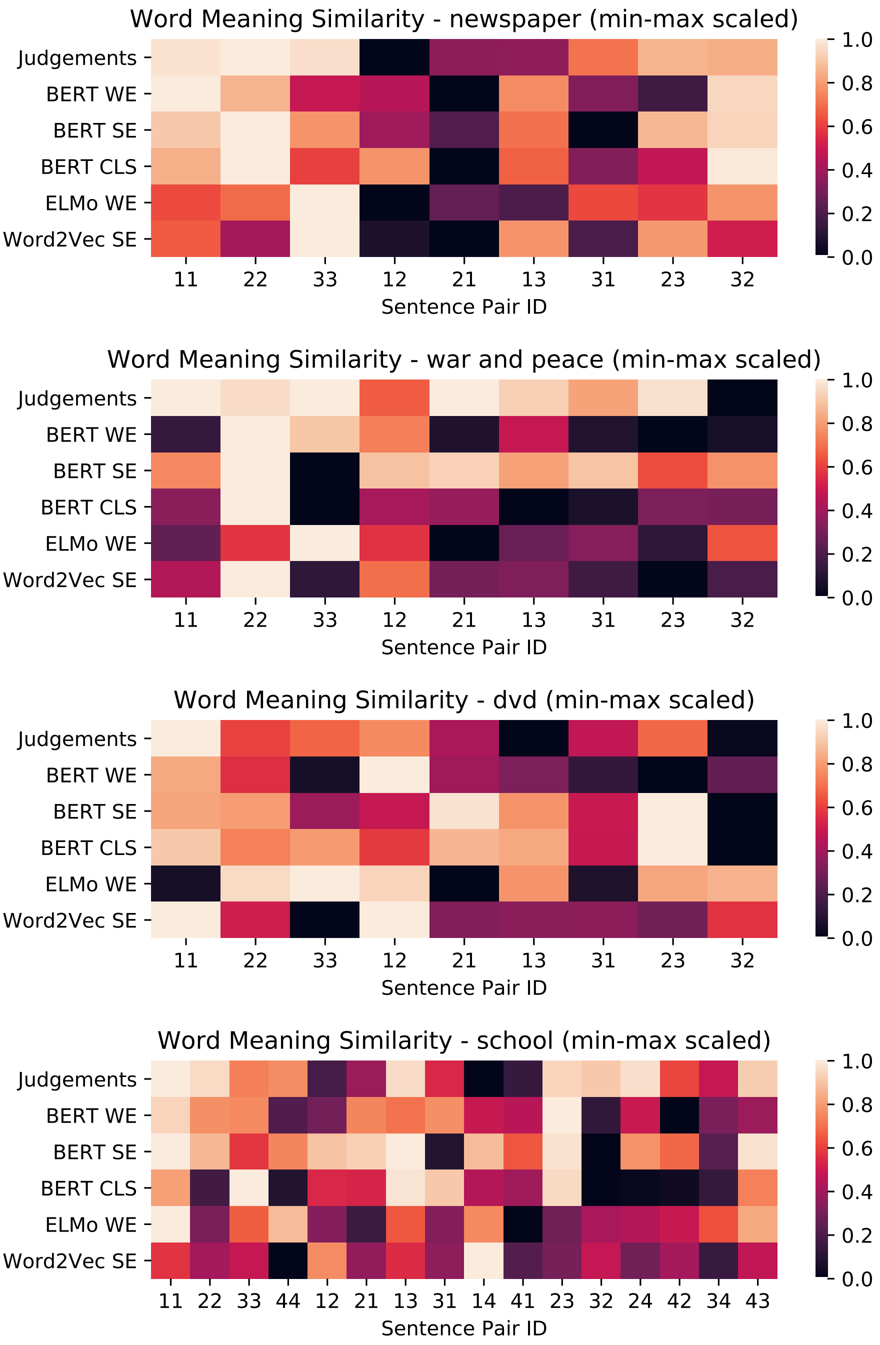
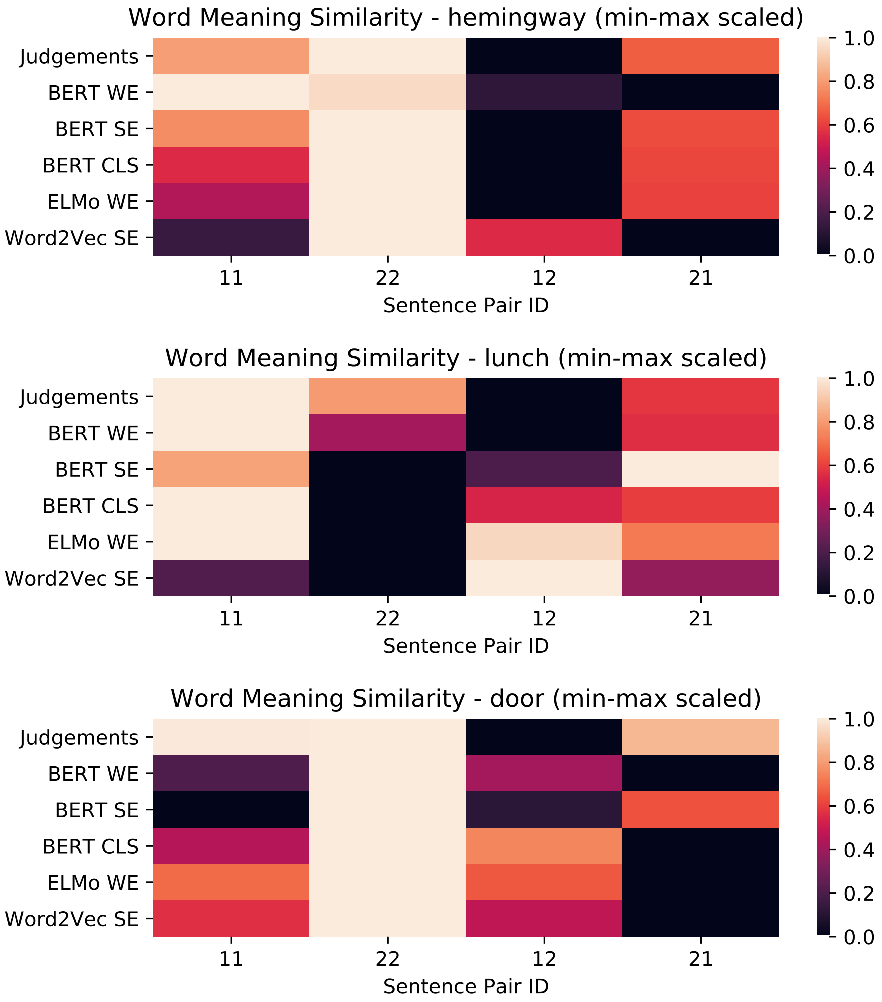


