8-17-2020 3:00 PM

On Polysemy: A Philosophical, Psycholinguistic, and Computational Study

Jiangtian Li, *The University of Western Ontario*

Supervisor: Robert J. Stainton, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Philosophy

Abstract

Most words in natural languages are polysemous, that is they have related but different meanings in different contexts. These polysemous meanings (senses) are marked by their structuredness, flexibility, productivity, and regularity. Previous theories have focused on some of these features but not all of them together. Thus, I propose a new theory of polysemy, which has two components. First, word meaning is actively modulated by broad contexts in a continuous fashion. Second, clustering arises from contextual modulations of a word and is then entrenched in our long term memory to facilitate future production and processing. Hence, polysemous senses are entrenched clusters in contextual modulation of word meaning and a word is polysemous if and only if it has entrenched clustering in its contextual modulation. I argue that this theory explains all the features of polysemous senses.

In order to demonstrate more thoroughly how clusters emerge from meaning modulation during processing and provide evidence for this new theory, I implement the theory by training a recurrent neural network (RNN) that learns distributional information through exposure to a large corpus of English. Clusters of contextually modulated meanings emerge from how the model processes individual words in sentences. This trained model is validated against a human-annotated corpus of polysemy, focusing on the gradedness and flexibility of polysemous sense individuation, a human-annotated corpus of regular polysemy, focusing on the regularity of polysemy, and behavioral findings of offline sense relatedness ratings and online sentence processing.

Last, the implication to philosophy of this new theory of polysemy is discussed. I focus on the debate between semantic minimalism and semantic contextualism. I argue that the phenomenon of polysemy poses a severe challenge to semantic minimalism. No solution is foreseeable if the minimalist thesis is kept, and the existence of contextual modulation is denied within the literal truth condition of an utterance.

Keywords: Philosophy of Language, Semantics, Pragmatics, Context, Lexical Meaning, Polysemy, Computational Modeling

Summary for Lay Audience

Some words have more than one related meanings. For example, “lock” can mean a mechanism to fasten something or a confined section of waterway, and these two meanings are related to the concept of restriction. This phenomenon is called polysemy. Polysemy has some interesting features. For example, some meanings of polysemous words are more regular, such as the two meanings of “lock” mentioned above, but some meanings are less prototypical but still connect with the core regular meaning, such as the technique — “lock” of limbs in wrestling. Furthermore, new meanings can be created as time goes, such as a digital “lock” in the form of a computer program.

I propose a theory of polysemy to explain what polysemy is and why it has these features. I draw my theory from two characteristics of language. First, contexts of language use often change and contribute new information to the meaning of a word in use. Second, some contextual uses of language are similar and frequent so they are strengthened in our memory to ease future use. Polysemy, in my theory, is the collection of strengthened contextual meanings, or more formally, entrenched clusters of contextual meaning modulations.

In this dissertation, I compare my theory with other theories of polysemy, demonstrate how my theory of polysemy could explain the features of polysemy better. I also implement my theory in a computational model and verify it with annotated corpus and behavioral findings to collect evidence for my theory. Last, I discuss the implication of my new theory of polysemy to philosophy of language.

Co-Authorship Statement

Chapter 3 of this dissertation is a manuscript submitted for publication to the journal *Cognitive Science* that was co-authored by Dr. Marc Joanisse. I was the first (lead) author and wrote the full draft of this chapter. I also designed the study, programmed the computer simulation, and collected and analysed the data. Marc Joanisse guided my research, revised the draft extensively, and provided a lot of feedback on it.

Acknowlegements

I would like to express my sincere gratitude to my supervisor, Professor Robert Stainton, for his guidance along my journey pursuing philosophy and linguistics. He makes me believe that the details of natural languages matter for philosophy and the study of language should incorporate a broad evidential basis. This dissertation would have been very different (indeed, even with a very different methodology), had I not met him.

I would also like to express my deep gratitude to Professor Marc Joanisse, for his generosity of allowing me to attend his classes and join his lab. He helped me acquire quickly the necessary skills and knowledge to conduct psycholinguistic experiments and computational simulations and provided me with numerus feedbacks during my four years here.

My appreciation also extends to Professor Chris Viger for his mentoring of my teaching and spending time discussing philosophy with me. Thanks also go to Professor Ken McRae and Professor Arthur Sullivan for their comments on my early drafts and Professor Jixuan Zhang for his guidance during my very early career in pursuing philosophy. I would also like to thank Dr. Chang Liu, Dr. Filippos Papagiannopoulos, Dr. Justina Diaz-Legaspe, William Laufs, Alastair Crosby, Dylan Hurry, Fangnin Zhou, Miyu Bao, Xiaoqin Yao, with whom I enjoyed countless good discussions.

Finally, I am indebted to my family, my father Zheyuan Li, my mother Xue Jiang, my aunts Yue Jiang and Yun Jiang, my uncles Rujun Wang and Jianhua Shao, and my cousins (and in-law) Lingxiao Wang, Xiaoxuan Yu, Ru Shao, Hao Hu. They have been supporting my academic pursuit since the very beginning. This degree could not be attained without their support.

Contents

Abstract i

Summary for Lay Audience ii

Co-Authorship Statement iii

Acknowlegements iii

List of Figures viii

List of Tables x

1 Introduction to Polysemy and Theories of Polysemy 1

1.1 Introduction 1

1.2 Features of Polysemy 2

1.2.1 Relatedness and Structuredness 2

1.2.2 Gradedness and Flexibility 3

1.2.3 Productivity 4

1.2.4 Regularity and Literalness 5

1.2.5 Behavioral Characteristics of Polysemy Processing 7

1.3 Polysemy Compared to Other Context-Sensitivities of Word Meaning 8

1.3.1 Homonymy 9

1.3.2 Indexicality 10

1.3.3 Ad-hoc Meaning Variation 11

1.4 Previous Theories of Polysemy 12
1.4.1 Static Account of Polysemy 12
Sense-enumerative Account 13
Semantic Network Account 14
Distributed Representation Accounts 15
1.4.2 Operational Accounts of Polysemy 17
Formal Semantic Account 17
Relevance Theory 19
Recanati 20
1.5 Conclusion 21

2 A New Theory of Polysemy **22**
2.1 Introduction 22
2.2 Two Basic Components: Meaning Modulation and Clustering 23
2.2.1 Meaning Modulation by Context 24
Prototypical Meaning Modulation in Context 24
Ad Hoc Linguistic Coercion 25
Metonymy 25
Metaphor 26
Summary 27
2.2.2 Clustering of Meaning Modulation 27
Semantic Space Model 27
Clustering in the Semantic Space Model 29
Entrenchment from Clustering 31
2.2.3 Example of Polysemy Emerging from Two Components 33
2.3 Learning, Representation, and Processing of Polysemy 35
2.4 Handling Features of Polysemy with the New Theory 37

3 A Computational Model and Experiments **39**

3.1 Introduction 39

3.2 Model Architecture — Long Short-Term Memory 42

3.3 Model Training 45

3.4 Model Assessment with Linguistic Corpora 47

3.4.1 Modeling Senses as Clusters in Meaning Modulations 48

3.4.2 Modeling Regular Polysemy as Evidenced by Geometric Patterns between Clusters 53

3.5 Model Assessment with Behavioral Experiments 58

3.5.1 Modeling Sense Relatedness Ratings 59

3.5.2 Modeling Online Sentence Processing of Polysemy 61

3.6 Comparison with Other Models of Polysemy 64

3.7 Conclusion 68

4 Philosophical Implications of Polysemy **70**

4.1 Introduction 70

4.2 Background for the Debate between Minimalism and Contextualism 70

4.3 The Problem of Polysemy for Semantic Minimalism 73

4.4 Strategy 1: No Need to Handle Polysemy for Minimal Proposition 74

4.5 Strategy 2: Handle Polysemy as Ambiguity 77

4.5.1 Flexibility of Polysemy 78

4.5.2 Productivity of Polysemy 79

4.5.3 Psycholinguistic Evidence against Polysemy Disambiguation 82

4.6 Strategy 3: Assimilate Polysemy into the Basic Set of Context-Sensitivity . . . 83

4.7 What is Needed to Account for Polysemy? 85

Bibliography **88**

List of Figures

2.1 Visualization of a Three Dimensional Semantic Space 28
2.2 Visualization of Clustering of Meaning Modulation 29
2.3 Visualization of Clustering of a Monosemous Word 31
2.4 Visualization of Clustering Properties 32
2.5 Visualization of Meaning Modulation with 4 clusters 33

3.1 Model architecture. Texts are presented to the model word-by-word. Each word is presented individually on the input layer, and the model's task was to predict the following word in the sequence, so the input and output layer has the same size as the training vocabulary (Including 267,735 words. Each different surface form of the "same word" is treated as a different word. "Walked" and "walk" are treated as separate word tokens.). This was achieved by passing activation through two intermediary layers, one static and one recurrent. 44
3.2 Unrolled Architecture depicting how the network represents a temporal sequence of word inputs, and uses prior information to predict upcoming words (and punctuations). Since word prediction is imperfect because of the creative and productive nature of language, the model learns to output a probability distribution of possible next words in a sequence. The model learns to form internal representations on the Recurrent layer that encode contextual information about an input word and its preceding inputs in order to maximize accuracy in predicting subsequent words. 46

3.3 The first two pictures are plotted according to the annotation in the corpus. The second two pictures present the result of the unsupervised Gaussian mixture model with blue contours representing bivariate Gaussian distributions. Each Gaussian distribution represents a sense of the word. The darkness of blue represents the probability density of the Gaussian distribution. So the darker the blue is, the more probable the word of this sense will have this particular modulated meaning. 50

3.4 Regular polysemy modeled as geometric relations between clusters. Each dot represents the mean of a word sense's meaning cluster. Each line denotes a two-dimensional simplification of the multidimensional relationship between the two senses of each word. Similar-direction lines suggest a common type of relatedness among different word pairs, reflecting sense regularity. We quantified this regularity by computing the angle of each pair of word senses in high-dimensional space, and then computed the degree of variability in this angle, where highly regular patterns should be reflected in very low degrees of variability among sense pairs. 57

3.5 Scatterplot of the correlation between model produced cosine similarity and human rated relatedness of polysemous words 60

List of Tables

1.1 Unambiguity, polysemy, and homonymy representations in attractor networks. Representations are coded using binary units of 1 or 0. Polysemy is coded as an identical orthographic code (0010) mapping onto two different but overlapping semantic codes, differentiated via a set of context features that condition the recognition of one or another. Unambiguous words have the very same activation patterns while homonyms have totally different activation patterns. Adapted from Armstrong and Plaut (2016). 15

3.1 Clusters of Polysemous senses. Standard deviation of all means is shown in parentheses. 52

3.2 Similarity between context-free meaning and modulated meaning of the target word and similarity between context-free meaning and previous context of the target word. Measured by Euclidean Distance in the semantic space. 54

3.3 Result of permutation test of regular polysemy 58

3.4 Materials of sentence processing in Foraker and Murphy's study 62

3.5 Result of sentence processing simulation. Standard deviation of all means is shown in parentheses. 62

3.6 Result of repeated-measures ANOVA of sentence processing simulation 63

3.7 Result of *post hoc* analysis of sentence processing simulation 63

3.8 Discrete polysemy representations in attractor networks. Representations are coded using binary units of 1 or 0. Polysemy is coded as an identical orthographic code (0010) mapping onto two different but overlapping semantic codes, differentiated via a set of context features that condition the recognition of one or another. Adapted from Armstrong and Plaut (2016). 66

Chapter 1

Introduction to Polysemy and Theories of Polysemy

1.1 Introduction

In some very loose sense, word meanings are context-sensitive, as meanings of a word can be different in different contexts[1]. In this dissertation, I focus on a specific kind of context-sensitivity of meaning — polysemy. Polysemy is the phenomenon that a single word type[2] has multiple related literal meanings, which we traditionally call senses of a word. For example, "cup" can either mean a container in "I broke my cup" or its contents in "I finished my cup," and which sense it means depends on the context. Note that the content sense and container sense are semantically related.

Before the review of the phenomenon of polysemy, I want to briefly contrast it with other similar context-sensitivity of word meaning. First, there is the phenomenon that two words

[1]There is a very radical sense of context-sensitivity that word meanings change diachronically very quickly, imagining the meaning of "awful" changes from "inspiring wonder (or fear)" to "very bad" in a second. There is another radical sense that word meaning does not change in context at all as there is a minimal shared core among all uses of the same word, as long as there is an intersection. I do not mean these radical senses of context sensitivity. Instead, I mean the context sensitivity in communication where a word can refer to different things in different contexts of use.

[2]Word type is a word understood as a type rather than a token. For example, there are three "cup" tokens but only one "cup" type in "cup, cup, cup."

have the same pronunciation, such as "loch" in "loch lake" and "lock" in "door lock," so a single sound can have different meanings, which is called **homophony**. Second, there is the phenomenon that two words share the same pronunciation and spelling such as "lock" as in "a lock of hair" and "lock" in "a door lock," so that a single spelling and sound can have different meanings, which is called **homonymy**. Third, a single word can have different literal meanings (senses) in different contexts. For example, the word "lock" can either mean a mechanism for keeping an item fastened in the context of "a door lock" or a confined section of canal or water in "a lock filled with water." This is the topic of this dissertation — **polysemy**. Lastly, there are more ad-hoc context-sensitivity of words such as **metaphor** and **metonymy**. For example, the word "lock" means differently from its literal meaning in "his heart is a lock without any key."

In this chapter, I discuss different features of polysemy and review the previous theories of polysemy. I also argue that most theories fail to capture all these features, which motivates my theory of polysemy that polysemy is clustering and entrenchment of meaning modulation in context, as discussed in the second chapter.

1.2 Features of Polysemy

In this section, I first focus on different features of polysemy that makes it interesting and distinctive and then discuss how it differs from other context-sensitivity of word meaning such as homonymy or metaphor. I propose that these features should be captured by a successful theory of polysemy.

1.2.1 Relatedness and Structuredness

One key feature of polysemy is that different senses of a polysemous word are semantically related. "Lock" as the confined compartment of a canal and the lock as a mechanism of fastening share the base root meaning of barrier, enclosure, confinement, which is separated from

the reading[3] of "lock" as a piece of hair coils or hangs together. As a result, linguists usually categorize "canal lock" and "door lock" as polysemy but "door lock" and "a lock of hair" as homonymy depending on the relatedness of different senses/readings of a word.

The relatedness between polysemous senses is intricate. Different senses are related in a structured way. Some polysemous words have a central sense, and other senses of it are derived from it. For example, "lock" has a central sense of confinement and enclosure. Senses of different kinds of confinement, such as canal lock, are derived from this central one. New senses can also derive from already derived senses. For example, "over" has a central sense of spatial relation of one is higher than another, such as "a tent over the bed." A spatial trajectory sense of going above and across is derived from this spatial relation sense, such as "we climbed over this mountain." Furthermore, a temporal trajectory/duration sense can be further derived from the spatial trajectory sense of "over," such as "we talked over coffee." As a result, different senses of a word can form an intricate radial network where peripheral senses are radiated from the central sense (Brugman, 1988).

1.2.2 Gradedness and Flexibility

On a coarse scale, we can represent polysemous senses as discrete pieces that form a network. However, on a finer scale, the identities of senses are more graded and difficult to determine. For example

(1) Trout is a baseball star.

(2) Leonardo DiCaprio is a movie star.

(3) Venus is the evening star.

(4) You cannot see any star in the daytime

(5) Venus is not a star.

[3]In this dissertation, "reading" refers to different meanings of homonyms while sense refers to different meanings of polysemous words.

It is easier to distinguish (1) and (2) from (3), (4), and (5), but questions arise when distinguishing more similar uses with each other. Does "baseball star" in (1) belong to the same sense as "movie star" in (2)? Is the more general sense of star (a luminous point in the night sky, excluding the sun) in (3) and (4) the same sense of fixed star (a large and fixed incandescent astronomical object) in (5)? Clearly, we can say loosely that Venus is a star, but we can also say strictly that Venus is not a star because it is a planet. So both (3) and (5) can be true depending on what is meant by "star." The borderlines between each different sense of a polysemous word are not always clear. The difference between senses is more likely to be a gradient, and these cases exhibit the gradedness of sense individuation. Ambridge (2019) introduces a similar notion called the lumping-or-splitting problem. When individuating senses, theorists who treat cases like this in the usual way have to either lump different uses into one sense or split them into different senses. Lumping is difficult because it may ignore the meaning difference within the lumped sense category. Splitting is also difficult because there is no principled way to stop further splitting.

This lumping or splitting problem marks the gradedness of polysemy: that different senses of polysemous words are not clear cut and set in stone. Instead, the individuation of senses is often graded. Some uses are more easy to categorize while some are not. On the flip side, the gradedness of polysemy makes polysemy flexible. Speakers are not limited to a finite number of discrete options when they use a polysemous word. Instead, polysemous words can be used in numerous less typical ways that are akin to the more prototypical senses.

1.2.3 Productivity

Polysemous words are not only flexible, but also extendable to new senses, and we understand perfectly these newly created senses without explicit instruction. So it is different from slang terms that mostly need to be taught. We call this property of polysemy *productivity*. For example, it is known that the names of authors can mean either the person or his or her work.

(6) Shakespeare was well educated.

(7) People don't read much Shakespeare hereabouts.

(8) Every library has some Shakespeares.

"Shakespeare" in (6) refers to the person Shakespeare, while "Shakespeare" and "Shakespeares" in (7) and (8) refer to the work or copies of work of Shakespeare. These two senses are very common and entrenched in English, but there are also cases in which "Shakespeare" acquires new senses relatively recently.

(9) The Shakespeare in my Kindle is locked.

Here, "Shakespeare" in (9) refers to a digital file that can be locked digitally — encryption. It is clear that this sense of a digital file did not exist in the past and was not explicitly instructed to English speakers. However, it can be readily understood by competent English speakers.

(10) Tom is a fast runner.

(11) Jack is a fast haircutter.

(12) Jim is a fast recoverer from COVID-19.

Similarly, the polysemy of the adjective "fast" is productive in that it can also be extended to new senses as fast in terms of making a hair cut in (11) and fast in terms of recovering from symptoms of COVID-19 in (12).

In sum, the productivity of polysemy is the feature that polysemous words acquire new senses without explicit instruction.

1.2.4 Regularity and Literalness

The regularity of polysemy means two things. First, polysemy is not an ad-hoc and one-off meaning variation, such as live metaphors. Polysemous senses are clearly used many times by many people in many places, so polysemy is different from the ad-hoc meaning variation

in context in terms of its regularity of uses. Second, regularity of polysemy refers to a particular category of polysemy – regular polysemy. It is not the case that only "Shakespeare" and "Jiangtian" can mean either the person or the person's written word. The pattern of author sense and work sense applies exists in a lot of other names, such as "Dumas." Therefore, not only a particular sense but a pattern of different senses can be regular. The latter is called regular polysemy. Regular polysemy is a subcategory of polysemous words sharing patterns with other polysemous words so that their senses follow a similar sense pattern, such as the AUTHOR-WORK pattern. A formal definition of regular polysemy was first given by Apresjan (1974).

> "Polysemy of the word A with the meanings a_i and a_j is called regular if, in the given language, there exists at least one other word B with the meanings b_i and b_j, which are semantically distinguished from each other in exactly the same way as a_i and a_j and if a_i and b_i, a_j and b_j are nonsynonymous."

Besides different names meaning both the person and his or her work, regular polysemy is also exemplified by words like "bottle" and "can." Both of them have the container and the content sense as in (13)/(14) and (15)/(16). These two senses have similar uses in "bottle" and "can."

(13) This bottle tastes better.

(14) This bottle is fragile.

(15) I ate a whole can.

(16) I threw out an empty can.

There are other categories of regular polysemy, such as INFORMATION-ARTICLE in "book" and "paper," and MEAT-ANIMAL in "chicken" and "fish." Within each category, there are many words retaining the same sense pattern. These regularities indicate that different

senses of polysemous words are organized in a structured way, and certain structures can be shared by other words so that they compose a regular category.

As a result, a theory of polysemy should capture the structure governing the different senses of a polysemous word and particularly the regular structure as exemplified in "bottle" and "can."

1.2.5 Behavioral Characteristics of Polysemy Processing

Polysemy has also been extensively studied in psycholinguistics with behavioral experiments, and it has been extensively contrasted with homonyms. People have found the polysemy advantage that polysemous words are activated faster than homonyms in lexical decision tasks (Klepousniotou, 2002; Rodd et al., 2002; Brown, 2008), because different senses of polysemous words are more related to each other.

In terms of polysemy itself, the most certain finding is the context effect. Polysemous senses are processed faster in a context that is consistent with the intended sense. Klein and Murphy (2001, 2002) first studied how the context of a polysemous word influences its processing in real-time. They found that the reaction time of the sensicality judgment of a phrase with a polysemous word is faster when primed with a consistent sense than a non-consistent sense. For example, "wrapping paper" is judged faster when primed by "liberal paper" than "shredded paper." This effect has also be discovered in more realistic settings (Foraker and Murphy, 2012) where participants read sentences containing both senses of a polysemous word, and they read the target sense faster in the consistent context. For example, "fabric" is read faster in the sentence-consistent context (17) than "crop" in the sentence with inconsistent context (18).

(17) the fashion designers discussed the cotton after the fabric ripped a second time.

(18) The fashion designers discussed the cotton after the crop failed a second time.

The next characteristic is the frequency/dominance effect of polysemous senses. Different

senses of a polysemous word are used unequally. For example, the thin plastic sense of "film" is used less frequently/dominantly than the movie sense of "film" in contemporary English. Polysemous senses that are more frequent are processed faster than less frequent senses when other conditions are controlled (Foraker and Murphy, 2012). The characteristic can be seen as the behavioral aspect of the regularity of polysemy discussed in the previous chapters. Polysemous senses are not created purely ad-hoc, but are regularly used so that the frequency of how regular a polysemous sense modulates how it is processed.

The third characteristic of polysemy is the semantic similarity effect. The semantic similarity between different senses will modulate the processing of this word in the context of an inconsistent sense. Klepousniotou et al. (2008) and Brown (2008) found that participants are faster to make sensicality judgment when the sense of the target is more semantically overlapped with the sense of prime. For example, it is faster to judge "best-selling book" primed with "heavy book" than to judge "movie admission" primed with "guilty admission." This behavioral effect echoes the feature of the relatedness of senses. It shows that polysemous senses are not simply encoded flatly in the lexicon, but there is a structure of how these different senses are related to each other.

1.3 Polysemy Compared to Other Context-Sensitivities of Word Meaning

Polysemy is not the only kind of phenomenon in the category of contextual variation of word meaning. There are other phenomena belonging to this category. In this section, I discuss homonymy, indexicality, ad-hoc context-sensitivity such as metaphor and metonymy and contrast them with polysemy.[4]

[4]Pure homophony is not included here because the spelling difference between two tokens, "too" and "to," can easily differentiate it from polysemy, where spellings of different tokens are the same.

1.3.1 Homonymy

First, polysemy is differentiated from homonymy, which is two words of the same spelling and sound but having very unrelated readings. The relatedness of reading is very important for distinguishing homonymy from polysemy. Without the help of it, we will not be able to decide whether two word tokens are of the same word type or two different word types with the same spelling.

(19) Bats are dangerous animal.

(20) Professional baseball players swing their bats everyday.

The animal reading of "bats" in (19) and the baseball stick sense of "bats" in (20) are semantically very distinct. Usually, these kinds of homonyms arise from historical accidents that two distinct words share the same word form.

However, the distinction between homonymy and polysemy is not very clear cut, because we do not always have a clear cut way to distinguish them. For example, etymology does not always serve as a perfect test for homonymy. For example, the two readings of the homonym "bank" as financial institution and riverside actually share the same Germanic origin, which means the long shaped bench or table (Renaissance Florentine bankers make transactions on them).

Besides the etymology, people have designed other tests to distinguish polysemy and homonymy. For example, there is the co-predication test where a homonym, such as "bat," predicated with properties of its two different readings, "nocturnal" and "made of wood," makes the sentence (21) a bizarre zeugma. Instead, a polysemous word predicated with its two different senses does not make the sentence semantically anomalous, as in (22).

(21) Bats are nocturnal and usually made of wood.

(22) Novels are enjoyable and usually very long.

(23) Huge amount of power is desired by any supercomputer and president.

Similarly, the co-predication test is not perfect. Typical polysemous words, such as "power" in (23), fail the test. There are other problems of testing polysemy and homonymy, which are well covered in Sennet (2016)'s summary.

As a result, the distinction between homonymy and polysemy is not clear cut. The relatedness of senses makes the difference between homonymy and polysemy to be better understood as a gradient. However, it does not mean that homonymy and polysemy are the same. It means that homonymy and polysemy differ in a continuous spectrum of semantic relatedness. The difference in relatedness of senses does make homonyms behave differently from polysemy, as summarized in the previous sections. For example, homonyms cannot be used flexibly and productively. On the contrary, there are no intermediary senses between the animal bat and baseball bat so that the "borderline" between them is easier to draw than that of homonyms.

1.3.2 Indexicality

Indexical terms include pronouns and demonstrative terms such as "I," "he," "now," or "this." In one sense, the meaning of these terms is constant across different contexts. "I" always means the speaker of the current utterance. "He" always means the previously mentioned or an easily identified man or boy. "Now" always means the time at present. However, in another sense, the meanings of these terms are very context-sensitive because the extensions of these terms, which are things referred to by these terms, vary in different contexts. Just a simple example,

(24) Tom says, "I am in love with you."

(25) Jack says, "I want to stay away from the danger."

The "I" in both (24) and (25) means the speaker of the utterance, but they refer to different persons. The extension of "I" in (24) is Tom, while it is Jack for "I" in (25). Kaplan (Almog et al., 1989) introduced the notion of character to capture this kind of context-sensitivity. The character of a word is a function from the context of the word to its extension, and this function can be captured by a uniform rule or criterion such as "the most obvious man in the context."

The character for indexical terms itself does not change in different contexts, but the output (extension) changes because the input into the character of indexical terms changes. This explains the intuition that "I" has a constant meaning (character) but means different persons in different contexts (extension).

Indexicality is different from polysemy because there is no uniform rule or criterion, such as *the current speaker of the utterance* of "I," that determines the extension of polysemous words in context. The extensions of "lock" in "door lock" and "canal lock" are determined differently with different criteria so that the context-sensitivity of polysemous words results from the variation of characters rather than only their extensions. As a result, indexicality is the phenomenon of extension variation without character variation, but polysemy is the phenomenon of character variation along with extension variation in context.

1.3.3 Ad-hoc Meaning Variation

Last, we have ad-hoc meaning variation, such as live metaphors.

(26) A work is a death mask of its conception.

In the literature of metaphor, "death mask" in (26) is called a source, and "work" is called a target. It is not very often one sees work described as a death mask, and "death mask" is also rarely used to mean an end. If it is the first time someone encounters (26), it may take a second to interpret what is meant by this sentence.

Metonymy is another kind of ad-hoc meaning variation wherein the extension of a word is mapped to another entity in very different categories. In (27), the extension of "ham sandwich" is mapped from the sandwich to the person who ordered a sandwich.

(27) The ham sandwich is getting impatient

In these cases, the frequency of use (regularity) marks the dimension that polysemy is different from ad-hoc meaning variation. Ad-hoc meaning variations are one-off cases involving

more creativity, which constitutes wholly novel use of an existing word. In contrast, polysemy is mostly conventionalized and hence used more frequent than live metaphors and metonymy.

In sum, homonymy, indexicality, and ad-hoc meaning variation are all different from polysemy. Homonymy is different from polysemy in terms of the relatedness of different senses of a word in different contexts. Indexicality results from uniform rule-like determination of referents of a word in context, which is absent in polysemy. Finally, ad-hoc meaning variation is less frequent and more novel than polysemy. Even though the differences between polysemy and other context-sensitivities may not always be clear-cut, it does not negate that most of the cases are different, and these differences should be captured by theories of polysemy.

1.4 Previous Theories of Polysemy

During the last 30 years, there have been some systematic studies on polysemy in theoretical linguistics, psycholinguistics, and philosophy of language. Falkum and Vicente (2015) provide a systematic review of these studies. Instead of reviewing different theories of polysemy based on different fields of studies, I focus on the general theoretical assumptions behind them. Therefore, I categorize these theories of polysemy into two groups, the *static* account and the *operational* account, based on whether they encode the polysemous senses statically into the semantic representation[5] posited in their theories. I also discuss how well each account explains the features proposed in the previous sections.

1.4.1 Static Account of Polysemy

The *static* account captures polysemy in terms of the type meaning of a polysemous word — the context-free semantic representation of a polysemous word statically encoded in the lexicon. In this representation, there are structures explicitly encoding different senses of a polysemous

[5]The term representation used in this dissertation is twofold. First, it refers to the mental representation as our knowledge of language. Second, it also refers to the representation in a linguistic theory. These two intended meanings are different if one does not equate a theory of linguistics with a theory of our mental knowledge of language.

word. The structure can be a list of possible senses, a network where senses correspond to different nodes in the network, or a set of different activation patterns of neuron-like units. The common theme among these accounts is that senses of a polysemous word are directly and statically encoded in the type meaning of the word. I will now consider several sub-varieties.

Sense-enumerative Account

The sense-enumerative account of polysemy is the most straightforward theory of polysemy. Different senses of a word are just encoded as a list of discrete entries in the lexicon. Consider the polysemous word "cup" as an example, the semantic representation of "cup" can be represented as

(28)
$$\begin{bmatrix} \text{Cup} \\ \textit{Semantic} \begin{bmatrix} 1 : \text{CUP-CONTAINER} \\ 2 : \text{CUP-CONTENT} \end{bmatrix} \end{bmatrix}$$

Here, the semantic representation of "cup" is a list with two discrete entries of senses, CUP-CONTAINER as in "I broke a cup" and CUP-CONTENT as in "I finished my cup." Other senses can be added in a similar fashion. The processing of "cup" in context is a process of disambiguation to one among the listed entries. In other words, polysemy is literally the one-to-many mapping between form and meaning, and there is no sophisticated relationship between different senses.

There are a couple of problems with this approach. First, listing senses like dictionary entries could not capture the flexibility of polysemy. There is no way to capture polysemy that does not align strictly with the encoded senses. Second, listing senses does not capture the structuredness of polysemy. A flat list does not represent the relationship between different senses of a word. Third, probably the worst problem is that a list of senses is of necessity very small so that it does not capture the productivity of polysemy. A list does not tell us the unbounded possibility of creative new uses.

Semantic Network Account

Another way to encode different senses of polysemous words is to represent them as a semantic network in which different senses are nodes and connected by edges of different relationships. The iconic work is Brugman (1988)'s analysis of the polysemous word "over."

(29) The painting is *over* the mantel.

(30) The plane is flying *over* the hill.

(31) Sam is walking *over* the hill.

(32) Sam lives *over* the hill.

(33) The wall fell *over*.

(34) Sam turned the page *over*.

(35) Sam turned *over*.

(36) She spread the tablecloth *over* the table.

(37) Guards were posted all *over* the hill.

(38) The play is *over*.

(39) Do it *over*, but don't *over* do it.

In Brugman (1988)'s work, different senses of "over" compose a network structure in which there is a prototypical sense "above cross sense" as in (30) and other senses are derived from this sense. The benefit of this network representation over the sense-enumerative account is that it captures the relationship among senses and how each sense derives from other senses.

The network model is clearly an improvement from the sense-enumerative account as it captures the sophisticated structure of senses. Different polysemous words could have different structures of sense organization encoded as the network topology. The regular pattern of the

polysemous word "bottle" can be captured by the topology of the network so that "bottle" shares with words such as "glass" similar network topologies. However, this account still does not deal with the flexibility of polysemous words. Senses have to be individuated as nodes in the network, so it faces the same final problem as the sense-enumerative account.

Distributed Representation Accounts

The third option is widely used in the literature of connectionist modeling. In connectionist modeling of polysemy (Rodd et al., 2004; Armstrong and Plaut, 2008, 2016; Rodd, 2020), each different sense of a polysemous word is encoded as the activation pattern of multiple neuron-like units. Thus, a sense is not represented as an entry in a list or a node in a network but distributed among multiple units. Each unit can represent a semantic feature so that a distributed representation can be understood to be a weighted list of semantic features. However, this is not mandatory, as most models utilizing distributed representation do not assume that a unit has to represent only one feature. Some models also use continuous activation values for units so that the semantic feature is not binary.

Item type	Orthography	Context	Semantics
Unambiguous	0100	10	00000000000000001111
	0100	01	00000000000000001111
Polysemous	0010	10	00000001111000000000
	0010	01	00000000111100000000
Homonymous	0001	10	00000001111000000000
	0001	01	11110000000000000000

Table 1.1: Unambiguity, polysemy, and homonymy representations in attractor networks. Representations are coded using binary units of 1 or 0. Polysemy is coded as an identical orthographic code (0010) mapping onto two different but overlapping semantic codes, differentiated via a set of context features that condition the recognition of one or another. Unambiguous words have the very same activation patterns while homonyms have totally different activation patterns. Adapted from Armstrong and Plaut (2016).

Rodd et al. (2004)'s and Armstrong and Plaut (2008, 2016)'s models are attractor networks. Attractors networks are highly interconnected networks, whose activity settles into a stable state after a period of time. Basically, their models have a layer of orthography repre-

senting the word form of polysemous words and another layer of semantics representing the senses of polysemy. These two layers are connected so that a word form can activate its word meaning. They train their models until the semantic layer learns to settle into the correct senses in different contexts. So different senses are represented as different activation patterns in the same layer of semantics.

One advantage of this account is that it captures the graded difference between homonyms and polysemy. The relatedness between different senses can be very intuitively captured as whether there is overlapping in the activation of units. It can also be quantified by formal distance metrics such as euclidean distance between the activation patterns. Similar to the semantic network account, this account could potentially capture sophisticated structures of different senses with the help of probing tasks such as hierarchical clustering. However, senses are still individuated as discrete rather than continuous patterns of activation. There are no intermediary senses between the two polysemous senses listed in table 1.1. The same worry of individuating similar but different senses still looms.

In sum, three different *static* accounts all have the problem of capturing the flexibility of polysemy, because they encode the senses directly into the semantics in the lexicon, so they all face the lumping-or-splitting problem as previously mentioned. No matter how sophisticated a theorist individuates the senses, there are always cases that do not fit into the represented sense repository. Furthermore, *static* accounts cannot capture the productivity of polysemy because all possible senses are already encoded in the semantic representation. There is a list of senses to choose from, instead of the on-the-fly generation of senses in a novel context. The specific one in the context has to be one of them. Without extra mechanisms of encoding new senses or some kinds of post hoc rule, *static* accounts lack the explanation of the productivity of polysemy.

1.4.2 Operational Accounts of Polysemy

The second type of account, the *operational* account, does not encode the polysemous sense directly in the type meaning of the word. Instead, each specific sense is produced through some kinds of operation in context. This makes the *operational* account of polysemy both a consititutive account that tell what polysemy is and a causal account of polysemy explains how polysemy arises and is processed because the caual account is part of what constitutes polysemy[6]. A word is polysemous if and only if its processing involves certain operations. These operations include formal semantic coercion and type-shifting, relevance-based pragmatic inference, and so on, so an *operational* account usually allows more flexible and productive senses than a *static* account.

Formal Semantic Account

This account draws upon the solution of the traditional problem of type-shifting, where the semantic types of two composing words do not match so that some operations to fix the semantic type are introduced. Pustejovsky (1998) first extended this approach to polysemous word meaning. For each word, its semantic representation is not just a list of sense entries but an organized repository of feature structures, which includes lexical typing structure, argument structure, event structure, and qualia structure. Below is the basic template of the semantic representation of a word "x."[7]

[6]This echoes the causal account of names (Kripke, 1980) which answers in virtue of what does a name has its meaning. The answer is the causal chain that determines how the initial meaning is baptized and later transmitted.

[7]Adapted from (Pustejovsky, 1998, p.101)

(40)
$$\begin{bmatrix} \text{x} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = x_1 \\ \dots \end{bmatrix} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E1} = e_1 \\ \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{CONST} = \text{what } x \text{ is made of} \\ \text{FORMAL} = \text{what } x \text{ is} \\ \text{TELIC} = \text{function of } x \\ \text{AGENT} = \text{how } x \text{ came into being} \end{bmatrix} \end{bmatrix}$$

Not only can different senses be stored in the semantic structure, but new senses can also be generated productively in context from the qualia structure because it contains rich information about the function and constitution of the object referred to by the word.

For example,

(41) Tom threw the book.

(42) Tom read the book.

(43) Tom began the book.

(42) and (41) can be disambiguated directly from the information in qualia structure in (44) because "book" is a information.physical object, which has both these aspects. However, the sense of "book" in (43) as reading the book or writing the book is not directly stored. Instead, it goes through an operation called type coercion so that an appropriate meaning based on qualia structure can be generated productively. The verb "began" takes an argument of an event type, so the meaning of "book" in (43) has to be coerced to the event type and here it can either be the event of reading books or writing books based on the qualia structure (the TELIC and AGENT value in this case).

(44)
$$\begin{bmatrix} \text{book} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = x : information \\ \text{ARG2} = y : phys\_obj \\ \ldots \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} information.phys\_obj\_lcp \\ \text{FORMAL} = hold(y, x) \\ \text{TELIC} = read(e, w, x.y) \\ \text{AGENT} = write(e', v, x.y) \end{bmatrix} \end{bmatrix}$$

Generative Lexicon theory is named "generative" because Pustejovsky (1998) considers the productivity of polysemy to be the central target for his theory to explain. It employs several operations to explain how new cases of polysemy can be interpreted correctly. However, generative lexicon theory still has the problem of explaining the flexibility of polysemy because of the limited operation options and encoded information in the qualia structure.

Relevance Theory

Instead of viewing linguistic communication purely as rule and code-based interaction, relevance theorists (Sperber and Wilson, 1996; Carston, 2008) highlight the pragmatic inference involved.

Inference utilizes various sources of information. Pragmatic inference is usually achieved with encyclopedia knowledge that is out of the realm of proper linguistic knowledge. For example, when someone interprets the metaphor "Juliet is the sun," he or she has to understand the encyclopedia knowledge of the sun as the center of the solar system, it being warm, perpetual, and so on. Relevance theorists also suggest that the inference involved in human language production and processing is relevance based. Speakers and listeners tend to be geared to the maximization of relevance, which is defined as how great positive cognitive effects are achieved and how little processing effort is expended.

For relevance theorists (Falkum, 2011, 2015), polysemy is also explained by relevance-

based pragmatic inference. During the production and processing of polysemous words, a context-free concept is first accessed, and then relevance based inference steps in to fulfill the interpretation as the construction of an ad-hoc concept.

(45) Kate began a book.

(46) KATE BEGAN $[_{VP}[_{V^0}$ **e**$][_{NP}$ a book]] [8]

The first stage of processing of (45) produces the logical form of the sentence in which "book" is a context-free concept. The verb "began" requires an argument of VP denoting an event so $[_{V^0}$ **e**] is interpreted through relevance-based inference to be "write" or "read." Hence, the polysemous sense of "book" as "writing a book" or "reading a book" in (45) is achieved on the fly through the operation of pragmatic inference.

Because of utilizing various sources of information, relevance theories can capture the flexibility of polysemy very well because each different use of the same encoded standing meaning will initiate different inferences with slightly different sources of information. However, relevance theories lack a full explanation of the regularity of polysemy because any appropriate ad-hoc concept for a word can be inferred in the context which does not explain where the regularity is from. Put otherwise, treating polysemy just as ad-hoc concepts may eliminate the difference between polysemy and metaphor.

Recanati

Recanati (2017a, 2019) notices that polysemy has both the sides of pragmatic modulation and homonymy. On the one hand, polysemous senses seem to be derived from meaning modulation in context. On the other hand, they seem to be stored in our memory the way homonyms are. As a result, Recanati takes polysemy to be conventionalized modulation. There is a single conventionalized type meaning of polysemous word that constitutes an abstract schema or a network (Langacker, 1987) and this abstract schema/network guides how polysemous

[8] Example from (Falkum, 2015, p.89)

senses are derived in context, which Recanati names conversion into sense (Recanati, 2019). For Recanati, conversion into sense inherits most of the properties of meaning modulation in context, but it is mandatory in interpretation. Every polysemous word has to be converted into a specific sense in every context, while meaning modulation is usually understood to be optional, occurring only when there is such a need for changing the meaning. Recanati's theory of polysemy comes closest in spirit to the theory proposed in the next chapter, where polysemous senses are understood to be clusters of meaning modulations in context. Setting aside some differences, this new theory can be seen as an improvement of Recanati's theory of polysemy and it offers a more detailed explanation of how conventionalization of meaning modulation is realized as entrenched clustering.