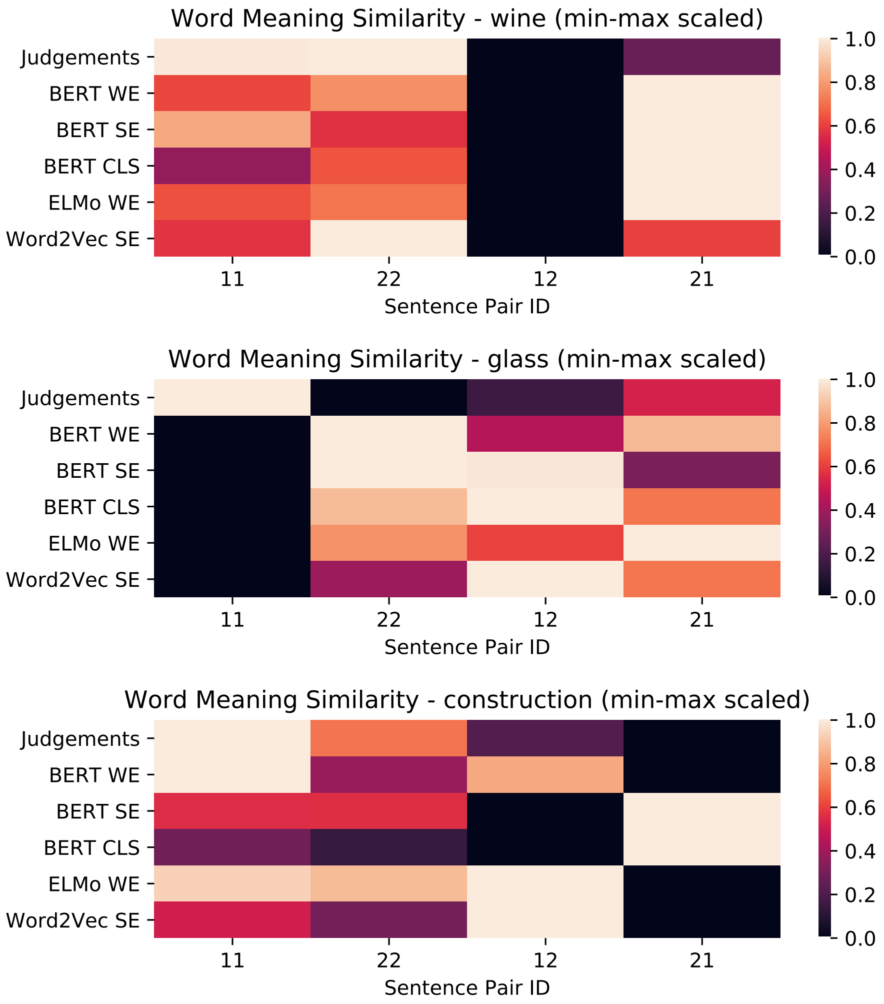
Figure A.4: Similarity scores assigned to the different sample combinations of target expressions with three or four sense extensions. Top rows show the mean explicit sense similarity judgement calculated based on the collected data, the following rows indicate similarity scores assigned by the different tested computational approaches. Scores are min-max scaled for clarity; low similarity scores with darker colours and high similarity scores with brighter colours.


Word Meaning Similarity - hemingway (min-max scaled)

Word Meaning Similarity - lunch (min-max scaled)

Word Meaning Similarity - door (min-max scaled)

Figure A.5: Similarity scores assigned to the different sample combinations of target expressions with two sense extensions. Top rows show the mean explicit sense similarity judgement calculated based on the collected data, the following rows indicate similarity scores assigned by the different tested computational approaches. Scores are min-max scaled for clarity; low similarity scores with darker colours and high similarity scores with brighter colours.



Word Meaning Similarity - wine (min-max scaled)

Word Meaning Similarity - glass (min-max scaled)

Word Meaning Similarity - construction (min-max scaled)

Figure A.6: Similarity scores assigned to the different sample combinations of target expressions with two sense extensions (continued)

Appendix B

Additional Materials for Chapter 5

B.1 Full Sample List

The materials for the second annotation run were created in the same way as in the pilot, with samples being combined as described in Section A.1. The full list of target samples used in the this study is as follows:

Newspaper Senses: (1) *organisation*, (2) *physical*, (3) *information*

1a The newspaper fired its editor in chief.
1b The newspaper was sued for defamation.
2a The newspaper lies on the kitchen table.
2b The newspaper got wet from the rain.
3a The newspaper wasn't very interesting.
3b The newspaper is rather satirical today.

Lunch Senses: (1) *food*, (2) *event*

1a Lunch was exceptionally delicious today.
1b Lunch got cold while we waited for someone.
2a Lunch took more than an hour yesterday.
2b Lunch is great for socialising and networking.

Door Senses: (1) *physical*, (2) *aperture*

1a The door was turned into a table top.
1b The door leads to a long hallway.
2a The door splintered when they hit it.
2b The door connects the two rooms.

DVD Senses: (1) *physical*, (2) *content*, (3) *medium*

1a The DVD has some scratches but looks fine.

1b The DVD got stuck in the player yesterday.

2a The DVD is a low resolution home movie.

2b The DVD wasn't very entertaining somehow.

3a The DVD will be replaced by BluRay soon.

3b The DVD has won the battle against VHR.

School Senses: (1) *building*, (2) *administration*, (3) *institution* (4) *students*

1a The school was painted during the holidays.

1b The school needs to be renovated soon.

2a The school requires students to wear a uniform.

2b The school informed parents about this year's events.

3a The school is well respected among researchers.

3b The school recently got a more modern website.

4a The school developed an important algebraic proof.

4b The school went on a field trip last summer.

Wine Senses: (1) *container*, (2) *content*

1a The wine lay in a padded wooden box.

1b The wine is a little dusty from storage.

2a The wine had a beautiful red tint.

2b The wine tastes great with fish.

Glass Senses: (1) *container*, (2) *content*

1a The glass broke when she dropped it.

1b The glass fits about 200 ml of liquid.

2a The glass had a thick layer of foam.

2b The glass was absolutely refreshing.

Construction Senses: (1) *process*, (2) *product*

1a The construction took far longer than expected.

1b The construction will begin in early September.

2a The construction has a solid steel frame.

2b The construction is larger than most in the city.

Magazine Senses: (1) *organisation*, (2) *physical*, (3) *content* (4) *storage*

1a The magazine lost a court battle against a former pop star.

1b The magazine got into serious money problems last year.

2a The magazine just kept falling off the small living room table.

2b The magazine was covered in paw prints after a cat sat on it.

3a The magazine featured a two-page poster of Justin Timberlake.

3b The magazine had a special report on David Guetta's world tour.

4a The magazine contained all kinds of defunct WW2 weaponry.