1.5 Conclusion

In this chapter, I reviewed the features of polysemy — relatedness, structuredness, gradedness, flexibility, regularity, and productivity — and how well previous theories capture these features. I divided previous theories into two groups, *static* accounts and *operational* accounts. *Static* accounts encode polysemous senses directly and statically into the semantic representation – the hearer simply selects among the options, however stored, while *operational* accounts represent the polysemous senses as the outputs of the creative and productive operation on the encoded primitive meaning of a word.

In general, *static* accounts have the problem of explaining the flexibility and productivity of polysemy because only a quite limited number of discrete senses are encoded. On the other hand, *operational* accounts cannot explain the regularity of polysemy and may blur the difference between polysemy and pure ad-hoc meaning variation, such as live metaphor or metonymy.

Chapter 2

A New Theory of Polysemy

2.1 Introduction

In this chapter, I introduce my theory of polysemy, which covers the learning, representation, and processing of polysemy. My theory is situated within the *operational* accounts introduced in the last chapter so it inherits the general feature of operational accounts that it serves both as a constitutive account to explain what polysemy is and as a causal account of how polysemy arises from linguistic use. The causal story of polysemy is part of what constitutes polysemy. Therefore, this theory is a new attempt at an *operational* account to capture simultaneously the relatedness, structuredness, gradedness, flexibility, productivity, and regularity of polysemy, as summarized in the last chapter.

In this theory, polysemous senses are understood to be clusters and entrenchment in the modulation of word meaning in context. The "context" here should be understood as the full-featured context rather than the Kaplanian context as a set of parameters (Almog et al., 1989) introduced to capture the context-sensitivity of indexical terms. This context includes both the linguistic context, such as the surrounding linguistic items, and the non-linguistic context, such as the broad situation where the utterance is produced.

There are two basic components in this theory. First, meaning modulation happens all the

time in the processing and production of language. Word meaning is modulated by the surrounding words and phrases (linguistic context), and where the utterance is going on, how the speaker perceives the world, and so on (non-linguistic context). Thus, there are various kinds of meaning modulations contributing to the variation of word meaning in context, including but not limited to linguistic coercion, metonymy, and metaphor. Second, modulated meanings in context cluster and are entrenched if they occur frequently. The entrenched clusters leave traces on our long-term memory so that some modulations are more readily accessed in future linguistic processing and production, which we perceive as the literal polysemous senses.

These two components combined together serve as a causal account of how polysemy emerges in linguistic communication and how polysemy is processed. What is more important is that the causal account of how polysemy arises also constitutes what polysemy is. A word is polysemous if and only if entrenched clustering arises in its meaning modulation during its use, as I argue that only a causal account with the two components explain the features of polysemy introduced in Chapter 1. As a result, this new theory of polysemy serves both a constitutive and causal account of polysemy.

In the following, I first discuss the two components, meaning modulation and clustering, in detail. Then, I discuss how they are integrated to explain the learning, representation, and processing of polysemy. Last, I present how features of polysemy are explained in my theory.

2.2 Two Basic Components: Meaning Modulation and Clustering

The two components of this new theory, meaning modulation and clustering, are motivated by the need to capture both all the features of polysemy and how polysemy differs from other context-sensitivity of meaning simultaneously. Previous theories, as discussed in the last chapter, have explained some of them instead of all of them simultaneously.

2.2.1 Meaning Modulation by Context

Meaning modulation in context is first introduced by Cruse (1986) as a particular kind of semantic variation in context. This is where the meaning of a word token being "modified in an unlimited number of ways by different contexts, each context emphasizing certain semantic traits, and obscuring or suppressing others." "The variation ... caused by modulation is largely continuous and fluid in nature." This is contrasted with the context-sensitivity of indexical terms, such as "I" and "now," where the context is modeled as a set of predefined parameters, because the token meaning (referent) of indexical terms varies with respect to the discrete rules specified as the character of the term. Cruse also contrasts meaning modulation with contextual selection, which is another kind of meaning variation in context, for example, in homonyms, that "proceeds in discrete jumps rather than continuously."

Although Cruse has different views on what polysemy is, I follow his analysis that meaning modulation is continuous and fluid and argue that this is the mechanism underlying polysemy. The continuity of modulation makes it possible that there is much finer meaning in context than the common taxonomy of polysemous sense allows. Furthermore, I suggest that there are different kinds of meaning modulation in context, and they all contribute to the phenomena of polysemy and sometimes even simultaneously.

The following sections cover four common cases of meaning modulation in context. However, it is not intended to be an exhaustive list of mutual-exclusive options but only as a demonstration of the breadth of meaning modulation and its nature of continuity.

Prototypical Meaning Modulation in Context

One prototypical meaning modulation in context is the enrichment of meaning (Recanati, 2004). It is where the meaning of a target word token is made more specific in the broad non-linguistic context. Thus, it is not linguistically driven (bottom-up from linguistic cues), but top-down contextually driven.

(1) There is still beer in the fridge.

First, consider a situation in which a party is going on and someone is asking for beer to drink. The answer (1) means that there is still bottled or canned beer in the fridge. Second, when (1) is uttered when someone is cleaning the mess in the fridge, it may mean that there are still beer stains on the inside surface of the fridge. In both situations, the meanings of the two tokens "beer" are made more specific to be a particular amount of beer by the non-linguistic context of utterance, partying, or cleaning. Note that this variation of meaning is not driven solely by linguistic cues. It is the context of utterance — when, where, and what is happening in these situations and what these situations usually look like — that initiates the process of specification, so it counts as top-down enrichment of meaning.

Ad Hoc Linguistic Coercion

In addition to pure top-down modulation, word meaning can also be modulated by the linguistic contexts, such as the grammatical constructions surrounding the target word. For example, the argument of a verb can be coerced into a different meaning preferred by the verb.

(2) Jim began the tree.

"Began" in (2) requires an argument that is an event, so "tree" has to be coerced to an event related to trees. Notice that this coercion also nudges the target meaning towards a more appropriate direction rather than forcing a discrete jump to familiar meanings. The result of the coercion can range from a variety of events related to trees such as cutting, planting, drawing, or even complex events that involve multiple instances of it.

Metonymy

Metonymy is another kind of meaning modulation which converts the token meaning (referent) of the target word from an entity of one category to another entity in a different category. A classical example comes from Nunberg (1995).

(3) The ham sandwich is getting impatient

When (3) is uttered by a waiter, "ham sandwich" refers to the customer who ordered a ham sandwich because the food ordered by the customer is his or her most salient feature to the waiter. Metonymy is different from enrichment because the modulated token meaning jumps to a different domain of entities (food to person) while the token meaning is only made more specific (beer in general to canned beer) in meaning enrichment. It is thus used as an ad-hoc and non-literal tool in communication to refer to things in a different category.

Metaphor

Last, we have a more creative ad-hoc meaning variation in context — metaphor. To revisit an example:

(4) A work is a death mask of its conception.

In the literature of metaphor, "death mask" in (4) is called the source, and "work" is called the target. The association between the source and the target is so rare that it requires some creative modulation of the meaning of the source. Here, "death mask" is creatively modulated to mean an end. If it is the first time for someone to read (4), it may take a second to interpret what is meant by this sentence. In history, there has been a debate on whether metaphor involves a new propositional meaning besides its literal meaning (Davidson, 1978). However, most people acknowledge the existence of propositional meaning variation in metaphor. They believe that the production and interpretation of metaphor modulate the meaning of the source term with the help of some kinds of interpretive mechanism, such as analogy between two domains of concepts (Bowdle and Gentner, 2005), categorization of one thing into a different category (Glucksberg and Keysar, 1993), or pragmatic inference based on some relevance principle (Wearing, 2006). For the purpose of introducing my theory, it is only necessary to note that metaphor forms another category of continuous meaning modulation, and it involves comparing two even very different domains of conceptual knowledge in order to arrive at the intended interpretation. Above all, it is probably the most flexible and least restricted category among all varieties of meaning modulation.

Summary

Meaning modulation in context is various. As suggested already in the previous chapter, the examples above clearly do not exhaust all possibilities. However, there are several points about these cases above that are worth emphasizing here. First, these categories are not mutually exclusive. For example, (3) can be seen both as a case of metonymy and linguistic coercion, because the linguistic context, the predicate "impatient," also coerces the interpretation of "ham sandwich" to be an animate person besides the non-linguistic context of the restaurant where the utterance occurs. Second, modulation of meaning can involve very different mechanisms. Some are more bottom-up linguistically driven, such as linguistic coercion, while some are more top-down driven, such as enrichment of meaning. Third, modulation of meaning in context is continuous or at least fine-grained enough to be counted as continuous when we discuss linguistic meaning. Even the same kind of modulation of the same word type can produce different token meanings so the token meanings of polysemous words are finer and more flexible than discrete individuations of polysemy senses. As a result, continuity is needed to capture the fact that polysemy is graded in nature and allows the flexibility of use, which is absent in representation with a fixed set of discrete senses.

2.2.2 Clustering of Meaning Modulation

Meaning modulation in context by itself is not equal to polysemy. For instance, polysemy isn't as ad hoc as other kinds of meaning modulation. As a result, the second component in this new theory has to be introduced, which is the clustering of meaning modulation in context.

Semantic Space Model

Before introducing the idea of clustering, it is necessary to first introduce the idea of semantic space (Churchland, 1993; Lenci, 2008). One way to think about meaning is to conceive it as a point in a high-dimensional space. Each dimension in this space represents a semantic

feature such as ANIMATE, CONCRETE, and so on. These feature dimensions construct a high-dimensional space in which each possible meaning occupies a point in the space. This way of thinking about meaning provides some powerful tools to characterize the structure and relationship of linguistic meanings. For example, geometric relationship, such as Euclidean distance, represents semantic relationship: the closer two points are, the more similar these two meanings are.

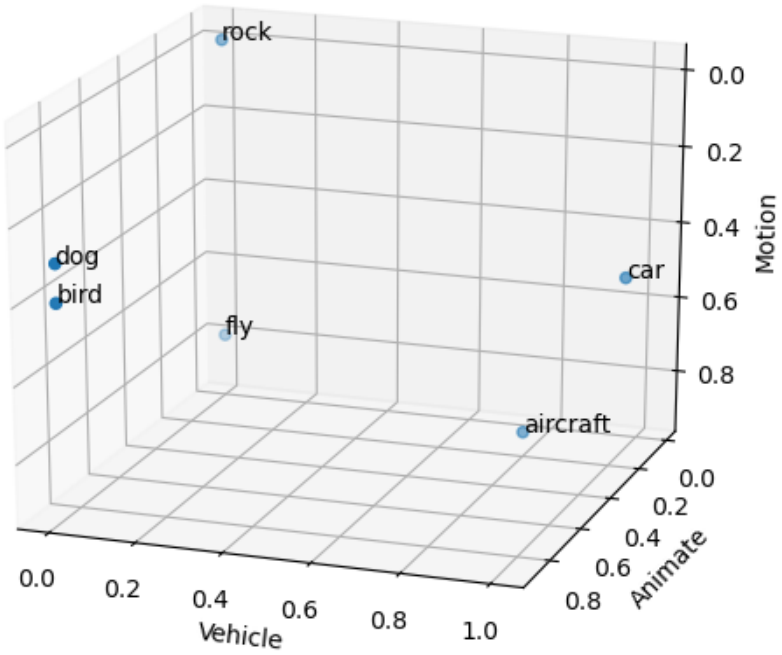


Figure 2.1: Visualization of a Three Dimensional Semantic Space

In Fig. 2.1, each point represents the meaning of a word. The coordinate represents the value of each semantic feature. For example, the meaning of "fly" has very high motion value but very low animate and vehicle values. The semantic similarity between "bird" and "dog" can be represented as the distance between the two points, which is closer than the distance

between “bird” and “rock.”

Clustering in the Semantic Space Model

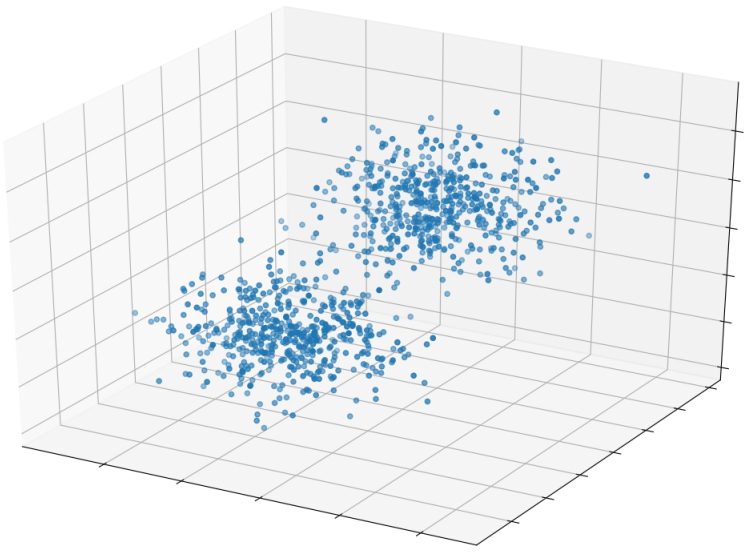


Figure 2.2: Visualization of Clustering of Meaning Modulation

When this semantic space model is applied to meaning modulation in context, each point in the space represents a particular instance of meaning modulation of the same word (token meaning). In Fig. 2.2, every point is an instance of meaning modulation of the same word (token meaning). Since clustering is defined to be similar things positioned closely together, the clustering of meaning modulation in context can be understood to be the existence of dense clouds of points in the semantic space. In Fig. 2.2, points are not distributed uniformly. Instead, they form two clusters of points. Each cluster represents a group of similar modulations in context, which corresponds to our intuitive notion of a polysemous sense. In addition to the individuation of senses, clustering provides a way to represent sense structures. It captures the relatedness between senses by measuring the distance between clusters. It can also capture

the hierarchical structure of senses and subsenses with the help of hierarchical clustering algorithms. As a result, the polysemous senses and their structures emerge from the clustering of modulated token meanings in the semantic space.

Notice that polysemous senses understood in this way are not discrete. First, different senses may overlap with each other just as clusters may overlap, so that there are modulations of meaning belonging to multiple polysemous senses. The "school" in (5) belongs to both the institution sense and the building sense. The "school" in (6) belongs to both the professional school sense and institution sense.

(5) The school has shut itself down.

(6) Tom went to medical school.

Second, there can be different ways to individuate senses based on how fine one wants the cluster to be.

(7) Mike Trout is a baseball star.

(8) Leonardo DiCaprio is a movie star.

Whether to categorize the "star" in (7) and (8) as the same sense does not have a determinate answer because it is based on how fine one wants the cluster of meaning modulation to be.[9]

Third, the distribution of meaning modulation of a word captures whether it is polysemous or monosemous. The meaning modulation of a polysemous word has a multimodal distribution as in Fig. 2.2, which is in contrast with a monosemous word in Fig. 2.3. The difference between a unimodal distribution and multimodal is also graded, which results in the graded difference between monosemy and polysemy.

[9]This echoes the practice in applying clustering algorithms that the parameter k (the number of clusters in k-means algorithm) is usually determined based on our goals of clustering instead of being an objective value.

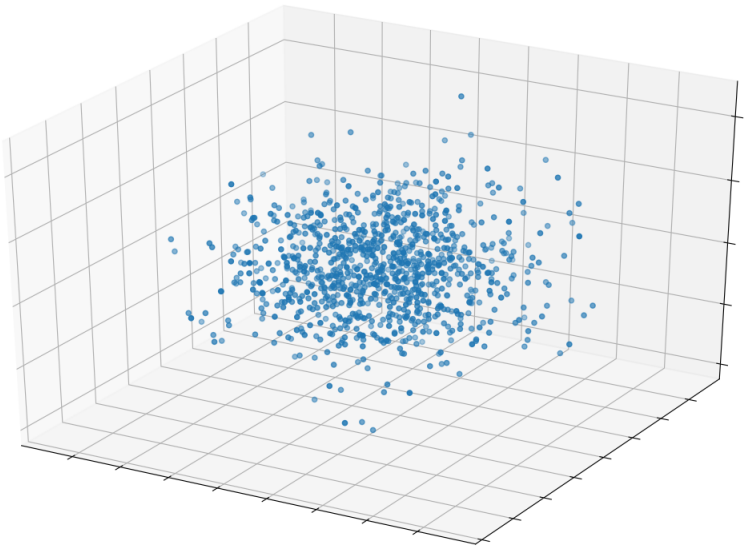


Figure 2.3: Visualization of Clustering of a Monosemous Word

Entrenchment from Clustering

The clustering of meaning modulation does not only emerge from uses of polysemous words, but also has an effect on future cognition in terms of processing and producing polysemous words, through the mechanism of entrenchment (Langacker, 1987). Entrenchment is the process of repetitive exposure and rehearsal of an item or items. It facilitates the future processing of these items and establishes the entrenched items as a unit (e.g., an idiom). As for polysemy, the frequency of the meaning modulations in one cluster (token frequency of a sense) will have an effect on how fast and effortless the sense is processed in the future. Thus, senses with higher frequency are more entrenched. This is well established in experiments on the frequency effect of polysemy (Klein and Murphy, 2001, 2002). However, clustering provides more than this token frequency metric. It captures the number of different clusters — type frequency and the variability of each different modulation within the cluster — cluster variability.

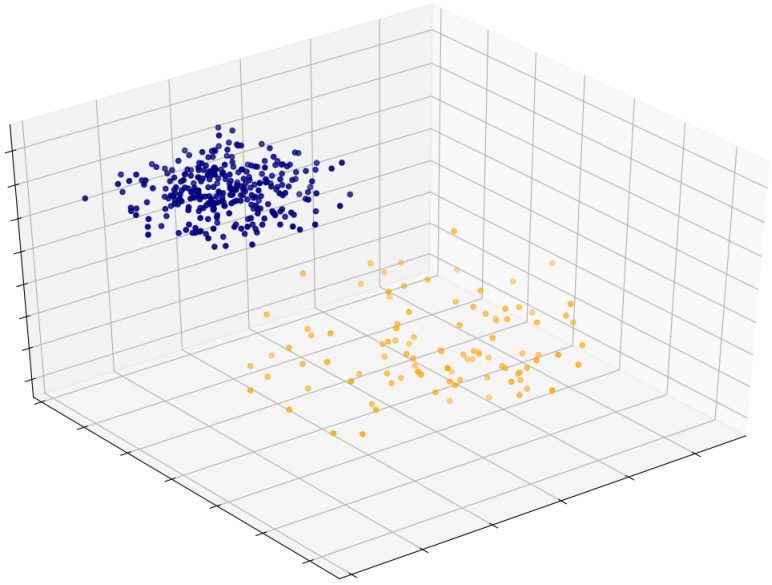


Figure 2.4: Visualization of Clustering Properties

These three properties of clustering have critical roles in explaining the productivity of polysemy, as studies in usage-based linguistics have shown (Tomasello, 2003; Bybee, 2010; Goldberg, 2019). A small and dense cluster (the purple cluster in Fig. 2.4) means low variability, which constrains future modulations of this polysemous word to be tightly consistent with previous modulations. Hence, future modulations of this sense will more likely lie in the range of this purple cluster. In contrast, a wide and loose cluster (the yellow cluster in Fig. 2.4) means high variability. Thus, future modulations will lie in a wider range in the semantic space. The type frequency of a cluster depends on the number of points in the cluster, so purple cluster has higher frequency while yellow cluster has lower frequency. Higher type frequency will constrain the future productive use more than the lower frequency one. Another way to think about this is based on dynamic system theory. A tight or frequent cluster forms a strong attractor in the process of meaning modulation, so that future modulation will gravitate towards the center of this cluster. Furthermore, a polysemous word with more clusters (higher type fre-

quency of 4) in Fig. 2.5 is more productive than a polysemous word with fewer clusters, such as the word represented in Fig. 2.2

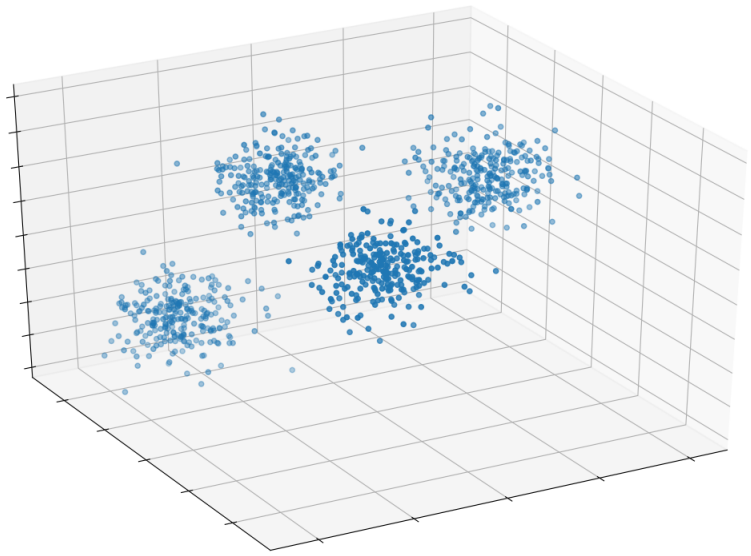


Figure 2.5: Visualization of Meaning Modulation with 4 clusters

2.2.3 Example of Polysemy Emerging from Two Components

A full analysis of polysemy in terms of the two components of meaning modulation and clustering cannot be achieved without the computational model introduced in the next chapter. However, I present some examples of polysemy in this section as a preliminary demonstration of how these two components work together for some polysemous words.