4b The magazine was originally designed for storing ballistic missiles.

CD Senses: (1) *physical*, (2) *medium*, (3) *content*

1a The CD had a hand-written label that was very difficult to read.

1b The CD was badly scratched because of the cheap case.

2a The CD has a much higher audio quality than the cassette tape.

2b The CD revolutionised how people listened to music at home.

3a The CD sounded like an eclectic mixture of different musical styles.

3b The CD was a lot more fun to listen to on headphones.

Photograph Senses: (1) *physical*, (2) *content*

1a The photograph was torn in two after a heavy fight last night.

1b The photograph looked like it was kept in a wallet for a long time.

2a The photograph showed a young couple next to a new car.

2b The photograph must have been taken sometime around 1975.

Book Senses: (1) *physical*, (2) *content*

1a The book had a leather dust jacket with embossed gold lettering.

1b The book had been used to prop open the old office door.

2a The book follows the adventures of the fictional captain Nemo.

2b The book was one of the first science-fiction stories ever to be written.

Chicken Senses: (1) *animal*, (2) *meat*, (3) *animal*$\oplus$

1a The chicken pecked for some food pellets in the new feeder.

1b The chicken sat on the roof of the coop all afternoon long.

2a The chicken was served with steaming hot potato wedges.

2b The chicken tasted like it had been marinated for at least 12 hours.

3a The chicken was bred by a well-known family of poultry breeders.

Pheasant Senses: (1) *animal*, (2) *meat*, (3) *animal*$\oplus$

1a The pheasant definitely gave the hunters a run for their money.

1b The pheasant was foraging for some seeds on a small clearing.

2a The pheasant was marinated in milk over night to make it tender.

2b The pheasant tasted much better than anything he'd ever eaten.

3a The pheasant had been shot by an experienced hunter with a long rifle.

Lamb Senses: (1) *animal*, (2) *meat*, (3) *animal*⊕

1a The lamb used to get stuck when trying to jump through the fence.

1b The lamb was fed with a milk bottle after its mother rejected it.

2a The lamb was seasoned to perfection by the head chef.

2b The lamb easily was my favourite dish at the banquet.

3a The lamb had been bred in the most animal-friendly way possible.

Seagull Senses: (1) *animal*, (2) *meat*, (3) *animal*⊕

1a The seagull kept circling over a food stall near the promenade.

1b The seagull stole a sandwich from an unsuspecting beachgoer.

2a The seagull definitely tasted better than anyone could have imagined.

2b The seagull had been roasted on a spit over a makeshift campfire.

3a The seagull was the only thing they were able to catch that day.

Dinner Senses: (1) *event*, (2) *food*

1a The dinner was the target of the celebrations yesterday.

1b The dinner was held at the restaurant of the Four Seasons Hotel.

2a The dinner tasted like it was microwaved leftovers from the day before.

2b The dinner got cold while everyone waited for the speeches to be over.

Door Senses: (1) *opening*, (2) *physical*

1a The door was just large enough for Tom to squeeze through.

1b The door let some light into the otherwise pitch-black room.

2a The door felt so heavy that it could withstand an atomic blast.

2b The door was painted with multiple layers of battleship grey.

Window Senses: (1) *opening*, (2) *physical*

1a The window offered a great view of the nearby town centre.

1b The window was the only source of fresh air for the stuffy office.

2a The window shattered when a gas tank exploded at a factory nearby.

2b The window was made out of four equally large rectangular panes.

Settlement Senses: (1) *process*, (2) *result*

1a The settlement took place after the region was declared neutral territory

1b The settlement was widely considered to be an act of political protest.

2a The settlement was home to more than 20 families from all over the country.

2b The settlement included a convenience store offering regional products.

Building Senses: (1) *process*, (2) *result*

1a The building took ten years longer than was originally planned.

1b The building was carried out by three construction companies.

2a The building was located at one of the busiest squares in the city.

2b The building housed a number of internationally acclaimed law firms.

Bank Senses: (1) *institution*, (2) *building*, (3) *branch* (4) *landscape*

1a The bank had to pay millions in fines after a money laundering scandal

1b The bank started an advertisement campaign to appeal to younger clients.

2a The bank was damaged during an attempted robbery involving explosives.

2b The bank had a beautiful facade with 19th-century Victorian design elements.

3a The bank gave a loan to Franklin so he could start a small business.

3b The bank hired a handyman to take care of the leak in the bathroom.

4a The bank wasn't marked on any of the commonly used nautical maps.

4b The bank had formed by sand accumulating due to a changed current.

University Senses: (1) *building*, (2) *directorate*, (3) *team* (4) *institute*

1a The university had been badly damaged during a flood two years ago.

1b The university was in dire need of some professional restoration.

2a The university banned radio-controlled drones after an accident.

2b The university rescheduled the date for the diploma ceremony.

3a The university had the best chances of winning the college trophy.

3b The university won over two thirds of last season's basketball matches.

4a The university was well respected for its research in theoretical physics.

4b The university obtained a 'Class A' distinction in the annual ranking.

Record Senses: (1) *physical*, (2) *content*, (3) *album* (4) *paperwork* (5) *achievement*

1a The record appeared to be badly warped due to improper storage.

1b The record had a white label with a handwritten note from the artist.

2a The record sounded amazing playing on the new stereo system.

2b The record quickly made everybody get together on the dance floor.

3a The record sold 12 million copies in the United States alone.

3b The record was shortlisted for a prestigious Mercury Award in 2018.

4a The record showed that the antique wardrobe was sold last week.

4b The record indicated that the buyer had paid the sum in full.

5a The record was set during the 1996 Olympic Games in Atlanta."

5b The record remained unbroken until the regulations changed in 2017.

Glass Senses: (1) *container*, (2) *content*, (3) *container*$\ominus$, (4) *content*$\ominus$

1a The glass left a ring of condensed water on the cardboard coaster.

1b The glass was filled much lower than the fill line on the label.

2a The glass tasted like a apple juice blended with forest fruits.

2b The glass thoroughly refreshed Ben after his 10k morning run.

3 The glass chipped when they accidentally hit it with a billiard cue.

4 The glass seemed to be some kind of high-caffeine energy drink.

Bottle Senses: (1) *container*, (2) *content*, (3) *container*$\ominus$, (4) *content*$\ominus$

1a The bottle was a lot larger than the other ones on the shelf.

1b The bottle had not been opened since it was made in 1986.

2a The bottle tasted exactly like Sue had always imagined.

2b The bottle made them talk a lot louder than they normally did.

3 The bottle was made out of recycled glass fished from the ocean.

4 The bottle was a fruit spirit produced by a family-run distillery.

Cup Senses: (1) *container*, (2) *content*,(3) *trophy*, (4) *container*$\ominus$, (5) *content*$\ominus$

1a The cup was decorated with drawings of exotic animals.

1b The cup normally was kept on the bottom shelf of the cupboard.

2a The cup tasted much sweeter than Jon remembered.

2b The cup was sweetened by naturally occurring fructose.

3a The cup is passed from one champion to the next every year.

3b The cup has last been won by the football club Liverpool FC.

4 The cup had a beautiful handle shaped to look like a snake.

5 The cup was made out of 100 % fresh, sun-riped fruit.

Beer Senses: (1) *container*, (2) *content*, (3) *container*$\ominus$,

1a The beer left a ring of condensed water on the cardboard coaster.

1b The beer was filled much lower than the fill line on the label.

2a The beer tasted exactly like Sue had remembered it.

2b The beer thoroughly refreshed Ben after his 10k evening run.

3 The beer was made out of recycled glass fished from the ocean.

Milk Senses: (1) *container*, (2) *content*, (3) *container*$\ominus$, 1a The milk was just too large to fit into the fridge's door compartment.

1b The milk had a beautiful drawing of grazing cows on the front.

2a The milk had been treated to stay fresh for almost a week.

2b The milk was the last ingredient for his perfect bowl of cereal.

3 The milk got squished when they dropped it on the kitchen surface.

Juice Senses: (1) *container*, (2) *content*, (3) *container⊖*,

1a The juice was just too large to fit into the fridge's door compartment.

1b The juice had drawings of exotic fruits Sue had never seen before.

2a The juice was made out of 100% fresh, sun-riped fruit.

2b The juice was sweetened by naturally occurring fructose only.

3 The juice got squished when they dropped it from the shelf.

B.1.1 Control Items

Instead of synonym and filler items, the second annotation run included two control items in each questionnaire used for filtering out spurious annotations in the data analysis. Control items were split into two categories: control-same, with items that should invoke exactly the same sense interpretation of the target word (or a perfectly acceptable sentence in the co-predication setting) and therefore lead to high ratings, and control-random, with random sentence combinations (and randomly scrambled word order in the co-predication setting). The control items used in the word sense similarity judgement task were as follows:

Control-same

The **bat** flew in through the open window.,
The **bat** flew in through the door.

The **jumper** didn't really fit her that well.,
The **jumper** didn't really suit her style.

The **club** only admits distinguished women over 50.,
The **club** only admits accomplished women under 30.

The **mole** dug tunnels all throughout the garden.,
The **mole** dug tunnels under the flower bed.

The **pitcher** threw a number of perfect curve balls.,
The **pitcher** threw at least three curve balls.

The **rocket** left the atmosphere at 2AM tonight.,
The **rocket** left the atmosphere yesterday at 5PM.

The **tank** was filled with over 500 liters of water.,
The **tank** was filled with a few hundred liters of water.

The **watch** slipped off his hand while he was swimming.,
The **watch** slipped off his hand while he was running.

The **yard** was overgrown with weeds.,
The **yard** was overgrown with scrub.

The **plane** landed with more than two hours delay.,
The **plane** landed with over three hours delay.

Control-random

The **spring** provides the oasis fresh water.,
The **mine** could be defused by an expert.

The **mine** had to close after an accident.,
The **order** was shipped two weeks later than expected.

The **order** gladly welcomed the new members.,
The **jumper** broke a long-standing regional record.

The **jumper** broke a long-standing regional record.,
The **letter** was signed by a famous magician.

The **plane** was a major feeding ground for buffalo.,
The **letter** looked like it could be an old-fashioned q.

The **mouse** stopped working in the middle of the presentation.,
The **plane** landed with more than two hours delay.

The **model** showed the proposed layout of the building complex.,
The **mouse** had chewed a hole into the bread basket.

The **model** wore a new dress designed by Versace.,
The **seal** indicated that the letter had never been opened.

The **seal** had gotten itself caught in an old fishing net.,
The **bass** added a thick, driving rhythm to the song.

The **match** ended without a clear winner.,
The **bass** managed to get off the hook.

The control items used in the co-predication acceptability either were conjunctive sentences without a co-predication (control-same), or conjunctive sentences where the second half of the sentence was randomly permutated (control-random). Since sample pairs were combined through conjunction reduction removing the first two words in the second sentence, each second sentence starts with two deletion markers (X X) that will be removed by the sample generator. The full list of control samples used in the co-predication acceptability task is as follows:

Control-same

A group of boys were playing Frisbee in the park.,
X X a girl tried to balance on a slack line.

The football players were getting ready for the game.,
X X the cheerleaders practiced their performance.

A muscular guy was changing the weights on the bench press.,
X X the gym instructor prepared a group session.

The car mechanic was changing a punctured tire.,
X X a police officer diverted the traffic.

The head chef was preparing the main course.,
X X a waiter brought Susan's starter.

The carpenter was fitting a large wardrobe.,
X X the painter rinsed out his brushes.

The guitarist was changing the strings of her guitar.,
X X the bass player tuned his instrument.

A pole vaulter was covering his hands with chalk.,
X X a track athlete was tying her running shoes.

The doctor was measuring a patient's pulse.,
X X a nurse added a note to the patient sheet.

The pilot was planning a new route to avoid a storm cloud.,
X X a steward was handing out refreshments to the passengers.

The teacher was explaining a difficult equation.,
X X the students were paying close attention.

Control-random

The spring provides the oasis fresh water.,
X X mine could by an expert the defused be.

The mine had to close after an accident.,
X X two weeks shipped was than expected later order the.

The order gladly welcomed the new members.,
X X record the a long-standing jumper broke regional.

The jumper broke a long-standing regional record.,
X X by letter signed a magician the was famous.

The plane was a major feeding ground for buffalo.,
X X q looked be the an it could like old-fashioned letter.

The mouse stopped working in the middle of the presentation., X X more with two landed the than delay hours plane.

The model showed the proposed layout of the building complex.,
X X had a chewed mouse the bread into the basket hole.

The model wore a new dress designed by Versace.,
X X opened letter the indicated never the that been had seal.

The seal had gotten itself caught in an old fishing net.,
X X added driving to the thick, the rhythm song bass a.

The match ended without a clear winner.,
X X the off the managed bass hook get to.

The club was made out of a single piece of wood.,
X X sheet to a note a nurse patient added the.

B.2 Similarity Pattern Heat Maps

Chapter 5.3.4 describes the observation of similarity patterns within certain types of polysemic alternations. Similarity patterns are visualised in heat maps, most of which are displayed in that chapter. The remaining heat maps for targets exhibiting an *event/food* alternation, an *physical/aperture* alternations or a *physical/content* alternation are shown here.

B.3 Sense Clustering Dendograms

In Chapter 5.3.5 we present a preliminary investigation of how well BERT Large's contextualised embeddings can distinguish polysemic word senses when fed to a hierarchical clustering algorithm. The chapter displayed a selection of dendogram visualisations of the results obtained, the remaining graphs are displayed here in Figures B.4 through B.9.