Clustered enrichment: the word "rabbit" usually refers to the rabbit animal. Sometimes, enrichment of meaning makes a token of "rabbit" to mean a specific part of a rabbit. For example, the "rabbit" in "rabbit coat" is enriched to be the fur of rabbits because of the existence of the practice of using rabbit fur as a clothing material. This contextual enrichment clusters and is entrenched so that future tokens of "rabbit coat" are more readily modulated to mean

the coat made of rabbit fur, so it becomes a new polysemous sense for the word "rabbit." In comparison, less entrenched "_____ coat" phrases allow more flexible interpretation, such as "panda coat" or "cat coat."

Clustered coercion: names are paradigmatic cases where linguistic coercion happens. "Reading Shakespeare is enjoyable" means that reading the works of Shakespeare is enjoyable rather than that reading him as a person is enjoyable (even though it is a possible reading at times). The verb "reading" coerces its argument into the interpretation as readable things. The names of authors are frequently coerced in this kind of verb bias context so that these kinds of meaning modulation cluster, are entrenched, and end up to be a literal polysemous sense of author names.

Clustered metonymy: The White House is the place where the president of the United States of America lives and works, so it serves as a symbol of the US government. However, the frequent use of "White House" as a metonym meaning the government of the USA occurs very frequently and clusters so that this meaning is entrenched as a polysemous sense of "White House," and is probably used as frequently as its original sense.

Clustered metaphor: "head" originally means the upper part of an animal. However, it also refers to the leader of an organization. Conceptual metaphor theory (Lakoff and Johnson, 2008) explains this metaphor as the mapping between two domains of an animal body and an organization where both the real head and the head person is the part that is in charge of the whole. This conceptual mapping between these two domains has been entrenched so that the word "head" acquires the polysemous sense of "person in charge of an organization."

In sum, polysemy can involve very different varieties of meaning modulations, but what unites polysemy as a category is the mechanism of clustering and entrenchment. A more detailed presentation of cases of polysemy requires a quantitative measure of the clustering effects of meaning modulation. This is discussed in the next chapter.

2.3 Learning, Representation, and Processing of Polysemy

With the two basic components being introduced, I now discuss how these two components are integrated into my theory of learning, representation, and processing of polysemy.

In terms of learning, a polysemous word is learned in the same way as a monosemous word. A different learning procedure is not required because the difference between polysemy and monosemy is a gradient of the distribution of meaning modulation in context rather than a categorical difference in the representation of meaning as in *static* accounts. Polysemous words can be learned through mechanisms of normal associative learning between words and concepts (Barak et al., 2019). They can also be learnt through prediction-based learning (Elman, 1990; Clark, 2013; Rabagliati et al., 2016), which is captured in the next chapter using a computational model. This learning task — predicting future words based on previous words — forces the learned semantic representation (type meaning of a word) to encode distributional information about polysemous words. These kinds of distributional information include word co-occurrence, which is seen as an important piece of linguistic meaning in most distributional semantic models (Lenci, 2008). They are based on the principle that, as a general rule, "difference of meaning correlates with difference of distribution" (Harris, 1954). Other approaches to word learning also play roles in polysemy learning. The key here is that this theory of polysemy does not require an additional procedure for learning polysemous words because the semantic representation for polysemous words is not categorically different from monosemous words as in the *static* account.

When it comes to representation, the semantic representation (type meaning) of a polysemous word is the superimposition of all its modulated token meanings in different contexts, which is the same for a monosemous word. The superimposition of different modulated token meanings utilizes distributed representations, which encode semantic information in multiple neuron-like units. Any piece of semantic information is represented by these units together instead of by a particular unit. As a result, this theory does not represent different senses statically as discrete entries as in the *static* account, so it avoids the lumping-or-splitting problem

of sense individuation, which says that assigning token meanings into discrete sense bins is impractical and difficult. This distributed superimposition contains but is not limited to the referential information between words and referred objects as well as the fine-grained word co-occurrence information. A polysemous word used in different contexts will have different referential and word co-occurrence information, and these different pieces of information are superimposed into a single representation rather than encoded separately. In sum, a polysemous word has a single distributed representation as its type meaning, which is abstracted from all its different modulated token meanings in context and contains both referential and distributional information.

What makes a word polysemous is *how* its context-free semantic representation (type meaning), as previously discussed, is modulated by context during production and processing. The comprehension and production of a word are achieved through contextual activation (modulation) of its superimposed semantic representation. This is the place where polysemy emerges. This activated semantic representation is modulated by both the linguistic contexts and non-linguistic contexts, so different kinds of meaning modulation could all shape the token meaning of a polysemous word in context, such as enrichment of meaning, linguistic coercion, and so on. Notice that these modulations occur at the same level as the semantic composition of word meaning, so polysemy does not involve a second stage of comprehension as in Grice (1991)'s model. The distribution of contextual modulations determines whether a word is polysemous and how polysemous it is. The modulations of polysemous words follow a multimodal distribution in which multiple dense clusters are formed by similar modulations of the same polysemous words. These clusters of similar modulated token meanings are entrenched in one's memory after exposure and rehearsal, so they are more easily and effortlessly activated in the future. As a result, clusters of modulations further strengthen the probability of future modulations lying in the entrenched clusters so that the multimodal distribution is constantly retained.

In sum, polysemy is neither directly encoded as multiple different senses in a semantic

representation as in *static* accounts nor just pure ad-hoc modulation of meaning as suggested in *operational* accounts. Instead, polysemy arises from clustering and entrenchment of different contextual activations (modulations) of a single distributed semantic representation in different contexts.

2.4 Handling Features of Polysemy with the New Theory

In this section, I briefly summarize how my new theory captures the features proposed in the first chapter and how my theory explains the difference between polysemy and other context-sensitivity of meaning.

First, the **relatedness** between senses is captured by the distance between the sense clusters in the semantic space. Different polysemous senses of a word are closer to each other than the different readings of a homonym. The **structure** of senses is captured by the geometric relationship between the sense clusters in the semantic space, which is further discussed in the computational modeling in the next chapter. Second, the **gradedness** of polysemy is captured by the continuity of meaning modulation in context. Modulation is very fine-grained and highly sensitive to nuanced details in contexts, so meaning modulation can cover the nuanced differences that are not capturable by a *static* account of polysemy. The **flexibility** of polysemy also arises from the continuity of meaning modulation: it allows uses of polysemous words with in-between senses. Third, the **regularity** of polysemy is captured by clustering and entrenchment of meaning modulation. Regular uses of polysemous senses cluster and are entrenched so that future processing of polysemy will more likely align with the already clustered modulations. Furthermore, regular polysemy can be captured as the geometric structure of clusters in the semantic space, which is discussed in detail in the next chapter with the help of computational modeling. Last, the **productivity** of polysemy is captured by the entrenchment of the clustering in meaning modulation. Statistical measures of clustering (type frequency, token frequency, variability) in the semantic space capture the patterns of different clustering

entrenchments, which contribute to the difference in the productivity of different polysemous words.

To come at the point another way, the difference between polysemy and other kinds of semantic context-sensitivity can also be explained in my theory of polysemy. As noted, the difference between **homonymy** and polysemy is captured by the distance between emerged sense clusters in the semantic space. The sense clusters in homonymy have a longer distance between each other than those in polysemy. The difference between **indexicality** and polysemy is explained by the plurality of meaning modulation involved in polysemy. There is no single mechanism for polysemy as in the parametrized context-sensitivity of Kaplanian indexicality. Last, the difference between **ad-hoc meaning variation** and polysemy is explained in terms of the clustering effects of meaning modulation in context. Ad-hoc meaning variation is usually one-off without repetitive and sustained uses so it does not cluster in the semantic space and is not entrenched in our long-term memory. That is why, in contrast to polysemy, ad-hoc meaning variation does not constitute a part of our literal meaning.

In sum, the modulation of word meaning in context, on the one hand, captures the different features of polysemous senses. On the other hand, the clustering of modulation provides an explanation of the regularity and productivity of polysemy with the help of the semantic space. As a result, this new theory of polysemy provides a way to capture the regularity of polysemy without losing the flexibility of polysemous meaning. An important consequence of this new theory is that a constitutive account of what polysemy is lies in, as least partially, the causal account how polysemous senses arise and are processed because only the causal components of meaning modulation and entrenched clustering offer an explanation of why polysemy, as a linguistic phenomenon, has the feature of relatedness, structuredness, gradedness, flexibility, regularity, and productivity. In the next chapter, the implementation of clustering and modulation in a computational model provides quantitative measures for empirical confirmation in both corpus and behavioral studies, which provides further evidence in support of this new theory of polysemy.

Chapter 3

A Computational Model and Experiments

3.1 Introduction

In this chapter, we[10] implemented the theory of polysemy proposed in the last chapter in a recurrent neural network (RNN) model to demonstrate this new theory of polysemy and assess it with corpus studies and behavioral experiments. The computational model provides a source of how different components of the theory work together causally, because the details of clustering go beyond the scope of verbal description. Second, we plan to collect evidence for the new theory proposed in Chapter 2 and against the *static* account of polysemy. The static account entails a different causal account of how polysemous words are processed, so the computational model could provide evidence in terms of which causal account fits the behavioral data of polysemy processing better. Third, we want to showcase that connectionist models like ours can learn rule-like patterns captured by symbolic models, such as the shared patterns in regular polysemy.

There has already been a large number of semantic models that focus on polysemy. First, there are distributional semantic models that have proven their capacity to capture semantic information and psycholinguistic findings (Lenci, 2018). These models range from classical

[10]This chapter is based on a collaboration with Dr. Marc F. Joanisse.

versions, such as Latent Semantic Analysis (Deerwester et al., 1990) and Hyperspace Analogue to Language (Burgess, 1998), to contemporary neural network based versions, such as Word2Vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014). Most distributional semantic models do not distinguish different polysemous senses of a single word because they abstract each context of each occurrence into a single vector representation as prototype models (Jones, 2019). Second, there are attractor network models representing polysemous senses as attractor basins in a high dimensional semantic space (Rodd et al., 2004; Armstrong and Plaut, 2008, 2016). These attractor models have both the representational aspect of encoding senses as attractors and the processing aspect of following a trajectory into the basin in the high dimensional space. However, most attractor networks, such as Hopfield network (Hopfield, 1984; Hopfield and Tank, 1985), implement point attractors that discretize polysemous senses, and they use hand-coded meaning vectors rather than learning them in a natural corpus (Details in the discussion section). Third, there are the examplar models of polysemy, such as the Instance Theory of Semantic Memory (ITS) (Jamieson et al., 2018). It differs from the previous two models in that it does not abstract instances of the meaning of each occurrence of a word, neither into a single word level nor into a polysemous sense level. Instead, this model stores a representation for each linguistic instance it is exposed to, and the meaning of a word token is constructed on-the-fly from these stored instances.

We apply Recurrent Neural Network to model this new theory of polysemy, because RNN implements naturally the idea that the meaning of a word token is modulated by the previous context (influenced by the previous recurrent layer). Furthermore, RNN represents the context and the modulated meaning in a continuous way without discretizing contexts as in the attractor models. In attractor models, the contexts are either biasing one sense or another sense without in-between contexts. The continuous representation of contexts that RNN utilizes enables us to examine the emergence of the clustering structure of both the contexts and word meanings by focusing on the internal representations the model develops as it processes different senses of polysemous words. Furthermore, RNNs have been widely used in psycholinguis-

tics to model language learning and processing (Elman, 1990; Christiansen and Chater, 1999; Linzen et al., 2016; Futrell et al., 2019). Through exposure to sentences, an RNN can learn and generate a probability distribution for the next word given a sequence of previous words $P(W_n \mid W_1, W_2...W_{n-2}, W_{n-1})$. This task can be seen as a computational implementation of the prediction based learning previously mentioned. In this task, a model learns knowledge of languages to predict incoming words accurately. More specifically, language models implemented by RNNs have been shown to be particularly powerful for learning word meanings (Elman, 2004, 2009, 2011; Kocmi and Bojar, 2017). Peters et al. (2018) have also shown its power for learning contextual meanings of words. Besides these studies, our RNN models have similarities with a particular branch of RNN model called Sentence Gestalt model (McClelland et al., 1989; Rabovsky et al., 2018; Hoffman et al., 2018). They applied simple recurrent networks to directly model sentence processing and concept learnings. Our model could be seen as a modern version of the sentence gestalt model trained on a larger and more natural corpus.

We also recognize that there are limitations to our choice of models. Our model implements word meaning based only on distributional information, so there is a traditional worry of the symbol grounding problem (Harnad, 1990). That is how words connect to things in the word if there is only co-occurrence based distributional information encoded in the model. We are aware of this problem, but this is not a problem of our own. Instead, It is a problem for all models based on distributional semantic information. We do not deny the importance of word-object referential information in semantics, and contextually driven modulation of meaning in context is not at odds with utilizing referential information in semantics. It is just that we have not embedded referential information in our computational model. On the other hand, implementing this grounded and referential information (McRae, 2004) is possible within our modeling framework, as shown in Xu et al. (2015)'s work on multimodal recurrent neural network.

In sum, we apply the RNN model to examine the meaning modulation of polysemous words and their clustering. By focusing on both the internal representations and the next word

prediction, which the model develops as it processes polysemous words, we can assess the new theory proposed in the last chapter with both human-annotated corpora and behavioral experiments.

3.2 Model Architecture — Long Short-Term Memory

The RNN used here is a Long Short-Term Memory model (Hochreiter and Schmidhuber, 1997), specifically based on Merity et al. (2017)'s version of LSTM because of its good performance in learning linguistic generalizations about individual items within a large corpus of exemplars. The formulae of LSTM are listed below in Equations (3.1).

$$
\begin{aligned}
i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \\
f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \\
g_t &= \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \\
o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t) \\
\sigma(x) &= \frac{1}{1 + \exp(-x)} \\
\tanh(x) &= \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}
\end{aligned}
\tag{3.1}[11]
$$

The model is presented with sequences of words forming sentences. At each step, a word is input to the model, and the model is trained to predict what word follows it by activating the correct representation of the word that is predicted to follow it. Notably, word prediction

[11] These are the standard LSTM formulae "where h_t is the hidden state at time t, c_t is the cell state at time t, x_t is the input at time t, h_{t-1} is the hidden state of the layer at time $t-1$ or the initial hidden state at time 0, and i_t, f_t, g_t, o_t are the input, forget, cell, and output gates, respectively." W and b are connection weights and biases respectively. σ is the sigmoid activation function, $\odot$ is the Hadamard product, and tanh is the hyperbolic tangent activation function. (Paszke et al., 2019)

is rarely accurate because of the productive nature of language. Thus each predicted output represents the network's best guess as to the identity of the following word. By presenting the network with many sentences over the course of training, it learns to output the corresponding probability distribution of predicted words in context, reflecting the probabilistic structure of the corpus that it is trained on. This training task of prediction forces the model to learn both the generalized semantic representations of words (which are operationalized here as its distribution in the context of all other words in a corpus) and how word meanings are modulated by context.

As shown in Fig. 3.1, word inputs and predictions are coded in the model using a localist scheme on the input and output layers where a single unit is used to uniquely encode each unique word in the corpus vocabulary. This model also has a 600-unit embedding layer representing the context-free meaning of each word presented to it. No single unit in it alone represents a specific word meaning. Instead, it is the activation pattern of all these 600 units that represents[12] the context-free meaning (type meaning) of each word as a distributed representation. Next, a 600-unit recurrent layer that is fully connected with the embedding layer represents the modulated meaning of each word token in its context (the words before the target polysemous words in a sentence), and these meanings are also distributedly represented. The recurrent layer is self-connected and parametrized as in Equations (3.1). The target for the output layer on each trial is a one-hot localist representation[13] denoting the actual next word in the sequence. Because the actual prediction cannot be achieved at a perfect level, the network ultimately learns to produce a set of softmaxed (3.2) activation levels, at the output, in which multiple words are predicted to varying degrees of certainty, reflecting the probability distribution of the next word given previous inputs.

[12]What we mean by *representing* here and other places in this chapter is that the activities of layers of units are operationalized as linguistic meaning in our model. The distributional information captured by these activities is strictly speaking not meaning, but it is useful for examining the modulation and clustering effects, and hence providing evidence for my theory of polysemy.

[13]One hot representation is a vector representation in which only one dimension is number 1 while all other dimensions are 0. For example, a five dimension one hot vector could be $(0, 1, 0, 0, 0)$ or $(1, 0, 0, 0, 0)$.

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{3.2}$$

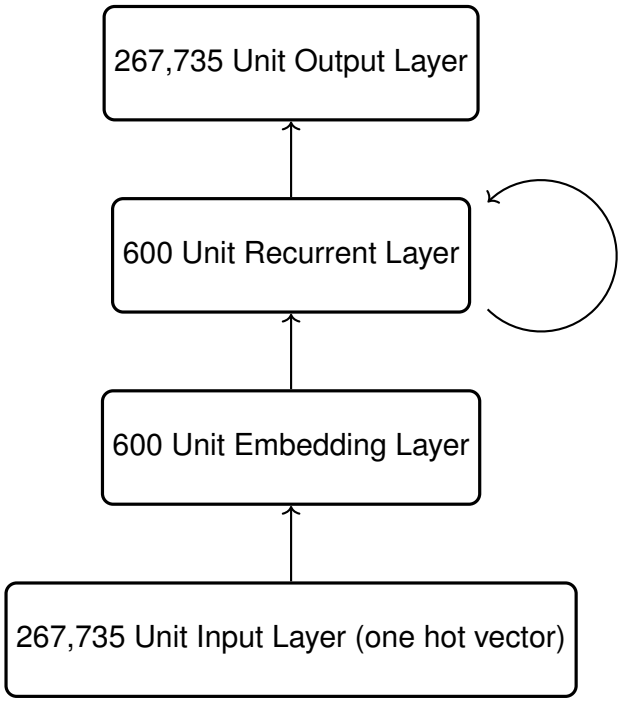


Figure 3.1: Model architecture. Texts are presented to the model word-by-word. Each word is presented individually on the input layer, and the model's task was to predict the following word in the sequence, so the input and output layer has the same size as the training vocabulary (Including 267,735 words. Each different surface form of the "same word" is treated as a different word. "Walked" and "walk" are treated as separate word tokens.). This was achieved by passing activation through two intermediary layers, one static and one recurrent.

The model runs sequentially, and therefore it can be unrolled along the temporal dimension as in Fig. 3.2. At each time step t, a word W_t in a text is fed into the embeddings layer as an input to activate the context-free semantic representation E_t of the word and then E_t feeds into the recurrent layer R_t. The recurrent layer R_t is fed by both the embeddings layer of time t, E_t, and previous recurrent layer of $t-1$, R_{t-1}. R_t represents the context-specific meaning of word W_t as a modulated activation of the semantic representation E_t. For the output layer, each node represents the probability of a word as the next word, so the activation of the output layer represents a discrete probability distribution with respect to every possible word at the

next word position. Each word in a sentence is presented in turn, such that the context-sensitive recurrent layer is influenced not only by the current input word but also the weighted influence of prior words in the sentence. An example of this LSTM model processing the sentence "The electrical power is ... a circuit" is also provided in Fig. 3.2. At the time step 1, the word "The" is input into the model as W_1 and it is "modulated" as R_1 but there is no previous context. The model produces an output O_1 as a softmax distribution of the predicted next word. The ground truth for this prediction is "electrical" so the predicted distribution should peak at "electrical." At the time step 2, the second word "electrical" in the sentence is input to the model. The context-free meaning E_2 is modulated to be R_2 by the previous context R_1 and an output of next word prediction O_2 is produced from R_2. The model continuously processes the sentence one word at a time until it reaches the end of the sentence at time t.

3.3 Model Training

The training corpus used in this model is Wikitext-103 (Merity et al., 2016). It is a collection of over 100 million word tokens from a set of verified Good and Featured articles (n=28,595) on Wikipedia. The included articles "have been reviewed by humans and are considered well written, factually accurate, broad in coverage, neutral in point of view, and stable" (Merity et al., 2018). The vocabulary of this corpus contains 267,735 words.[14] This corpus is already preprocessed and divided into training, validation, and test sets by Merity et al. (2016). The validation set is used to check the progress of the model during training, and the test set is used to test the model when all training is done. Neither the validation nor the test set is used for model training. The training set contains 28,475 Wikipedia articles of 103 million word tokens. Both the validation and testing sets contain 60 Wikipedia articles of around 200 thousand word tokens.

During the model training, the training set was divided into 9322 batches of equal-sized

[14]Each different surface form of the "same word" is treated as a different word, so "walked" and "walk" are treated as separate word tokens.

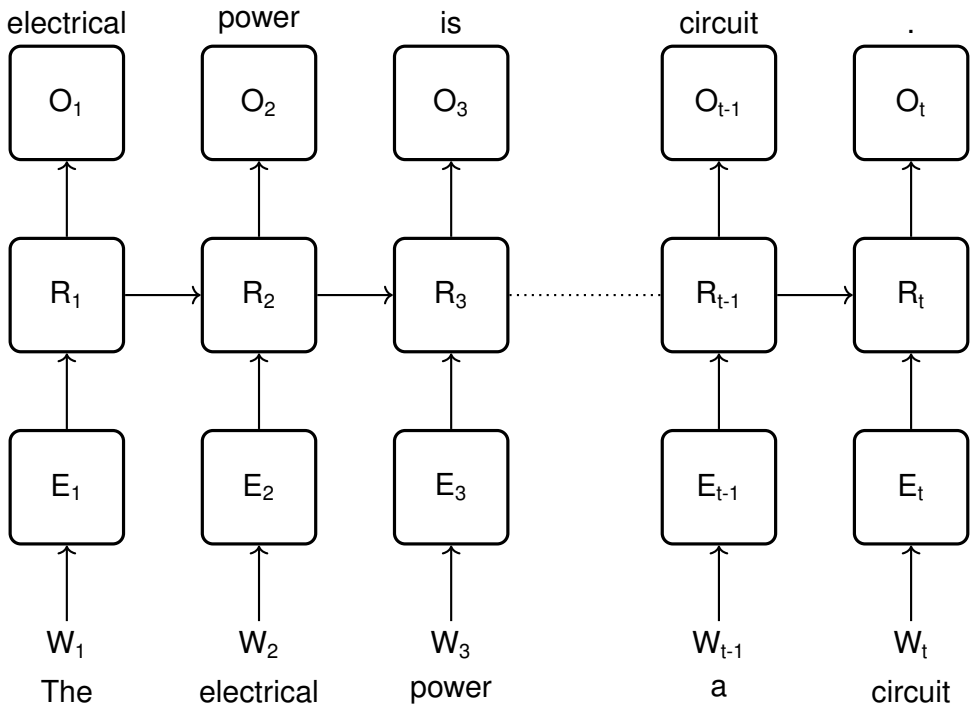


Figure 3.2: Unrolled Architecture depicting how the network represents a temporal sequence of word inputs, and uses prior information to predict upcoming words (and punctuations). Since word prediction is imperfect because of the creative and productive nature of language, the model learns to output a probability distribution of possible next words in a sequence. The model learns to form internal representations on the Recurrent layer that encode contextual information about an input word and its preceding inputs in order to maximize accuracy in predicting subsequent words.

chunks without random shuffling and the parameters of the model were updated after each batch is presented. The model was initialized with random weights with uniform distributions bounded from −0.1 to 0.1 and trained to predict the next word for each word in the set by adjusting weights using the backpropagation-through-time algorithm. A cross-entropy loss function (3.4) was used to compute the difference between the predicted next word and the actual next word in the corpus, so that the model could update its predicted next word probability distribution in order to maximize the probability of the actual next word. During the first 36 epochs, the model was optimized using the Adam algorithm with 0.001 learning rate. In the next 32 epochs, average stochastic gradient descent was used to fine-tune the model. We trained the model until the prediction perplexity (3.3) was no longer decreasing, so the model could not predict words better. The training was stopped after 68 epochs, at which point the model had reached perplexity of 52.68.

$$\text{Perplexity}(W) = \sqrt[n]{\frac{1}{P(w_1 w_2 ... w_n)}} \tag{3.3}$$[15]

$$\text{CrossEntropyLoss}(x, class) = -\log\left(\frac{\exp(x_{class})}{\sum_j \exp(x_j)}\right) \tag{3.4}$$[16]

3.4 Model Assessment with Linguistic Corpora

We first assessed our trained LSTM model against corpus data in terms of the clustering of meaning modulation. We examined what kinds of clusters emerge and how well these emerged

[15] In the formula, W represents the target sentence, and w represents each word in the sentence, so the perplexity of a sentence is defined as the inverse probability of the target sentence, normalized by the number of words. In general, perplexity is an information theory measure of how well a probability distribution (here, produced by the trained model) predicts a sample. A low perplexity of language modeling indicates good performance of a model in predicting words. Because the model is trained to predict the training data, good training leads to assigning high probability to sentences in the training data (low perplexity). If the model also generalizes well to data unseen (test data), it should also yield low perplexity for the test data.

[16] This loss function is the cross entropy combined with the softmax activation. The variable *class* denotes the positive class.

clusters explain the gradedness, structuredness, and regularity of polysemy senses.

3.4.1 Modeling Senses as Clusters in Meaning Modulations

We assessed (1) whether the model had developed internal representations of polysemous word meanings, (2) how it would process the polysemous word in context to achieve the correct meaning modulation, and (3) whether the clustering of modulations would match the human intuition of sense individuation. For each word in the corpus vocabulary, the model learned[17] a 600-dimensional vector in the 600-unit embeddings layer, which represents the context-free meaning or type meaning of a word. In contrast, for each word fed into the model, the model activated a potentially unique 600-dimensional vector in the recurrent layer as its modulated meaning in context because the activation is a function of both the previous context and the context-free meaning of the current input word. Different tokens of the same word will have different representations within the recurrent layer but not within the embeddings layer. Activation in the recurrent layer reflects both the information of the previous context and the current word. In this modeling experiment, we assessed both how the previous context before the target polysemous word (the recurrent layer at $t-1$, t is when the target polysemous is input to the model) and the target polysemous modulated by the previous context (the recurrent layer at t) to examine the roles of contexts and context-free word meanings.

In order to probe the learned representation of polysemous words and how polysemy emerges from meaning modulations during processing, we tested our model against a sense annotated corpus (Evans and Yuan, 2017) on which our model had not been trained. This corpus is composed of MASC (Passonneau et al., 2012) and SemCor (Mihalcea, 1998) datasets and manually annotated with sense definitions according to the New Oxford American Dictionary.

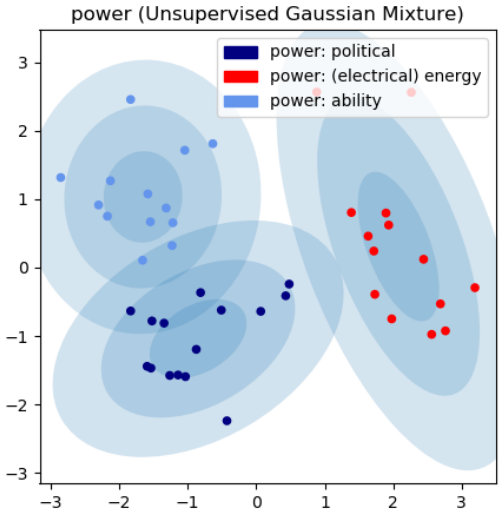
[17]Representation learning is a terminology used in the field of machine learning. It means that, by training a computational model on a specific task, the model extracts the features useful for the task from the data and **generalizes to data not trained on**. Here, we mean the recurrent neural network develops a set of weights so that given a fragment of a sentence, the model correctly predicts the next word in the sentence. The weights also produce activations of units in the intermediate layer, as the representation of useful distributional features.

We fed the corpus into the model as sequences of words and extracted the recurrent layer at the time of annotated polysemous words as 600-dimensional vectors (activation values of the units in the recurrent layer), reflecting their modulated meaning in their contexts. We also extracted the recurrent layer before the annotated polysemous words come in, reflecting the contexts modulating this target word. With the annotation of senses in the corpus, we obtained the ground truth distribution of each sense of each polysemous word and sense-biasing context.

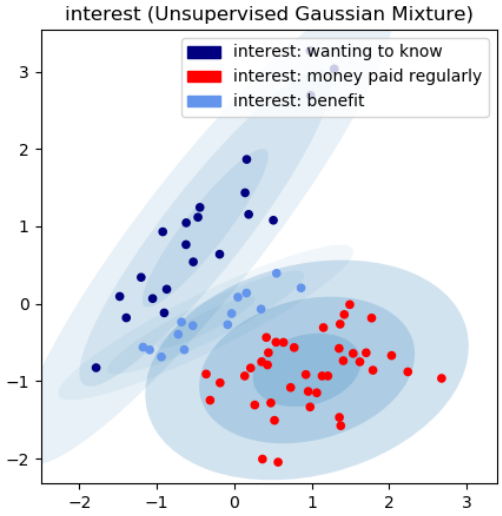
To examine the clustering structure of both the meaning modulations and previous contexts, we applied unsupervised Gaussian mixture models to cluster the meaning modulation vectors (recurrent activation at t when the target polysemous word is input) and context vectors (recurrent activation at $t - 1$). The Gaussian mixture model will find the n most likely Gaussian distributions underlying the data (the distributions from which the data are sampled from), so it finds the n mostly likely sense clusters given the contextual token meaning vectors provided to the model. This tested the theory that polysemous senses emerged naturally as clustered modulations from linguistic processing and these emerged clusters are consistent with human intuition. We tested quantitatively how accurately these clusters were formed according to the ground truth annotations in the corpus. The sense clusters were obtained by unsupervised Gaussian mixture modeling of n components (n equals the numbers of senses) with the Expectation-Maximization algorithm. In order to make the clustering algorithm work properly[18], we randomly sampled the meaning vectors so that there was an equal number of words belonging to each sense (cf. during training, where sense bias was allowed to reflect what was present in the training corpus). As a result, each sense cluster obtained was a 600d Gaussian distribution, and it represented how likely this sense had this particular meaning in a context. We also applied the same Gaussian mixture model to the PCA compressed 2-dimensional meaning vectors for visualization as shown in Fig. 3.3.

In order to quantitatively test the hypothesis that clustering consistent with human intuition

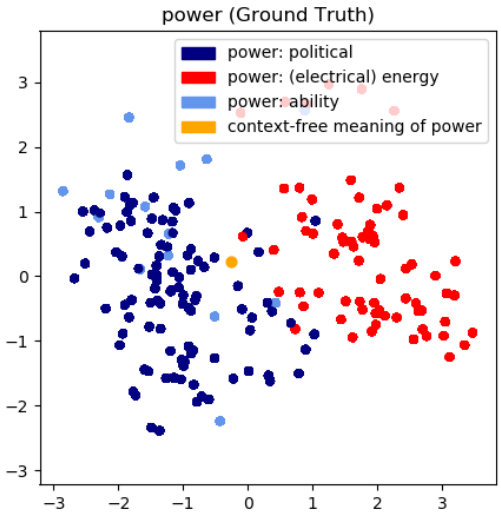
[18]Accurately clustering imbalanced datasets is a challenging problem in statistics and machine learning (Krawczyk, 2016). The standard approach in the field is to preprocess the data to make the dataset balanced. In order to prevent the clustering algorithm performing poorly and assess the clustering accurately, we preprocessed the data to make them balanced.



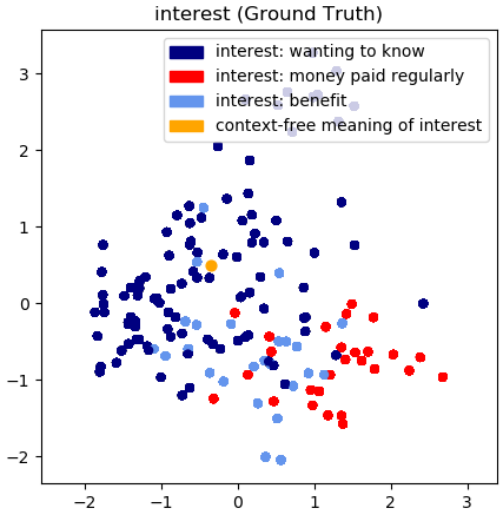
(a) Gaussian Mixture model of "power"



(b) Gaussian Mixture Model of "interest"



(c) Distribution of the modulated meaning of "power"



(d) Distribution of the modulated of "interest"

Figure 3.3: The first two pictures are plotted according to the annotation in the corpus. The second two pictures present the result of the unsupervised Gaussian mixture model with blue contours representing bivariate Gaussian distributions. Each Gaussian distribution represents a sense of the word. The darkness of blue represents the probability density of the Gaussian distribution. So the darker the blue is, the more probable the word of this sense will have this particular modulated meaning.

emerges in both the previous contexts and the target polysemous word meanings modulated by context, we applied permutation tests[19] in which we compared the model's clustering accuracy (compared with the ground truth label derived from the annotated corpus mentioned previously) to a random distribution of labels, allowing us to test the hypothesis that sense clusters consistent with human intuition emerge in our model. We sampled a set of random labels for each polysemous word, and retrieved two sets of labels created by the Gaussian mixture model for the target polysemous word and previous context. Then we calculated the accuracy of these two pairs of labels against the ground truth annotations in the corpus using V-measure (Rosenberg and Hirschberg, 2007)[20] We repeated this process 5000 times for item and performed a Welch's t-test to compare these accuracy distributions of the clustered labels and the random labels. The accuracy scores of both the clustering of polysemous words and previous contexts are all significantly above chance ($p < 0.001$). From the result in Table 3.1, we can see that polysemous words with senses dissimilar to each other had higher accuracy in clustering, while lower accuracy is among words with senses very similar to each other.

[19]In permutation tests, we simply run the clustering algorithms n times to acquire the distribution of clustering accuracy so that we can test whether the obtained result is a chance or reflects some real pattern.

[20]The upper bound of V-measure is 1, suggesting a perfect matching with the ground truth clustering labeling, and 0 is as bad as it can be.

Words	n(Senses)	n(Tokens)	Mean(Random)	Mean(Context)	t	p	Mean(Target)	t	p
right	2	54	0.01(0.01)	0.54(0.26)	-138.33596	$< 0.01*$	0.57(0.23)	-165.7140002	$< 0.01*$
way	2	306	0.00(0.00)	0.29(0.05)	-399.2738453	$< 0.01*$	0.20(0.05)	-264.0938988	$< 0.01*$
man	2	256	0.00(0.00)	0.47(0.07)	-418.2969306	$< 0.01*$	0.44(0.04)	-691.0091362	$< 0.01*$
life	3	150	0.01(0.00)	0.12(0.03)	-194.5065742	$< 0.01*$	0.09(0.03)	-172.4208147	$< 0.01*$
problem	2	24	0.03(0.04)	0.13(0.12)	-58.4666684	$< 0.01*$	0.14(0.11)	-66.22472595	$< 0.01*$
case	3	24	0.09(0.06)	0.29(0.11)	-109.1536451	$< 0.01*$	0.29(0.09)	-123.4706991	$< 0.01*$
school	2	24	0.03(0.04)	0.12(0.13)	-44.63753452	$< 0.01*$	0.09(0.11)	-36.56969444	$< 0.01*$
head	3	36	0.05(0.04)	0.71(0.10)	-410.0651209	$< 0.01*$	0.71(0.09)	-448.5218042	$< 0.01*$
country	2	34	0.02(0.03)	0.13(0.13)	-55.34390305	$< 0.01*$	0.13(0.13)	-58.62186267	$< 0.01*$
company	3	33	0.06(0.04)	0.30(0.13)	-115.3920944	$< 0.01*$	0.31(0.13)	-127.4535993	$< 0.01*$
room	2	42	0.01(0.02)	0.47(0.26)	-121.2398073	$< 0.01*$	0.43(0.26)	-109.4820975	$< 0.01*$
figure	3	30	0.06(0.05)	0.36(0.13)	-145.6454068	$< 0.01*$	0.42(0.13)	-178.7518013	$< 0.01*$
power	3	42	0.04(0.03)	0.31(0.13)	-139.2664879	$< 0.01*$	0.38(0.11)	-205.3802755	$< 0.01*$
interest	3	75	0.02(0.01)	0.33(0.11)	-181.6764711	$< 0.01*$	0.27(0.09)	-182.2931466	$< 0.01*$
building	2	42	0.01(0.02)	0.24(0.21)	-74.45572442	$< 0.01*$	0.29(0.23)	-81.24912504	$< 0.01*$
term	2	68	0.01(0.01)	0.19(0.14)	-89.84207278	$< 0.01*$	0.14(0.11)	-72.21249896	$< 0.01*$
film	2	18	0.04(0.06)	0.26(0.16)	-89.88498736	$< 0.01*$	0.27(0.13)	-86.4059051	$< 0.01*$
date	2	18	0.04(0.06)	0.24(0.20)	-64.81268349	$< 0.01*$	0.27(0.22)	-74.69170083	$< 0.01*$
issue	2	28	0.02(0.03)	0.13(0.12)	-59.33096106	$< 0.01*$	0.14(0.16)	-58.1165849	$< 0.01*$
matter	3	33	0.06(0.04)	0.43(0.18)	-141.7609424	$< 0.01*$	0.39(0.16)	-131.9548785	$< 0.01*$
back	2	78	0.00(0.01)	0.22(0.16)	-94.4371559	$< 0.01*$	0.21(0.1)	-104.0218819	$< 0.01*$
world	2	194	0.00(0.00)	0.03(0.04)	-45.91502646	$< 0.01*$	0.02(0.00)	-49.7401746	$< 0.01*$
point	4	28	0.14(0.06)	0.39(0.10)	-149.2831587	$< 0.01*$	0.44(0.19)	-166.9493663	$< 0.01*$
system	2	182	0.00(0.00)	0.12(0.08)	-107.9401929	$< 0.01*$	0.12(0.00)	-209.4022598	$< 0.01*$
hand	2	26	0.03(0.04)	0.10(0.08)	-52.29789509	$< 0.01*$	0.09(0.03)	-47.85721836	$< 0.01*$
form	3	87	0.02(0.01)	0.32(0.11)	-175.9811898	$< 0.01*$	0.34(0.08)	-226.2852028	$< 0.01*$
office	2	42	0.01(0.02)	0.11(0.11)	-62.64026067	$< 0.01*$	0.10(0.05)	-59.55746128	$< 0.01*$
position	3	42	0.04(0.03)	0.26(0.10)	-134.381395	$< 0.01*$	0.27(0.18)	-132.2234076	$< 0.01*$

Table 3.1: Clusters of Polysemous senses. Standard deviation of all means is shown in parentheses.

Both the meaning modulations of the target polysemous words and contexts before the target words are significantly similar to the ground truth human intuition of sense individuations. This means that meaningful clustering emerges in both of them. However, this does not mean that both of them are polysemous senses of words. According to this new theory, contexts modulate the target word meaning into different contextual meanings. It is the modulating contexts that make the modulated meanings different, but it is the target word being modulated that pulls these contextual meanings together to be relevant as the polysemous senses. In sum, clustering emerges in both the previous contexts of polysemous words and the modulated target word meanings, but it is the modulated meanings that are the polysemous senses.