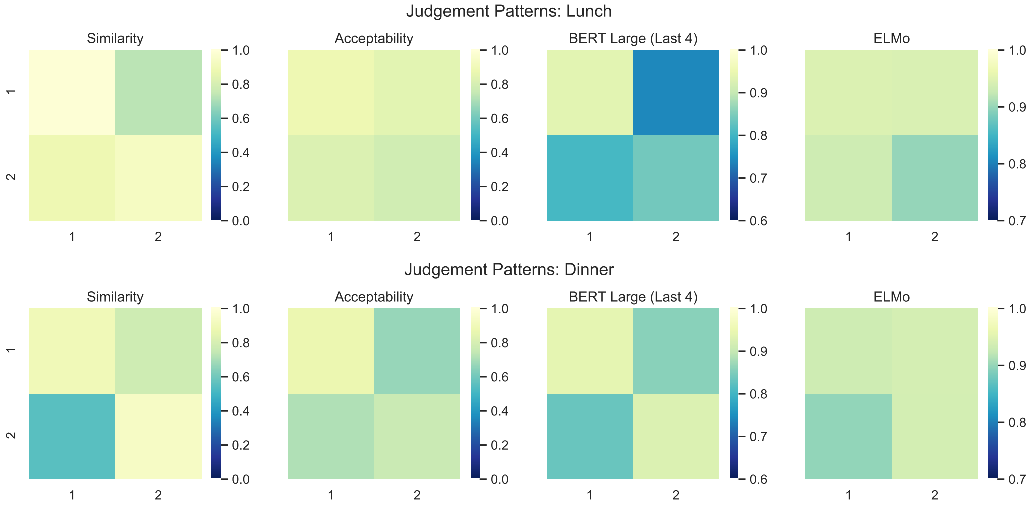


Figure B.1: Similarity patterns in the sense similarity ratings for *event/food* alternation polysemes. Senses: 1-, 2-. Colour-scales adjusted for computational measures.

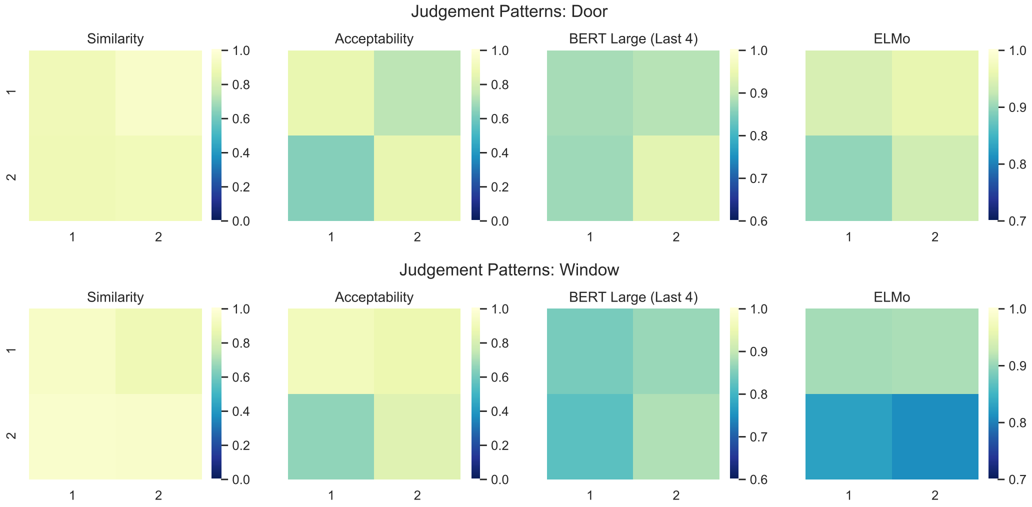


Figure B.2: Similarity patterns in the sense similarity ratings for *opening/physical* alternation polysemes. Senses: 1-, 2-. Colour-scales adjusted for computational measures.

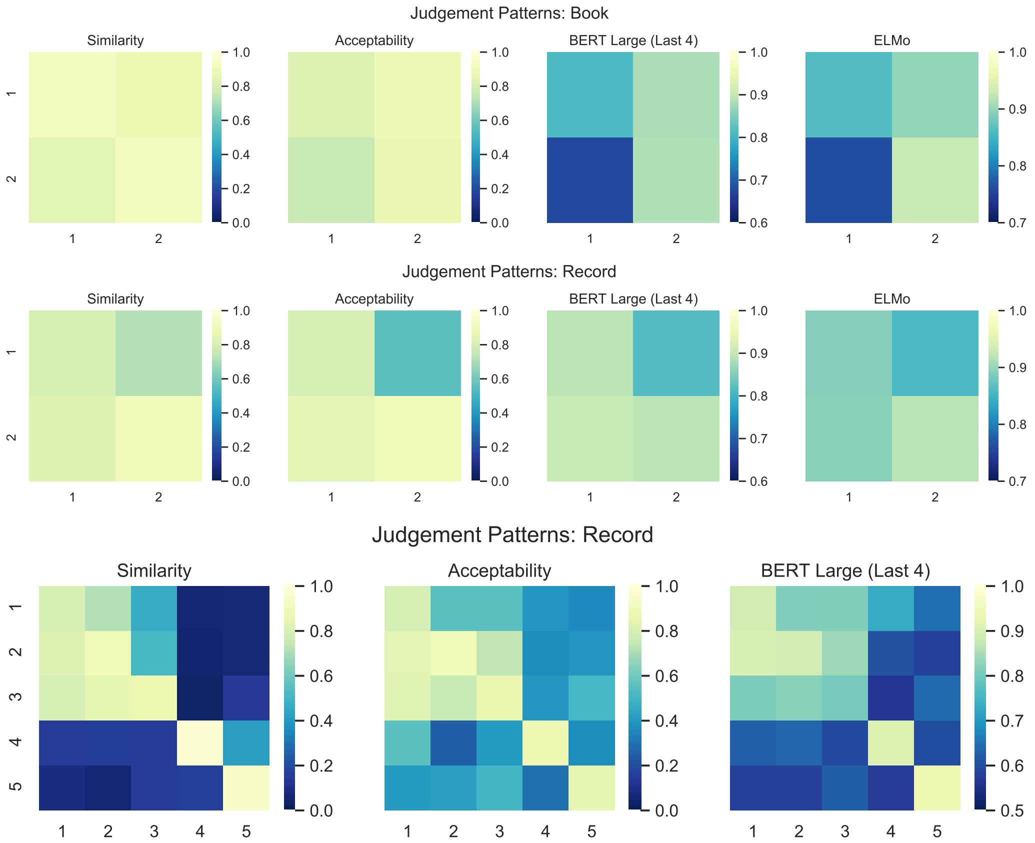


Figure B.3: Similarity patterns in the sense similarity ratings for *physical/information* alternation polysemes, with the last row showing all tested alternations of *record*, including homonymic readings. Polysemic senses: 1-, 2-. Colourscales adjusted for computational measures.