In order to distinguish the roles of previous contexts and modulated word meanings, we measured the similarity between the modulated meanings of target words and context-free meaning of the target words and also the similarity between previous contexts of target words and context-free meaning of the target words. The similarity was computed as the euclidean

distance between these two vectors. The context-free meanings of a word were extracted from the recurrent layers when nothing but the target words were input into the previously trained model. Hence, there were no previous contexts modulating the target word meanings. The modulated meanings of target words and the previous contexts of target words were extracted in the same way as the recurrent layers when and before the target words are input. The result was shown in Table 3.2. The distances between modulated meaning and context-free meaning ($M = 4.49, SD = 0.09$) were significantly smaller than the distances between the previous context and context-free meaning ($M = 6.49, SD = 0.05), t(50) = -27.9, p < .01$. This suggests that modulated meanings in context are more suitable than the previous contexts as the polysemous senses because different polysemous senses of a word have to be the sense of and relevant to this word. As a result, both the contexts before the target word and the target word itself work together to create the phenomena of polysemy. The modulating contexts initiate clustering. The modulated target word ensures its different modulations relevant as the polysemous senses.

3.4.2 Modeling Regular Polysemy as Evidenced by Geometric Patterns between Clusters

Next, we tested whether our model captured the pattern of regular polysemy. Regular polysemy is a subcategory of polysemous words sharing patterns with other polysemous words so that their senses follow a similar sense pattern. A formal definition of regular polysemy was first given by Apresjan (1974).

> "Polysemy of the word A with the meanings a_i and a_j is called regular if, in the given language, there exists at least one other word B with the meanings b_i and b_j, which are semantically distinguished from each other in exactly the same way as a_i and a_j and if a_i and b_i, a_j and b_j are nonsynonymous."

Regular polysemy is exemplified by words like "bottle" and "can." Both of them have the

Word	Distance between context-free meaning and modulated meaning	Distance between context-free meaning and context
right	5.055466068202052	7.150261994065909
way	4.687675223293074	6.687803947987121
man	4.307380996009176	6.321435841109402
life	4.325067264539701	6.5610595586823255
problem	4.008743804867796	6.38097418900813
case	4.346037136441539	6.409790044807526
school	4.605996442128377	6.301127676982956
head	4.891251127642414	6.746920891106129
country	4.603889402649212	6.428128490985279
company	4.444797497077529	6.129696076114972
room	4.295097446245421	6.195505795655428
figure	4.751714542388916	6.649252591067797
power	4.683839912623005	6.8758788278116985
interest	4.211841853041398	6.617945945169043
building	4.370523995161056	6.265177508195241
term	5.088193865105657	6.684964482146914
film	4.611371386158574	6.297269761025369
date	4.037615131159297	6.311933908306185
issue	4.246251006336773	6.43766249628628
matter	4.077568163610485	6.377540948256007
back	4.889421946124027	6.543625492798655
world	4.786468000577484	6.711315628763393
point	4.6397463893427435	6.626270512530678
system	4.254914183312274	6.40063029147209
hand	4.559083682903345	6.662872217703557
form	4.584657266616821	6.278816229418704
office	4.2958116324650755	6.451713302921008
position	4.0207778228226525	6.288376815983506

Table 3.2: Similarity between context-free meaning and modulated meaning of the target word and similarity between context-free meaning and previous context of the target word. Measured by Euclidean Distance in the semantic space.

container and the content sense as in (1)/(2) and (3)/(4). And these two senses have similar uses in "bottle" and "can."

(1) This bottle tastes better.

(2) This bottle is fragile.

(3) I ate a whole can.

(4) I threw out an empty can.

Traditionally, people explain regular polysemy by positing some kinds of explicit structures into their semantic representations as in the *static* accounts. These structures are accordingly shared by polysemous words in the same category, which explains the regularity of such patterns. However, regular polysemy poses a problem for most *operational* accounts of polysemy because these accounts do not posit explicit structures in the semantic representation, and they see polysemy as a phenomenon of semantic/pragmatic operation. Therefore, this requires an extra explanation of why these meaning modulations are regular and shared by other polysemous words in the same category.

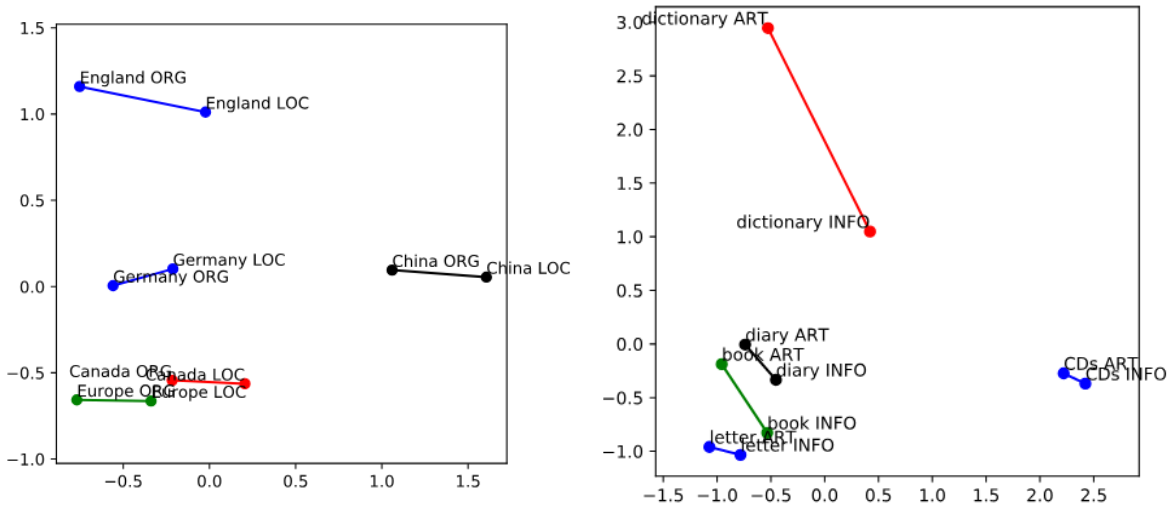
We argue that the structure of regular polysemy also emerges from the processing and production of polysemous words instead of being directly and statically encoded in their semantic representations. In our model of polysemy, the activation patterns in the embedding layer is a weighted superimposition of different contexutal modulated activation in the recurrent layer. And the connection weights between the embedding layer and the hidden layer store the probabilistic knowledge of how polysemous words are modulated in different contexts. Some patterns of meaning modulations occur more frequently so they are more entrenched in the connection weights (long term memory) to facilitate future processing. As a result, meaning modulations follow these regular patterns through entrenchment and the regularity of these patterns can be assessed by analyzing the geometric relationships between modulated meaning in the semantic space, which is similar to the word analogy relationship discovered in

natural language processing research (Mikolov et al., 2013b). For example, the regularity of CONTENT-CONTAINER polysemy can be captured by the vector from $\text{bottle}_{\text{container}}$ modulations to $\text{bottle}_{\text{content}}$ modulations being parallel to the vector from $\text{can}_{\text{container}}$ modulations to $\text{can}_{\text{content}}$ modulations.

In order to test our LSTM model, we used an annotated corpus of regular polysemy (Alonso, 2013) that includes five categories of regular polysemy: ANIMAL-MEAT e.g., "chicken" as in "raise a chicken" and "eat chicken," ARTICLE-INFORMATION e.g., "book" as in "buy a book" and "revise a book," CONTAINER-CONTENT e.g., "glass" as in "break a glass" and "finish a glass," LOCATION-ORGANIZATION such as "Canada" in "went to Canada" and "Canada negotiated a treaty," PROCESS-RESULT such as "construction" in "finish the construction" and "a rigid construction." Each category includes 500 sentences with regular polysemous words annotated as, for example, ARTICLE or INFORMATION sense. We fed this regular polysemy corpus as sequences of words into the LSTM model pre-trained on the Wikitext-103 corpus. As such, the LSTM model had not seen this regular polysemy corpus during the training and was instead asked to generalize what it had learned to these novel sentences. We extracted the hidden activity patterns of each annotated polysemous word as a 600-dimensional vector, which represents the modulated meaning of this word token in its context. Then we averaged the vectors of the same sense to be the word's sense vector and defined the sense pattern of a polysemous word to be the vector from its one averaged sense to the other averaged sense.

Ideally, polysemous words belonging to the same category have parallel sense patterns with each other as in Fig. 3.4, which are two successful cases captured by our model. In Fig. 3.4, each dot is a 2d projection (2 highest-ranked dimensions by PCA) of a 600d vector averaged from meaning modulations of one sense. The line segments connecting these dots represent the sense patterns of different regular polysemous words. We see similar sense patterns among regular polysemy in both LOC-ORG and ART-INFO in terms of their angles. So $\vec{V}_{ChinaLOC} - \vec{V}_{ChinaORG} \approx \vec{V}_{EnglandLOC} - \vec{V}_{EnglandORG} \approx \vec{V}_{GermanyLOC} - \vec{V}_{GermanyORG} \approx \vec{V}_{CanadaLOC} - \vec{V}_{CanadaORG}$ and

similar relationship holds for ARTICLE-INFORMATION regular polysemy.



(a) Regular Polysemy of LOCATION-ORGANIZATION

(b) Regular Polysemy of ARTICLE-INFORMATION

Figure 3.4: Regular polysemy modeled as geometric relations between clusters. Each dot represents the mean of a word sense's meaning cluster. Each line denotes a two-dimensional simplification of the multidimensional relationship between the two senses of each word. Similar-direction lines suggest a common type of relatedness among different word pairs, reflecting sense regularity. We quantified this regularity by computing the angle of each pair of word senses in high-dimensional space, and then computed the degree of variability in this angle, where highly regular patterns should be reflected in very low degrees of variability among sense pairs.

For hypothesis testing, we calculated the sense pattern of each polysemous word in the corpus. For the sense pattern of each word, we calculated the angle between itself and that of another word in the same category (for example, the angle between the sense pattern of "England" and the sense pattern of "Germany") on the one hand, and the angle between itself and that of a word out of the same category (for example, the angle between the sense pattern of "England" and the sense pattern of "line") on the other. This process was repeated for 15000 times as a permutation test. If the angles are significantly smaller within the same category than out of the same category, it serves as evidence that our model captures polysemy regularity.

Category	Within Category Means	Out of Category Means	t	p
ANIMAL&MEAT	86.49(10.85)	89.64(4.20)	-33.17	$< 0.01*$
CONTAINER&CONTENT	93.32(13.11)	90.22(5.01)	27.06	> 0.99
LOCATION&ORGANIZATION	77.06(14.60)	90.11(3.96)	-105.59	$< 0.01*$
ARTICLE&INFORMATION	68.87(9.48)	90.17(4.81)	-245.33	$< 0.01*$
PROCESS&RESULT	90.56(12.14)	89.93(4.08)	6.01	> 0.99

Table 3.3: Result of permutation test of regular polysemy

Among the five categories of regular polysemy, three had significantly smaller within-category angles than out-of-category angles, which shows that regular polysemous words within these categories share similar patterns. This serves as preliminary evidence that regular sense patterns emerge from the linguistic processing as the geometric relationship between clusters of modulated meanings in context, here as the degree of the angle. It shows that our connectionist model captures rule-like patterns of regular polysemous words without explicit supervision.

The other two categories don't show significant similar patterns. We propose that this reflects the fact that our model only captures the context *before* the polysemy word, hence only W_1 to W_{t-1} but not W_{t+1} (and the following), so the modulation of the polysemous words has to be reached in the previous contexts in our model. This would predict that certain types of polysemy regularity are captured during online processing in which contextual cues regularly occur prior to word presentation (e.g., LOCATION-ORGANIZATION), whereas other types of regularities reflect offline judgments that are largely meta-linguistic and thus not well captured within a processing-based model like the present one. It could also be that the regularity of polysemy is much more complicated than parallel relationships between sense patterns. We suggest that further research should be done.

3.5 Model Assessment with Behavioral Experiments

In this section, we assessed our trained LSTM model with human behavioral data — offline ratings, eye-tracking data, and reaction time of priming tasks. We tested how well our trained model captured the sense relatedness and the behavioral characteristics of polysemy.

3.5.1 Modeling Sense Relatedness Ratings

We tested our model in terms of how well it captures previous behavioral findings. We first calculated polysemous senses similarity from our previously trained RNN model and compared it with Klepousniotou et al. (2008)'s empirical sense rating data.

In Klepousniotou et al. (2008)'s study, they selected 72 polysemous words as their stimuli. Each word was accompanied by four modifiers divided into two senses. For example, "admission" was accompanied by "movie," "concert," "guilty," and "false." The first two modifier-word pairs suggest the sense of the entering process, while the last two suggest the sense of acknowledging truth. They collected the empirical semantic overlap ratings between the two senses suggested by the modifiers. This provided a rating of 1 to 5 (1 = low overlap, 5 = high overlap). Words like "admission" in "movie admission" or "guilty admission" were rated low since these two senses of "admission" have low overlap and relatedness. High overlap words included words like "book" as in "best-selling book" or "heavy book."

For the simulation, we preprocessed Klepousniotou et al. (2008)'s data by deleting one polysemous word "key" because one of its modifiers "backspace" does not exist in our model training data. Hyphenated words like 'best-selling' were coded as two words (e.g., "best-selling" was coded as "best selling"), reflecting the format of our training data. Then we input all the polysemous words preceded by each of their four modifiers into our trained RNN model and collected the vector of hidden layer activation when the polysemous word was input. For each polysemous word, we collected four vectors of two different senses. We then averaged the two vectors belonging to the same sense and calculated the cosine similarity between the two average vectors.

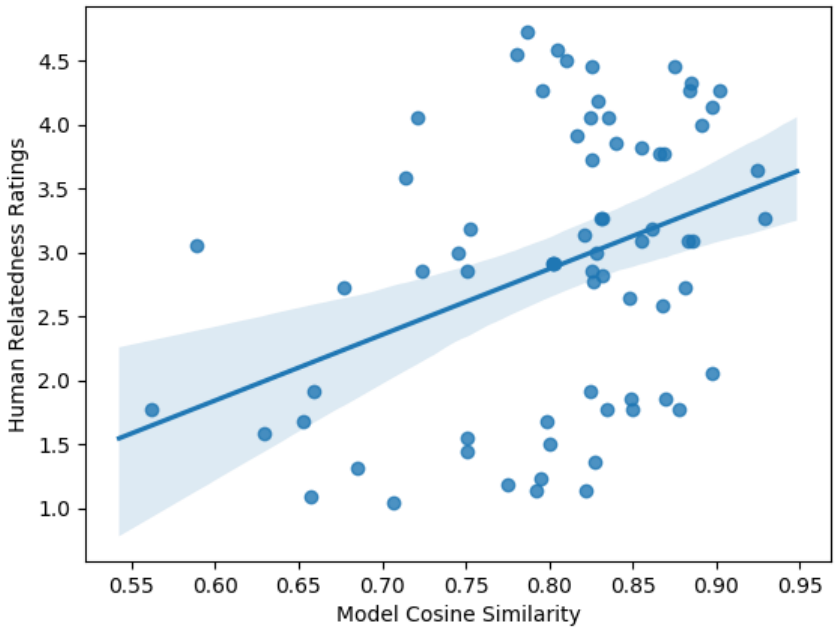


Figure 3.5: Scatterplot of the correlation between model produced cosine similarity and human rated relatedness of polysemous words

We calculated the correlation between the model generated cosine similarity and the human relatedness ratings. There was a positive correlation between these two variables, $r = 0.37$, $p \leqslant 0.001$, n = 71. A scatterplot summarizes the result (Fig. 3.5). As a result, our model captured the polysemous sense relatedness from learning a natural corpus of English texts to a moderate degree.

We are aware that this is not a very strong effect. There could be multiple reasons. First, there are only two contexts for each sense, and the context modulating the polysemous target word in Klepousniotou et al. (2008)'s dataset is only one word long. It may not be enough for the model to reach a complete modulation as the polysemous sense that human raters are thinking of. The correlation could be improved if we use longer contexts or more contexts to modulate the target polysemous word. Second, our model was trained on one set of experiences (Wikitext-103) that resembled some aspects of how a language learner encounters English but clearly not everything. It is reasonable to expect the model will approximate knowledge, but

closer approximations could be captured if its experiences more closely matched the variety of utterances/sentences a person experiences over their lifetime. Third, there is no referential or multimodal information in our training data. Distributional information only explains one aspect of linguistic meaning.

3.5.2 Modeling Online Sentence Processing of Polysemy

Next, we tested how well our model captured the behavioral findings of previous sentence processing tasks of polysemy. There is a large literature in computational linguistics and psycholinguistics discussing the relationship between reading times and word probability given previous contexts (Hale, 2001; Ferreira and Henderson, 1990; Altmann et al., 1992). It is known that if a word is predicted to be more likely in context, it is read faster in human sentence processing. Here, we simulated the sentence processing of polysemy in Foraker and Murphy (2012)'s self-paced reading and eye-tracking study based on the log probability produced by our model's prediction.

Foraker and Murphy investigated the reading time of disambiguating phrases of polysemous words in both biasing contexts and neutral contexts. In neutral contexts, participants read sentences like "they discussed the cotton after the fabric ripped/crop failed a second time," where the meaning of "cotton" was neutral between the crop and fabric until the disambiguating phrase "fabric ripped/crop failed" appear. In biasing contexts, participants read sentences like "the fashion designers/the farm owners discussed the cotton after the fabric ripped/crop failed a second time." "Fashion designers" biased the interpretation of "cotton" towards the fabric sense, while "farm owners" biased it towards the crop sense. The disambiguating phrase "fabric ripped/crop failed" might either be consistent or inconsistent with the biasing context.

We used the same stimuli in Foraker and Murphy (2012)'s Experiment 2 and 3. They consisted of 25 items, and each item had three contexts — neutral, dominant, and subordinate and two disambiguating phrases — dominant and subordinate, so there was a total of six conditions for each item (see Table 3.4). The main goal of the simulation was to test whether our model

captured the same human behavioral effects: that disambiguating phrases were read faster in consistent conditions (dominant context and dominant disambiguating phrase or subordinate context and subordinate disambiguating phrase). And dominant disambiguating phrases were read slightly faster compared to subordinate disambiguating phrases in neutral contexts.

Context	Sense Completion	Sentence
Dominant	Dominant	The fashion designers discussed the cotton after the fabric ripped a second time.
Dominant	Subordinate	The fashion designers discussed the cotton after the crop failed a second time.
Subordinate	Dominant	The farm owners discussed the cotton after the fabric ripped a second time.
Subordinate	Subordinate	The farm owners discussed the cotton after the crop failed a second time.
Neutral	Dominant	They discussed the cotton after the fabric ripped a second time.
Neutral	Subordinate	They discussed the cotton after the crop failed a second time.

Table 3.4: Materials of sentence processing in Foraker and Murphy's study

We input each item into the previously trained RNN model as six sentences in six different conditions and applied a log-softmax function to the output layer activation before the target disambiguating phrase was input in order to extract the conditional log probability of the target phrase being predicted given the previous contexts. If the disambiguating phrase had more than one word, we summed the conditional log probabilities of each disambiguating word as the log probability of the disambiguating phrase.

Context	Disambiguating Target	Log Probability
Neutral	Dominant	-12.25(5.05)
Neutral	Subordinate	-12.69(4.56)
Subordinate	Dominant	-12.28(4.82)
Subordinate	Subordinate	-12.28(4.41)
Dominant	Dominant	-11.93(5.05)
Dominant	Subordinate	-12.76(4.59)

Table 3.5: Result of sentence processing simulation. Standard deviation of all means is shown in parentheses.

The summary of the simulation result is shown in Table 3.5. We applied a repeated-measures 3×2 ANOVA to analyze the result (Table 3.6). Mauchly's test indicated that the assumption of sphericity had been violated for the main effects of Context, $W = 0.769$, $p < .05$, and Context-Target Interaction, $W = 0.650$, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse–Geisser estimates of sphericity. There was a significant main context

effect, $F(1.6, 38.4) = 3.55$, $p < 0.05$, and a significant interaction effect between the type of context and the type of disambiguating target, $F(1.36, 32.64) = 11.38$, $p < 0.001$, which was similar in Foraker and Murphy (2012)'s study.

Effect	DFn	DFd	F	p
Context	2	48	3.56	0.04∗
Target	1	24	0.25	0.63
Context:Target	2	48	11.32	< 0.001∗

Table 3.6: Result of repeated-measures ANOVA of sentence processing simulation

Tukey *post hoc* tests (Table 3.7) revealed that disambiguating phrases of dominant senses were more probable following dominant contexts than subordinate contexts, $t(92) = 3.124$, $p < 0.05$. Disambiguating phrases of subordinate senses were less probable following dominant contexts than subordinate contexts, $t(92) = -4.268$, $p < 0.001$. However, we did not find the significant difference between the dominant and subordinate disambiguating phrases in neutral contexts, $t(24.7) = 0.516$, $p = 0.9950$.

Contrast Conditions	Estimate	SE	DF	t	p
DominantDominant - NeutralDominant	0.32199	0.115	92.0	2.811	0.0646
DominantDominant - SubordinateDominant	0.35777	0.115	92.0	3.124	0.0280*
DominantDominant - DominantSubordinate	0.84090	0.862	24.7	0.976	0.9214
DominantDominant - NeutralSubordinate	0.76643	0.860	24.5	0.891	0.9451
DominantDominant - SubordinateSubordinate	0.35208	0.860	24.5	0.409	0.9983
NeutralDominant - SubordinateDominant	0.03578	0.115	92.0	0.312	0.9996
NeutralDominant - DominantSubordinate	0.51891	0.860	24.5	0.603	0.9898
NeutralDominant - NeutralSubordinate	0.44444	0.862	24.7	0.516	0.9950
NeutralDominant - SubordinateSubordinate	0.03009	0.860	24.5	0.035	1.0000
SubordinateDominant - DominantSubordinate	0.48313	0.860	24.5	0.562	0.9926
SubordinateDominant - NeutralSubordinate	0.40866	0.860	24.5	0.475	0.9966
SubordinateDominant - SubordinateSubordinate	-0.00569	0.862	24.7	-0.007	1.0000
DominantSubordinate - NeutralSubordinate	-0.07447	0.115	92.0	-0.650	0.9867
DominantSubordinate - SubordinateSubordinate	-0.48882	0.115	92.0	-4.268	0.0007*
NeutralSubordinate - SubordinateSubordinate	-0.41435	0.115	92.0	-3.618	0.0063*

Table 3.7: Result of *post hoc* analysis of sentence processing simulation

In sum, we found the consistent effects that polysemous words are processed faster in consistent contexts than inconsistent context. We did not find the dominance effect that following

neutral contexts, polysemous words in dominant senses are read faster than subordinate senses. However, this dominance effect in neutral contexts has been a controversial effect. Frisson (2015) replicated Foraker and Murphy (2012)'s sentence processing experiments but failed to find the significant difference in reading times between dominant and subordinate sense following neutral context as us. Therefore, our model failing to replicate the dominance effect could be due to that this dominance effect is itself small. Moreover, a danger of computational simulation is overfitting a single set of findings which may themselves not be good estimates of population behavior, so it is better that our model successfully simulated the more robust finding — consistency effect rather than the less robust dominance effect.

3.6 Comparison with Other Models of Polysemy

In this section, we discuss our model of polysemy in comparison with other related models. We first compare our model with other connectionist models of polysemy (Rodd et al., 2004; Armstrong and Plaut, 2008, 2016; Rodd, 2020), then discuss the connection of our model with an exemplar model of language processing and acquisition (Jamieson et al., 2018; Ambridge, 2019).

Rodd et al. (2004)'s and Armstrong and Plaut (2008, 2016)'s models are attractor networks. Attractors networks are highly interconnected networks, whose activity will settle into a stable state after a period of time. Basically, their models have a layer of orthography representing the word form of polysemous words and another layer of semantics representing the senses of polysemy. These two layers are connected so that a word form can activate its word meaning. Attractor network modelers train their models until the semantic layer learns to settle into the correct senses. Thus, their models present how semantic activation of polysemous words unfolds in the temporal dimension.

There are a lot of similarities in spirit between attractor models of polysemy and our model. We all recognize the importance of context and how it modulates the activation of stored poly-

semous word meaning in context. And the entrenchment of clustering very likely form attractors in the activation of polysemous word meanings in the recurrent layer to constrain possible polysemous word meanings.

However, we do want to point out some difference between our model and attractor network models of polysemy that is theoretically meaningful. First, our model represents contexts differently. Our model represents contexts as a sequence of words before the target polysemous word so it can represent a wide range of different contexts. In contrast, attractor network models represent contexts in a discrete way as context units in 3.8 representing either context biasing one sense (01) or another sense (10).

Second, our model represents polysemous senses differently. In an attractor network model, a polysemous word form is directly mapped on to multiple discrete sense representations as in Table 3.8. The two senses are represented as two binary vectors overlapping in some dimensions, which corresponds to the relatedness among senses. This creates a distinct boundary between different senses of one word. By contrast, in our model, these quasi-discrete senses emerge as the clusters of contextual modulations of the activated word meaning, instead of being statically encoded as the word meaning or being represented as any particular activation in the model. To put more specifically, each instance of a word is represented as a discrete pattern of activation across many units. However, this exact pattern of activation varies somewhat from one instance to another because of the difference in the modulating contexts. Differences in word senses occur as a result of these individual senses forming quasi-regular clusters in activation space. The distance and strength of the boundary between each of these clusters reflect the relative distinctness of these competing word senses. These quasi-discrete senses can be quantified as Gaussian clusters by unsupervised Gaussian mixture modeling as in the result section. So in one sense, our model can be seen as a more detailed implementation of Rodd et al. (2004)'s and Armstrong and Plaut (2008, 2016)'s models, in which the discrete sense vectors in Table 3.8 from their models could be seen as a simplification of our model of polysemy and describing higher-level emergent characteristics of polysemous senses. Cru-

Item type	Orthography	Context	Semantics
Polysemous	0010	10	00000000111100000000
	0010	01	00000000111100000000

Table 3.8: Discrete polysemy representations in attractor networks. Representations are coded using binary units of 1 or 0. Polysemy is coded as an identical orthographic code (0010) mapping onto two different but overlapping semantic codes, differentiated via a set of context features that condition the recognition of one or another. Adapted from Armstrong and Plaut (2016).

cially though, there is no absolute distinction between different senses of a polysemous word — the distinction emerges only as a result of differences in terms of how words are modulated in context, where the contextual modulations of words vary in the extent to which they form quasi-discrete clusters in multidimensional distributional space. Third, this difference of representation leads to the potential difference between the attractors in attractor network models and our LSTM model. The attractors in attractors network models are point attractors so a new activation is attracted to one of many discrete point attractors depending on the context. However, in our LSTM model, the attractors are not points but cloud-like clusters. A particular activation of a polysemous sense in context can be any point in the cloud depending on the nuances in contextual difference. As a result, our model could represent the nuance between the same senses of a word in a slightly different context.

Third, our model explains polysemy from the perspective of emergentism — how quasi-discrete senses emerge from contextually-driven meaning modulation. Polysemous senses are not statically encoded in the lexical entry but learned without supervision from a large English corpus[21] instead of artificially hand-coding them as in Rodd et al. (2004)'s and Armstrong and Plaut (2008, 2016)'s model.

In terms of the different psycholinguistic predictions between these previous models and ours, their models focus on the temporal dimensions of semantic activation i.e., the tendency for words to vary in their speed of recognition as a result of activity settling into a stable activation pattern. Polysemous words have a different profile of semantic activation from homonyms

[21]This corpus does not represent the current status of English fully but it provides a snapshot of an important style of written English and also includes a wide range of topics.

and monosemous words, which is called polysemy advantage. So their models capture this polysemy advantage of semantic activation (Rodd et al., 2002; Klepousniotou and Baum, 2007) in the single word lexical decision task. Instead, our model processes polysemous words within the context of full sentences, so it can yield predictions of human polysemy processing in the context of sentences (Foraker and Murphy, 2012) as shown in the section of simulating polysemy processing. For example, sentence contexts can constrain recognition in a way that can enhance or attenuate ambiguity effects (McRae et al., 1998). Furthermore, our model could indirectly predict the priming effect for polysemous phrases in the sensicality judgment task (Klein and Murphy, 2001; Klepousniotou et al., 2008) based on the sense similarity produced as shown in the section of simulating sense relatedness ratings.

Continuous with comparisons, we want to discuss the relationship between our model of polysemy with Jamieson et al. (2018); Ambridge (2019)'s exemplar model of language acquisition and processing. Jamieson and Ambridge argue that listeners do not store abstract linguistic representations such as abstract type meanings or senses. Instead, they maintain representations of every exposed exemplar, and thus process and produce language on the fly by analogy with stored exemplars.

Our theory shares some commonalities with Jamieson's and Ambridge's exemplar models. We both recognize the difficulty of drawing a clear-cut boundary between different senses to delineate an appropriate abstract sense representation, as Ambridge points out, "it is not possible to posit abstractions that delineate possible and impossible form; ... that ... rule in pool *tables* and data *tables*, but rule out chairs." We point this out in the section that summarizes previous theories, and we both recognized the importance of exemplars in language processing, as we argue that abstract and discrete senses are metalinguistic clusters of meaning modulations of word exemplars. Each instance of a polysemous word is processed individually and hence modulated by its own context. Polysemous senses as abstract and discrete entities are not directly stored or represented in our linguistic processing. Instead, they are metalinguistic features from different modulations of individual words.

However, our approach is not as radical as the one put forward by Jamieson et al. (2018), which rejects any abstraction in semantic theorizing. We still reserve an abstract representation for each word as the superimposition of all its different meanings, which serves as the base for meaning modulation in context, so that constraints for possible modulations of this word can be learned, stored, and exercised on polysemy processing in the future. That is, recognizing a new word in context is not the process of matching it against all prior experiences. Instead, its computation is achieved through weighted connections that are established/abstracted through these prior experiences. As a result, we recognize the importance of exemplars, particularly in polysemy representation and processing, but we still retain a minimal amount of abstraction for polysemy.

3.7 Conclusion

This computational modeling of polysemy implements the two essential causal components of the new theory of polysemy: meaning modulation in context and entrenched clustering. The meaning modulation by context is realized by the recurrent layer, whose activation is modulated by the previous context. The clustering effect is obtained by statistical analysis on the internal representations extracted from the recurrent layer representing the same polysemous words in different contexts. These causal components provide a causal account of how polysemy emerges from linguistic communication and processing. They also constitute what polysemy is as the new theory of polysemy proposed in Chapter 2 argues that a word is polysemous if and only if the meaning modulation of this word has entrenched clusters.

The clusters obtained in the model are examined against our intuitive individuation of polysemous senses through a sense-annotated corpus. Furthermore, the geometric relations among these clusters are examined with another sense-annotated corpus of regularly polysemous words. Both analyses suggest that abstract internal representations within this model nevertheless match the polysemous features of graded, flexible, but structured sense individu-

ation. Furthermore, we simulated an offline polysemy relatedness rating study and an online sentence processing study and assessed how well our model captured these behavioral data. These analyses provide positive evidence that the new theory of polysemy proposed in Chapter 2 captures important features of polysemy and polysemy processing.

In terms of the limitation of our modeling, it is clear that humans use a richer and more grounded set of information about semantics derived from sensory and referential information (e.g., sensory and functional features, (Cree et al., 1999)) than our model. Our model does not preclude this perspective, and this richer set of information could be easily incorporated into the model, for instance, by encoding word representations using feature information. Although this may serve to improve the realism of the model, we are struck by the degree to which our simulation captures the target phenomena using only distributional information about word co-occurrence.

As a final remark, our model belongs to the category of connectionist or neural network models of semantics. However, we do not intend our models to be a competition to symbolic theories to semantics. We think connectionist models capture the same semantic knowledge as symbolic models but in different size of grain. This difference results from the fact that connectionist models learn the rule-like knowledge from the linguistic data directly instead of the theorizing from human experts. As shown in the experiments of modeling regular polysemy, our model learned and captured the rule-like regular patterns of polysemy, such as LOCATION-ORGANIZATION of country names. This result supports how we view our model of polysemy as a connectionist model.

Chapter 4

Philosophical Implications of Polysemy

4.1 Introduction

In this chapter, I discuss the philosophical implications of polysemy and my theory of polysemy proposed in the previous chapters. I first situate polysemy within the debate between semantic minimalism and contextualism in philosophy of language and focus on what polysemy contributes to the debate. I argue that polysemy poses a serious problem for semantic minimalism because it only allows a very limited range of context-sensitivity in the literal truth conditions of an utterance, with which polysemy does not fit well. On the other hand, my theory of polysemy, which provides a solution to this problem, fits well within contextualism.

4.2 Background for the Debate between Minimalism and Contextualism

There is no doubt that contexts contribute a lot to what is meant in an utterance, but my question is how much contextual contribution belongs to the literal meaning of an utterance. Hence, the debate between minimalism and contextualism is focused on the context-sensitivity of the literal truth condition of an utterance. Grice (1991) first distinguishes what is said and what

is implicated by an utterance. While an utterance can convey an open-ended possibility of information through what is implicated, Grice claims that what is said by an utterance is tightly constrained and closely related to the conventional meaning of the expression uttered so that what is said can be separated from what is implicated in context. This idea that a minimal truth-condition of an utterance can be isolated from other highly contextual meanings of this utterance is further developed by semantic minimalists such as Cappelen and Lepore (2008), Borg (2012), and Devitt (2013). They claim that contexts only contribute what is called the minimal proposition (or minimal truth condition). Hence,

> (1) a truth-evaluative minimal proposition can be recovered from an utterance as soon as the values of grammatically triggered context-sensitive expressions are assigned, and ambiguous words are disambiguated.

These grammatically triggered context-sensitive expressions include indexical, such as "I" and "now," demonstratives, such as "this" and "those," and other similar constructions, and they form the set called the Basic Set of Context-Sensitive Expressions (Cappelen and Lepore, 2008). Semantic minimalists do not deny the context-sensitive roles of these grammatically triggered expressions. The truth-conditional contribution of the expressions in the basic set, such as indexicals, to the literal truth condition is clearly determined in the context. They usually adopt a Kaplanian approach (Almog et al., 1989) to restrict the relevant context to a limited set of contextual parameters, such as the speaker, the place, and so on. These parameters are objects instead of subjective intentions of the communicators. The contribution of this narrow context can be understood as the value of a function (character) that takes this narrow context as input. On the other hand, semantic minimalists do not deny that full-blown homonyms such as "bank" are ambiguous so that contexts are needed to disambiguate which of them it refers. However, minimalists deny that any more context-sensitivity than the basic set and disambiguation exists when recovering a truth-conditional minimal proposition from an utterance.

On the other hand, semantic contextualists (Recanati, 2004, 2010; Carston, 2008) argue that it is rarely the case that a literal truth condition can be recovered from a declarative utterance when supplied with contexts even if indexicals and ambiguity are already properly dealt with in context (moderate contextualism). Some even argue that it is never the case that a truth-conditional literal meaning can be recovered from an utterance without knowing the full context, which is everything and anything that is relevant to what a speaker is on about in contrast to Kaplanian narrow context (radical contextualism). Instead, contextualists argue that even the most normal and literal uses of words are context-sensitive in terms of their semantic contribution to the literal truth condition. For example, there are context shifting cases in which a single declarative sentence without indexicals and ambiguities has different truth conditions in different contexts of utterance.

(2) There is no beer in the fridge

Consider a scenario in which (2) is uttered at a party when someone wants to drink something cold, an utterance of (2) means that there is no canned or bottled beer in the fridge. If (2) is uttered when someone is cleaning the inside surface of a fridge, an utterance of (2) probably means that there are no beer stains on the inside surface of the fridge. In these two scenarios, "beer" refers to different things, bottled beer, or beer stains, which are determined in the context. However, there is no overt grammatical trigger for this context-sensitivity. "Beer" is just a regular open-class noun. And in both scenarios, "beer" is used literally rather than figuratively or metaphorically. Contextualists argue that the truth conditional difference in these two scenarios is the result of semantic modulation, which does not require an overt bottom-up grammatical trigger as proposed by semantic minimalists. Similar cases of semantic modulation are pervasive in linguistic communication. As a result, contextualists argue that context-sensitivity of the literal truth condition is way more pervasive than the basic set of indexical and ambiguity. Hence, semantic minimalism is false.

4.3 The Problem of Polysemy for Semantic Minimalism

Polysemous words, as we discussed in the previous chapters, are words with multiple related meanings such as "power" and "book." "Power" can mean either the political power as in "China's increasing power in the global stage" or the electric power as in "the power of the house is off." These related senses need to be specified in the context of the utterance. It is obvious that both "power"s can be used in a literal way that alters the truth condition captured in semantics. As a result, polysemy seems to be an obvious case of context-sensitive semantic contribution to the literal truth condition of an utterance.

There are three strategies for semantic minimalism to deal with polysemy. First, they could argue that the context-sensitivity of polysemous words does not need to be handled at all in terms of recovering the minimal proposition from an utterance. The semantic contribution of polysemous words to the minimal proposition is insensitive to context, while the context-sensitive part belongs to the realm of the conversational implicature or speech act content. Second, minimalists could treat polysemy in the same way as ambiguous expressions, such as homonyms, as they acknowledge that ambiguous expressions need to be disambiguated into the correct interpretation based on the context. Third, minimalists could argue that polysemy is handled in the same way as the expressions in the basic set of context-sensitivity so that polysemous words are treated like indexicals such as "I," and "here."

In the following three sections, I discuss these three strategies in detail and argue that none of them work for minimalism. Therefore, polysemy remains a problem for semantic minimalism.

4.4 Strategy 1: No Need to Handle Polysemy for Minimal Proposition

In this section, I discuss the first option that the context-sensitivity of polysemy does not need to be handled for the sake of recovering the minimal proposition. My objection is that even though it may be plausible to claim that there is a minimal meaning of a polysemous word irrespective of contexts, the minimal proposition containing it will cease to be truth-evaluable. Therefore, minimalists cannot hold their thesis (1), if they accept this option.

Consider the polysemous word "window,"

(3) Tom jumps through the window.

(4) Tom breaks the window.

"Window" can either refer to the window pane as in (4) or the window aperture as in (3), so it is a context-sensitive polysemous word. However, a minimalist can respond that both tokens of "window" make the same semantic contribution to each minimal proposition. "Window" just means window in both (3) and (4). The difference in the interpretation between a window pane and a window aperture lies in non-minimal speech act content of the utterance of them. Minimalists will say that polysemy is just not something that occurs within the minimal proposition of an utterance.

A similar case that semantic minimalists have replied directly to is the context-sensitivity of "tall,"

(5) Mount Everest is tall.

(6) Kobe Bryant is tall.

Kobe Bryant is tall as a human but not tall as a mountain. Contextualists hold that contexts influence the truth conditional contribution of "tall" in (5) and (6). However, minimalists such as Cappelen and Lepore (2004) deny the existence of these kinds of context-sensitivity within

literal truth conditions. They argue that the literal truth condition of (5) is just that Mount Everest is tall and same for (6).

They back their position with context-sensitivity tests such as the conjunction test. In a conjunction test, two different true sentences containing the alleged context-sensitive expression are put into conjunction. If the conjunction is also true and non-bizarre, it shows that this expression is not context-sensitive.

(7) Mount Everest and Kobe Bryant are tall.

That the conjunction (7) built from the true (5) and the true (6) is also true shows that "tall" passes the conjunction test. Therefore, as argued by semantic minimalists, "tall" is not context-sensitive.

This reply could be adapted to argue that "window" is not context-sensitive within the literal truth condition in (3) and (4). And luckily, "window" belongs to a special category of polysemy that passes the conjunction test. If we build a conjunction from (3) and (4), we get

(8) What Tom jumped through and Jack broke yesterday is this window.

In (8), "window" is predicated with two different properties, being jumped through and being broken. According to the traditional interpretation, these two properties belong to two different senses of "window," window aperture and window pane. However, (8) does not lead to hampered interpretation and is still true, which could serve as support for minimalists' strategy that polysemy is not context-sensitive within the minimal proposition.

However, I want to argue that this strategy does not generalize to all polysemy. The "window" example works because "window" belongs to a special category within regular polysemy. For this kind of polysemous word, different senses of it refer to different aspects of a single unified object. Panes and apertures are two aspects with which to conceptualize windows. They are two spatial-temporal components of this metaphysically complex object — window — so that both senses can be derived from this object. However, not all polysemous words are

like this in terms of how different senses are related to each other. The semantic relatedness between senses of different polysemous words is wildly various. For example, the word "line" in

(9) The data can be fitted with a curved line.

(10) The prince is removed from the royal line.

(11) It is unprofessional for an actor to forget his line.

(12) Every company here wants to expand its business line.

All these tokens of "line" are related in certain ways but not spatial-temporally connected as the pane and aperture of a window. Instead, these senses refer to very different entities of very different domains. If semantic minimalists want to continue their strategy of "window," they have to argue that "line" contributes semantically the same to the literal truth conditions of all four utterances, which seems absurd. The problem for the semantic minimalists is to decide what is the minimal truth conditional contribution of "line" to each minimal proposition — the minimal meaning of "line."