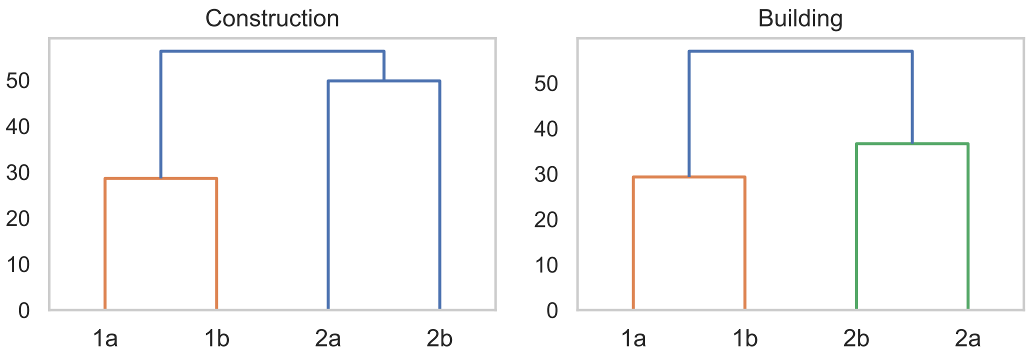


Figure B.4: Dendograms of BERT Large contextualised embedding similarity for targets *construction* and *building* based on hierarchical Ward clustering. Senses: 1-process, 2-result.

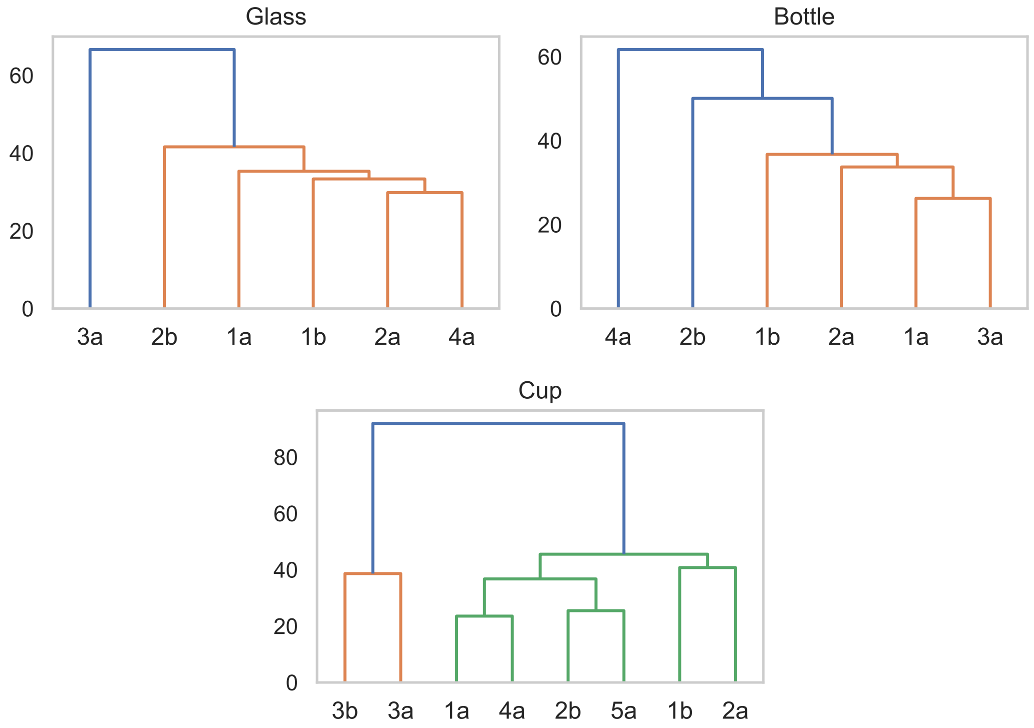


Figure B.5: Dendograms of BERT Large contextualised embedding similarity for targets *glass*, *bottle* and *cup* based on hierarchical Ward clustering. Senses for *glass* and *bottle*: 1-container, 2-content, 3-container⊖, 4-content⊖. Senses for *cup*: 1-container, 2-content, 3-trophy, 4-container⊖, 5-content⊖.

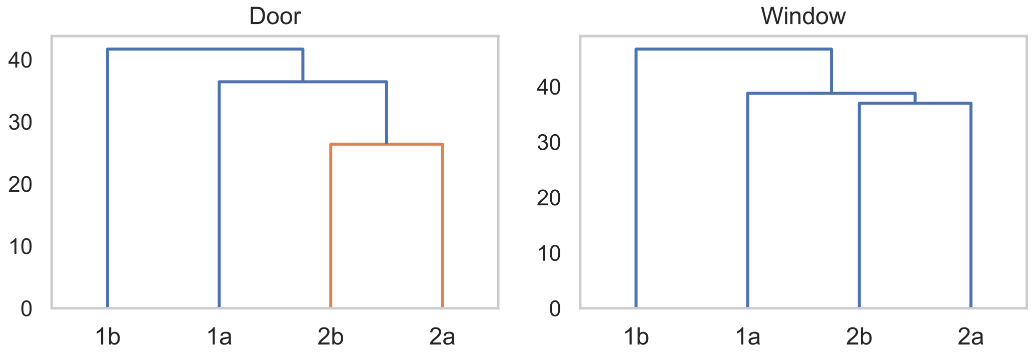


Figure B.6: Dendograms of BERT Large contextualised embedding similarity for targets *door* and *window* based on hierarchical Ward clustering. Senses: 1-opening, 2-physical.

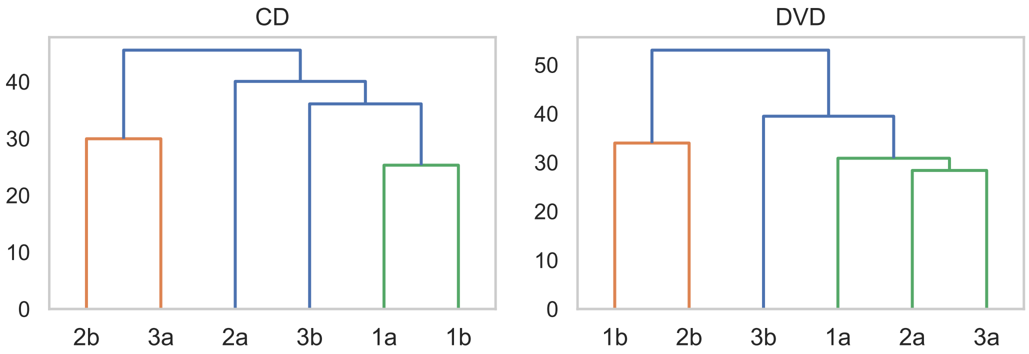


Figure B.7: Dendograms of BERT Large contextualised embedding similarity for targets *CD* and *DVD* based on hierarchical Ward clustering. Senses: 1-physical, 2-medium, 3-content.

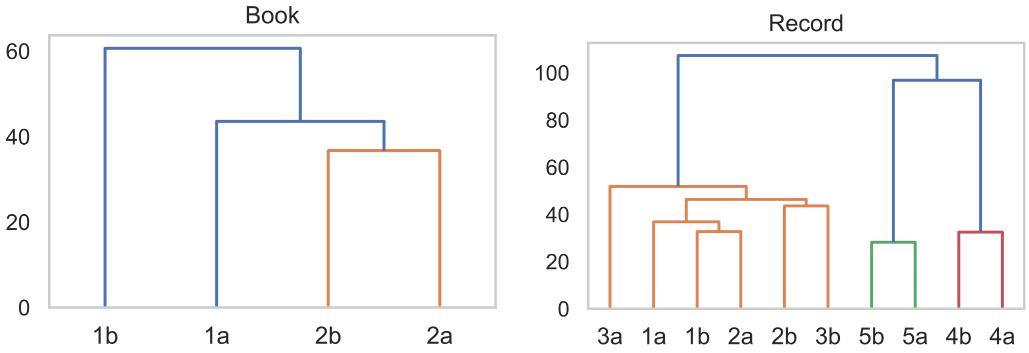


Figure B.8: Dendograms of BERT Large contextualised embedding similarity for targets *book* and *record* based on hierarchical Ward clustering. Senses: 1-physical, 2-content, 3-album, 4-paperwork, 5-achievement.

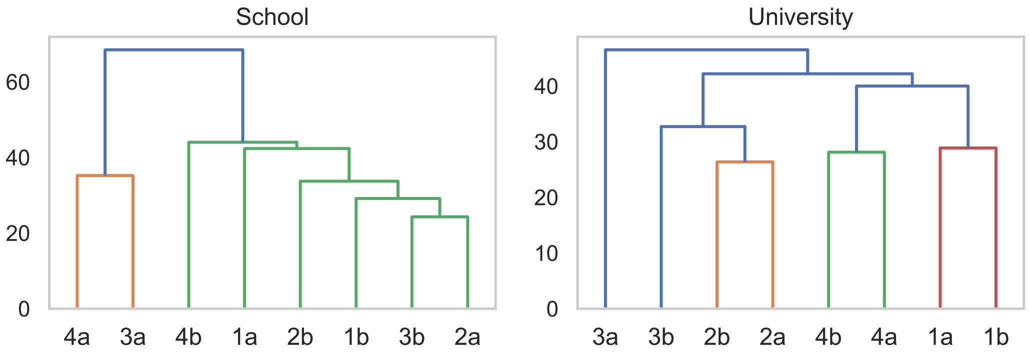


Figure B.9: Dendograms of BERT Large contextualised embedding similarity for targets *school* and *university* based on hierarchical Ward clustering. Senses: 1-building, 2-administration, 3-institution, 4-students.

Appendix C

Additional Materials for Chapter 6

In Chapter 6 we presented a pilot algorithm to automatically detect words that might allow for the same set of sense alternations as a given target word. The following tables contain the results of this pilot run, listing proposed substitutes for a given target and indicating those substitutes' polysemic potential as determined by the algorithm.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
fish	26.46	0.866	1	0.03135		X
meat	22.36	1.079	2	0.03569		
pork	21.44	1.104	2	0.03448		
rice	19.29	0.988	None	0.04358		X
beef	16.99	1.047	None	0.03992		
pig	11.62	0.873	1	0.03449		
food	10.79	1.037	None	0.02391		
duck	10.16	0.821	1	0.03649		X
bird	9.96	0.803	1	0.02377		

Table C.1: Evaluation of corpus-based substitutes for reference word *chicken*. Senses 1: *animal*, 2: *food*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
chicken	25.27	0.964	None	0.0407		
beef	23.03	0.982	None	0.0279		
pork	22.12	1.069	2	0.0254		
meat	21.54	1.021	None	0.0249		
goat	20.63	0.912	1	0.0353		X
fish	18.14	0.983	None	0.0260		X
sheep	16.24	0.903	1	0.0374		X
cow	15.41	0.875	1	0.0292		
pig	13.50	0.927	1	0.0366		X
dog	11.68	0.844	1	0.0286		

Table C.2: Evaluation of corpus-based substitutes for reference word *lamb*. Senses 1: *animal*, 2: *food*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
wine	31.49	0.890	1	0.0223		X
alcohol	19.48	0.935	1	0.0224		
liquor	18.99	0.935	1	0.0210		X
beers	18.95	0.928	1	0.0254		X
ale	17.58	0.947	1	0.0173		X
water	17.04	0.909	1	0.0193		
coffee	14.31	0.923	1	0.0311		X
drink	14.16	0.996	None	0.0389		X
food	13.48	0.865	1	0.0240		X

Table C.3: Evaluation of corpus-based substitutes for reference word *beer*. Senses 1: *container*, 2: *content*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
beer	22.90	0.961	None	0.0141		
wines	22.71	1.016	None	0.0181		
grape	17.24	1.012	None	0.0203		
water	16.75	0.934	1	0.0105		
food	16.36	0.925	1	0.0113		
liquor	12.89	0.942	1	0.0141		
champagne	9.81	0.953	None	0.0125		

Table C.4: Evaluation of corpus-based substitutes for reference word *wine*. Senses 1: *container*, 2: *content*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
glass	30.62	1.019	None	0.0193		
cup	16.94	1.039	None	0.0122		
container	13.38	1.012	None	0.0170		
box	12.74	0.995	None	0.0221	X	X
case	12.35	1.013	None	0.0117		
bag	12.21	1.052	2	0.0199		X
barrel	11.38	0.993	None	0.0261		
jar	11.33	1.068	2	0.0196		

Table C.5: Evaluation of corpus-based substitutes for reference word *bottle*. Senses 1: *container*, 2: *content*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
stone	23.97	0.993	None	0.0217		
metal	20.61	0.967	None	0.0260		
steel	18.65	0.940	1	0.0225		X
crystal	16.80	1.044	None	0.0196		
wood	16.60	0.959	None	0.0232		
water	12.65	1.101	2	0.0183		
wooden	12.55	0.936	1	0.0150		
window	10.89	1.004	None	0.0257		
windows	10.06	0.977	None	0.0241		
plastic	9.91	0.930	1	0.0204		

Table C.6: Evaluation of corpus-based substitutes for reference word *glass*. Senses 1: *container*, 2: *content*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
building	50.93	1.161	2	0.0270		
development	36.04	0.988	None	0.0353		
completion	18.55	0.925	1	0.0393		X
creation	17.29	1.090	2	0.0276		
reconstruction	16.89	1.034	None	0.0374		
design	16.50	1.098	2	0.0344		
work	16.36	1.046	None	0.0327		
renovation	12.26	1.006	None	0.0279		
expansion	12.16	1.043	None	0.0274		
erection	11.33	1.018	None	0.0338		

Table C.7: Evaluation of corpus-based substitutes for reference word *construction*. Senses 1: *process*, 2: *result*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
house	42.14	1.136	2	0.0221		
structure	38.87	1.099	2	0.0344		
church	25.54	1.068	2	0.0165		
tower	22.51	1.158	2	0.0183		
site	20.90	1.064	2	0.0184		
complex	19.63	1.103	2	0.0225		
hall	19.19	1.167	2	0.0156		
facility	18.02	1.128	2	0.0173		
buildings	14.36	1.131	2	0.0264		
school	14.16	1.060	2	0.0159		

Table C.8: Evaluation of corpus-based substitutes for reference word *building*. Senses 1: *process*, 2: *result*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
deer	20.64	1.178	2	0.0281		
duck	18.77	1.196	2	0.0234		
bird	17.43	1.179	2	0.0160		
pigeon	16.89	1.192	2	0.0198		
owl	15.55	1.192	2	0.0212		
eagle	15.01	1.190	2	0.0201		
hawk	12.87	1.169	2	0.0241		
frog	12.87	1.234	2	0.0146		

Table C.9: Evaluation of corpus-based substitutes for reference word *pheasant*. Senses 1: *animal*, 2: *food*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
dinner	43.60	1.124	2	0.0210		
breakfast	38.38	1.025	None	0.0232		
food	20.46	1.060	2	0.0156		
meal	20.36	1.065	2	0.0206		
tea	17.72	0.988	None	0.0161		
meals	16.41	1.072	2	0.0177		
coffee	14.65	0.990	None	0.0157		
school	10.84	0.913	1	0.0080		

Table C.10: Evaluation of corpus-based substitutes for reference word *lunch*. Senses 1: *event*, 2: *food*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
lunch	29.64	1.167	2	0.0374		X
breakfast	23.34	1.180	2	0.0400	X	X
party	20.31	0.819	1	0.0287		
tea	14.31	1.083	2	0.0238		X
supper	13.53	1.085	2	0.0519		
meeting	12.70	0.833	1	0.0356		
ceremony	12.06	0.776	1	0.0405		
event	11.38	0.884	1	0.0294		

Table C.11: Evaluation of corpus-based substitutes for reference word *dinner*. Senses 1: *event*, 2: *food*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
window	37.70	0.948	1	0.0365		
doors	36.67	0.968	None	0.0458		
gate	32.42	0.941	1	0.0336		X
doorway	25.73	1.057	2	0.0418		
entrance	23.97	0.976	None	0.0291		
wall	16.65	0.920	1	0.0327		
house	10.79	0.926	1	0.0248		

Table C.12: Evaluation of corpus-based substitutes for reference word *door*. Senses 1: *physical*, 2: *aperture*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
windows	35.06	0.926	1	0.0514		
door	34.77	0.869	1	0.0319		
doorway	18.51	0.952	None	0.0268		
wall	15.87	0.933	1	0.0251		
period	10.89	1.241	2	0.0220		
roof	10.69	0.872	1	0.0246		
glass	10.25	0.926	1	0.0300		

Table C.13: Evaluation of corpus-based substitutes for reference word *window*. Senses 1: *physical*, 2: *aperture*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
novel	53.96	1.139	2	0.0226		
work	41.75	1.001	None	0.0222		
film	29.44	1.055	2	0.0277		X
story	29.20	1.134	2	0.0280		
series	23.49	1.114	2	0.0253		
album	18.31	1.043	None	0.0152		
collection	17.72	0.970	None	0.0296		
books	17.19	0.971	None	0.0431		
volume	17.04	1.008	None	0.0290		
play	16.80	1.040	None	0.0203		

Table C.14: Evaluation of corpus-based substitutes for reference word *book*. Senses 1: *physical*, 2: *information*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
records	45.51	0.817	1	0.1027	X	X
mark	33.01	0.875	1	0.0188		
music	13.57	1.190	2	0.0328		
season	13.48	1.007	None	0.0171		
recording	12.89	1.117	2	0.0519		X
album	11.91	1.353	2	0.0261		
total	11.72	0.881	1	0.0151		
best	11.47	0.985	None	0.0126		

Table C.15: Evaluation of corpus-based substitutes for reference word *record*. Senses 1: *physical*, 2: *information*.

Target	**Subst.**	**Slope**	**Dom.**	**std.**	**M.M.**	**Polys.**
magazine	62.79	1.066	2	0.0296		X
paper	47.85	1.124	2	0.0288		X
publication	38.38	1.050	None	0.0289		
journal	36.28	1.092	2	0.0330		X
daily	28.56	1.118	2	0.0185		
weekly	26.32	1.140	2	0.0175		
newspapers	24.66	1.046	None	0.0368		
periodical	22.31	1.106	2	0.0201		
news	13.96	0.942	1	0.0268		
press	11.62	0.954	None	0.0316		

Table C.16: Evaluation of corpus-based substitutes for reference word *newspaper*. Senses 1: *physical*, 2: *organisation*.

Target	Subst.	Slope	Dom.	std.	M.M.	Polys.
magazine	62.79	0.986	None	0.0398		X
paper	47.85	0.949	1	0.0216		X
publication	38.38	0.967	None	0.0232		
journal	36.28	0.948	1	0.0227		X
daily	28.56	0.957	None	0.0269		
weekly	26.32	0.952	None	0.0217		
newspapers	24.66	0.990	None	0.0405		
periodical	22.31	0.961	None	0.0198		
news	13.96	0.940	1	0.0336		X
press	11.62	1.036	None	0.0427		

Table C.17: Evaluation of corpus-based substitutes for reference word *newspaper*. Senses 1: *physical*, 2: *information.*

Target	Subst.	Slope	Dom.	std.	M.M.	Polys.
magazine	62.79	0.930	1	0.0406		X
paper	47.85	0.863	1	0.0330		X
publication	38.38	0.935	1	0.0333		
journal	36.28	0.883	1	0.0390		X
daily	28.56	0.865	1	0.0240		
weekly	26.32	0.851	1	0.0224		
newspapers	24.66	0.950	1	0.0451		
periodical	22.31	0.884	1	0.0229		
news	13.96	1.043	None	0.0332		
press	11.62	1.052	2	0.0412		X

Table C.18: Evaluation of corpus-based substitutes for reference word *newspaper*. Senses 1: *organisation*, 2: *information.*