Minimalists could opt for an underspecified meaning of "line" which somehow generalizes to all different senses of "line," but this move makes the minimal proposition of sentences containing "line" cease to be truth-conditional, because this underspecified meaning of "line" does not pick out any particular object in the world if this underspecification has to include all four different senses. On the other hand, minimalists can opt for one sense of "line" among all of them. However, this creates another problem. That is, it eliminates certain analyticity or semantic entailment.

Consider the word "square," which is polysemous. "Square" either means a geometric object with four right angles and four equal sides or a big open area surrounded by buildings.

(13) All squares have four sides.

(14) A line is a one-dimensional object.

(13) is analytically true in a context where "square" refers to a geometric object, but not analytically true if "square" refers to open space. Similarly, (14) is analytically true if "line" refers to the mathematical object, but analytically false if it means other senses such as drama line. As a result, if minimalists choose any particular sense as the minimal meaning of a polysemous word, it may eliminate the possibility for a sentence to be analytic based on which sense is deemed as essential.

In sum, if semantic minimalists hold that polysemy is not context-sensitive in terms of its contribution to the minimal proposition, this strategy only works for some words like "window." For most words, minimalists have to either posit some underspecified meanings as the minimal meanings or choose one sense among all the polysemous senses as the minimal meanings. However, none of these options work out. Therefore, I argue that this strategy of denying the context-sensitivity of polysemy in contributing to literal truth condition is not plausible.

4.5 Strategy 2: Handle Polysemy as Ambiguity

The second strategy for semantic minimalists to handle polysemy is to simply treat polysemy as ambiguity. Everyone agrees that contexts are needed for disambiguation. For example, homonyms, which are different words with the same spelling or pronunciation, need to be disambiguated according to the context. Whether "bat" is disambiguated into the animal bat or the baseball bat depends on the context, both the linguistic context and broader context of the utterance. Therefore, minimalists could argue that polysemy is just another case of ambiguity so that sentences containing so-called polysemy are first disambiguated into the correct senses before a definite truth-evaluative proposition is assigned. For example, Cappelen and Lepore (2008) suppose that one needs to "disambiguate every ambiguous/polysemous expression in S. (Cappelen and Lepore, 2008, p.145)" in order to recover the minimal proposition from an utterance.

My general objection to this strategy is that polysemous senses are not discrete entities that can be disambiguated into. First, I argue that flexible uses of polysemy make it difficult to individuate and encode discrete polysemous senses. Second, encoding polysemous senses as discrete forms is incompatible with the fact that polysemy is productive. Third, I present some psycholinguistic evidence to show that disambiguation does not seem to exist for polysemy, which is different from homonyms.

4.5.1 Flexibility of Polysemy

First, our uses of polysemous words are very flexible. It is extremely difficult to individuate different uses of polysemous words into discrete bins of senses, because different polysemous senses are closely related. It is often arbitrary to set a threshold on the continuous scale of semantic relatedness, which is very different from categorizing different meanings of homonyms because their meanings are more clearly separated.

The individuation of polysemous senses is aggravated by the fact that we don't perceive word meaning in different contexts categorically. This is in sharp contrast with the categorical perception of speech sound. It is well known in psycholinguistics since the work of Liberman et al. (1957) that humans perceive phonemes categorically even though the acoustic features of different phonemes are continuous. For example, when we continuously increase the voice onset time of the vowel, our perceptual identification of the phoneme will change abruptly from /da/ to /ta/, hence categorical perception. However, we are not aware of this kind of abrupt change in the domain of word meaning but homonymy is like this. The continuity of polysemous senses makes the individuation of polysemous sense more arbitrary.

On the other hand, the difficulty and arbitrariness of sense individuation have been testified by centuries of lexicographical work. As a matter of fact, different dictionaries and lexical databases often give different sense individuations.

For example, for "paper:"

- WordNet

(15) S: (n) composition, paper, report, theme (an essay (especially one written as an assignment)) "he got an A on his composition."

(16) S: (n) paper (a scholarly article describing the results of observations or stating hypotheses) "he has written many scientific papers."

- The Oxford Dictionary of English

(17) an essay or dissertation, especially one read at an academic lecture or seminar or published in an academic journal: he published a highly original paper on pattern formation.

WordNet separates essays submitted in the class and essays submitted to journals and assigns them into different bins of senses, while The Oxford Dictionary of English only has one sense bin for class paper and journal paper, so it assigns them into the same one.

There is a deeper problem of linguistic categorization underlying sense individuation. Ambridge (2019) call it the lumping-or-splitting problem as discussed in Chapter 1. When individuating senses, theorists have to either lump different uses of a word into one sense or split them into different senses. Lumping is difficult because it may ignore the difference of meaning within the lumped sense category. Splitting is also difficult because there is no principled way to stop further splitting. This lumping or splitting problem marks the flexibility of polysemy that different senses of polysemous words are not clear cut and set in stone. People use different senses of polysemous words in widely different and nuanced ways.

The arbitrariness of sense individuation makes the strategy of disambiguating polysemy very impractical because it is hard to come up with a list of polysemous senses which can be disambiguated into the minimal proposition.

4.5.2 Productivity of Polysemy

Even if we can arrive at a workable manner of polysemous sense individuation, encoding polysemy as discrete senses does not fit with the productivity of polysemy.

One distinctive feature of human language is its productivity. We can produce and understand an unlimited number of linguistic expressions. Thus, there will always be new expressions that were not produced or understood in the past but could be produced and understood in the future. Productivity occurs at different levels of language. For example, we can combine morphemes to create new words, we can combine words to create new phrases, and we can combine phrases to form new sentences. The same feature of productivity applies to polysemy as well. In the following, I present examples of productivity that is at odds with the strategy of treating polysemy as ambiguity, because new senses are not encoded in the lexicon.

For example, names are polysemous among the person being named, their work, and so on.

(18) Shakespeare is well educated.

(19) People don't read Shakespeare here.

(20) Every library has some Shakespeare.

"Shakespeare" in (18) refers to the person Shakespeare, while "Shakespeare" and "Shakespeare" in (19) and (20) refer to the work or copies of work of Shakespeare. However, "Shakespeare" can also be used in productive ways as in (21) which refers to a digital file that can be encrypted.

(21) The Shakespeare in my Kindle is locked.

Not only can new senses come into existence, regular patterns of polysemous senses can be transferred to new words. Regular patterns can be seen among a category called regular polysemy. For example, there is ARTICLE-INFORMATION polysemy, such as "book" and "dictionary." Both of them have the ARTICLE sense and the INFORMATION sense as in (22)/(23) and (24)/(25). And these two senses have similar cases of use.

(22) This book is very heavy.

(23) This book is very difficult.

(24) I threw the dictionary to him.

(25) I revised the dictionary by myself.

However, these patterns are not only summarizations of old uses, but also generalizable to new words. For example, e-books as electronic books are digital files that emerged in the late 20th century. The use of the word "e-book" catches up with the ARTICLE-INFORMATION pattern very quickly. Even though an e-book cannot be literally thrown or torn, it has distinctive features that only belongs to it as an article.

Therefore,

(26) I deleted the e-book on my system.

(27) This is the best selling e-book this year.

There are other categories of regular polysemy, such as MEAT-ANIMAL polysemy that can be applied to new animals or animals that are never eaten before.

(28) Armadillo is an interesting animal.

(29) You shouldn't eat armadillo.

Even if the word "armadillo" has never been used to express the MEAT sense before, there would be no obstacle to understand that (29) means eating the meat of armadillo rather than the whole animal.

In sum, both individual senses and regular patterns of senses are productive. However, this productivity is at odds with the strategy of treating polysemy as ambiguity because new senses cannot be encoded in the lexicon beforehand. Therefore, new senses cannot be activated through disambiguation because disambiguation assumes a list-like representation of senses, which does not capture well the productivity of polysemy.

As a result, productivity makes it implausible for polysemous senses to be encoded as a list of discrete entities. Instead of encoding discrete senses of each animal's meat, maybe a

better approach is to posit lexical rules as in Pelletier (1975)'s and Copestake and Briscoe (1995)'s work or some kinds of generative mechanism such as co-composition as in Pustejovsky (1998)'s theory of Generative Lexicon, which leads to the third strategy in the next section.

4.5.3 Psycholinguistic Evidence against Polysemy Disambiguation

In this section, I am going to provide some psycholinguistic evidence that polysemy processing does not involve disambiguation. Disambiguation is the process of decision-making — choosing one among several interpretations. In general, disambiguation requires more cognitive processing, so polysemous words are slower to process compared to monosemous words.

There are several lexical decision tasks investigating polysemy processing. In lexical decision tasks, subjects are shown real words and fake words on a computer screen. The task is to press a button to decide whether the word is a real word or not, hence, a lexical decision. A lot of variables will influence the reaction time of lexical decisions such as the frequency of the word, length of the word, or priming conditions. Azuma and Van Orden (1997), Rodd et al. (2002), and Klepousniotou and Baum (2007) studied the relationship between reaction time and the relatedness of word senses. They found that words with related meanings (polysemy) have quicker reaction times than monosemous words, while words with unrelated meanings have a slower reaction time than monosemous words. (homonymy). Hence, polysemous words are processed faster than monosemous words, while homonyms with unrelated meanings are processed slower than monosemous words. This indicates that polysemous words are processed differently from homonyms in real-time processing.

Furthermore, the sensicality judgment experiment conducted in the previous chapters shows that human semantic processing is sensitive to the minute difference between tokens of polysemous words belonging to the same bin of sense. This indicates that human utilizes more nuanced continuous meaning than discrete bins of senses.

In sum, I argue that polysemy cannot be simply treated as ambiguity because (1) polysemy

is difficult to be classified into discrete senses; (2) disambiguating from a list of discrete senses is not compatible with the productivity of polysemy; (3) psycholinguistic evidence shows that polysemy is not processed like ambiguity. Therefore, semantic minimalists cannot resort to this strategy to handle polysemy.

4.6 Strategy 3: Assimilate Polysemy into the Basic Set of Context-Sensitivity

The last strategy for semantic minimalists is to assimilate polysemy into the basic set of context-sensitive expressions so that polysemy can be handled in a similar way as indexicals. As a result, different polysemous senses can be taken to be the outputs of some context-sensitive semantic operations, such as the character in Kaplan's theory of indexical (Almog et al., 1989), Stern (2000)'s Mthat operation or Pelletier (1975)'s lexical rule of grinding. The output of the context-sensitive operation is then integrated into the minimal proposition to capture its context-sensitivity.

On the one hand, lexical rules such as grinding (Pelletier, 1975; Copestake and Briscoe, 1995) are introduced to capture some specific rule-like patterns in polysemy such as ANIMAL-MEAT, CONTENT-ARTICLE polysemy. These rules convert a polysemous word from one of its senses to another before being contributed to the literal truth condition. For example, the meaning of "chicken" can be converted by the grinding rule from the animal sense to the meat sense in sentences such as "bodybuilders eat chicken every day." The problem with these kinds of approaches is similar to what I have discussed in the previous sections. Polysemous words with these kinds of regular patterns are only a small fraction of all polysemous words. Other polysemous words such as "line" or "case," have unique sense patterns that do not conform to rule-like patterns. As a result, the approach of lexical rules is not extensible to all polysemous words.

On the other hand, more general approaches for semantic minimalists can be based on

Kaplan's theory of indexicals. Kaplan develops his original approach to capture the apparent context-sensitivity of demonstratives and indexicals in semantics, such as "I," "here, " and "he." His approach includes distinguishing the character and the content of a word. The character of a word is a function from its context of evaluation to its content in the context. For example, the character of "I" is a function that takes the context of the utterance as an input and outputs the speaker of the utterance as its content. As a result, the content of "I" is context-sensitive, but the character of "I" itself is context-insensitive. Stern (2000, 2011) adapted Kaplan's approach by inventing the Mthat[t] operator. Mthat[t] "expresses a set of properties P presupposed to be m-associated with t in a context c," which are properties expressed by metaphors. "The metaphorical expression 'Mthat[t]' has a non-constant character, i.e., a meaning or rule (or function) that in different contexts yields different contents, i.e., different truth-conditional factors (for predicates, properties)." Similar adaptation can be made to design an operator — Pthat[t], that takes the context of a polysemous as an input and outputs the correct sense of the polysemous word. Hence, it also has a non-constant character to capture the context-sensitivity of polysemy. However, this approach faces the problem of being not explanatory enough. Even though it is certain that there exists such a function from the context of a polysemous word to its correct interpretation, giving out the function itself does not explain what the context-sensitive process of polysemy actually is. Particularly, providing the extension of the character function, the mapping between context and content, lacks the explanation of the mechanisms of the context-sensitive operation. Instead, providing the intension of the character as rules face the problems of comprehensibility as discussed in the last paragraph. As a result, a more flexible and comprehensive way to define the intention of this character function is needed.

Last but most importantly, it is doubtful that the approaches of assimilating polysemy into the basic set square with the thesis of semantic minimalism. Polysemous words are ubiquitous in any natural language. Based on Rodd et al. (2002)'s analysis of CELEX lexical database (Baayen et al., 1993), 84% of the words in English have more than one senses. Assimilating 84% of words into the basic set will make the basic set cease to be basic. Furthermore, most

polysemous words do not form a distinctive grammatical category as indexicals, so it goes against the tenet of semantic minimalism that every context-sensitivity in literal truth conditions is triggered by grammar. In sum, I argue in this section that minimalist approaches that assimilate polysemy into the basic set are not satisfactory, and they even contradict the thesis of semantic minimalism by allowing a wide range of context-sensitivity within the minimal proposition.

4.7 What is Needed to Account for Polysemy?

In this section, I want to discuss the reason why semantic minimalism does not handle polysemy well. It is because of its denial of semantic modulation in the literal truth condition of an utterance.

Polysemy, on the one hand, is largely literal, conventional, and regular in use. This makes polysemy different from non-literal uses of words such as metaphors, so it is hard to deny the semantic contribution of polysemy to the literal truth condition of an utterance. On the other hand, polysemy is very flexible and productive. They are used in very nuanced ways, which makes them difficult to be classified into discrete sense bins for future disambiguation, and it can be used in productive ways that have not been used before. I argue in the previous sections that these features of polysemy make it ill-suited to be handled within a minimalist approach, because polysemy requires theories to handle its flexible and context-sensitive contribution to the literal truth conditions of utterances.

As a result, a context-sensitive process within semantics is necessary to capture the phenomena of polysemy, which is semantic modulation. Modulation is the process where the meaning of a word is affected by the meanings of other words or situations in the context.[22] It belongs to the primary pragmatic process (Recanati, 2004), whose content is first available to our consciousness during comprehension. It is in contrast with the secondary pragmatic

[22]The terminology "modulation" is used in roughly the same way as Cruse (1986), Ruhl (1989), and Recanati (2004).

process, which involves inferential processes starting from the content derived in the primary pragmatic process. In the previous chapters, I argue that semantic modulation is various in terms of underlying mechanisms and continuous in terms of how far meaning is changed in context. These two features make semantic modulation fine-grained enough to capture the nuanced flexibility of polysemy within the literal truth condition of an utterance.

Furthermore, in order to capture the regularity and productivity of polysemy, I argue that polysemy is not just semantic modulation. Instead, It is the clustering and entrenchment of frequent semantic modulations from repetitive uses. When similar modulations repetitively occur, they form a cluster, which corresponds to our intuitive concept of polysemous senses. Furthermore, these patterns of clustering are entrenched in our memory so that they can be more readily used in the future. The productivity of polysemy comes from analogy with these entrenched clusters. The frequency and variability of these entrenched senses clusters determine how productive each word will be and where is the constraint on certain productive uses. In sum, words that are modulated in a certain way — forming multiple clusters, are polysemous. These meaning clusters are entrenched in memory to facilitate future productive uses through analogy.

However, admitting the existence of semantic modulation contradicts with (1),

> (1) a truth-evaluative minimal proposition can be recovered from an utterance if the values of grammatically triggered context-sensitive expressions are assigned and ambiguous words are disambiguated.

In principle, any word can be modulated by context, and most modulations are not bottom-up driven by grammar, neither syntax nor morphology. Instead, it is a free pragmatic process that occurs during linguistic production and comprehension. Therefore, (1) is false if admitting the existence of semantic modulation.

Recanati (2017b, 2019) provides a similar argument for the conclusion that polysemy supports semantic contextualism. Recanati more specifically points out that the mandatoriness of polysemy modulation supports radical contextualism because every polysemous word needs to

be modulated by the context in order to recover the literal truth condition of an utterance. In contrast, modulation of word meaning by context is optional in moderate contextualism. The type meaning of a word can either directly serve as the literal meaning of the utterance or being modulated. However, because the type meaning of a polysemous word is a combination of its different senses (Recanati adopts Langacker (2008)'s network representation of polysemy type meaning.), the modulation of a polysemous word is mandatory. Therefore, Recanati argues that the existence and pervasiveness of polysemy support the view of radical contextualism.

In sum, polysemy cannot be well handled by the semantic minimalist. Instead, one has to adopt semantic modulation, which is contradictory to the semantic minimalist thesis (1). As a result, polysemy goes against semantic minimalism and supports semantic contextualism.

Bibliography

Almog, J., Perry, J., and Wettstein, H. (1989). *Themes from Kaplan*. Oxford University Press.

Alonso, H. M. (2013). *Annotation of Regular Polysemy: An Empirical Assessment of the Underspecified Sense : PhD Dissertation*. Faculty of Humanities, University of Copenhagen.

Altmann, G. T., Garnham, A., and Dennis, Y. (1992). Avoiding the garden path: Eye movements in context. *Journal of Memory and Language*, 31(5):685–712.

Ambridge, B. (2019). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, page 0142723719869731.

Apresjan, J. D. (1974). Regular Polysemy. *Linguistics*, 12(142):5–32.

Armstrong, B. C. and Plaut, D. C. (2008). Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Computational and behavioral evidence. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

Armstrong, B. C. and Plaut, D. C. (2016). Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences. *Language, Cognition and Neuroscience*, 31(7):940–966.

Azuma, T. and Van Orden, G. C. (1997). Why SAFE Is Better Than FAST: The Relatedness of a Word's Meanings Affects Lexical Decision Times. *Journal of Memory and Language*, 36(4):484–504.

Baayen, R. H., Piepenbrock, R., and Van Rijn, H. (1993). The CELEX lexical database (CD-ROM). Linguistic data consortium. *Philadelphia, PA: University of Pennsylvania*.

Barak, L., Floyd, S., and Goldberg, A. (2019). Modeling the Acquisition of Words with Multiple Meanings. *Proceedings of the Society for Computation in Linguistics*, 2(1):216–225.

Borg, E. (2012). *Pursuing Meaning*. OUP Oxford.

Bowdle, B. F. and Gentner, D. (2005). The Career of Metaphor. *Psychological Review*, 112(1):193–216.

Brown, S. W. (2008). Polysemy in the mental lexicon. *Colorado Research in Linguistics*, 21(1):1–12.

Brugman, C. M. (1988). *The Story of over: Polysemy, Semantics, and the Structure of the Lexicon*. Taylor & Francis.

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers*, 30(2):188–198.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge University Press.

Cappelen, H. and Lepore, E. (2004). A tall tale: In defense of semantic minimalism and speech act pluralism. *Canadian Journal of Philosophy*, 34(sup1):2–28.

Cappelen, H. and Lepore, E. (2008). *Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism*. John Wiley & Sons.

Carston, R. (2008). *Thoughts and Utterances: The Pragmatics of Explicit Communication*. John Wiley & Sons.

Christiansen, M. H. and Chater, N. (1999). Connectionist Natural Language Processing: The State of the Art. *Cognitive Science*, 23(4):417–437.

Churchland, P. M. (1993). State-Space Semantics and Meaning Holism. *Philosophy and Phenomenological Research*, 53(3):667–672.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204.

Copestake, A. and Briscoe, T. (1995). Semi-productive Polysemy and Sense Extension. *Journal of Semantics*, 12(1):15–67.

Cree, G. S., McRae, K., and McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science*, 23(3):371–414.

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

Davidson, D. (1978). What metaphors mean. *Critical inquiry*, 5(1):31–47.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Devitt, M. (2013). What Makes a Property "Semantic"? In Capone, A., Lo Piparo, F., and Carapezza, M., editors, *Perspectives on Pragmatics and Philosophy*, Perspectives in Pragmatics, Philosophy & Psychology, pages 87–112. Springer International Publishing, Cham.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, 8(7):301–306.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, 33(4):547–582.

Elman, J. L. (2011). Lexical knowledge without a lexicon? *The mental lexicon*, 6(1):1–33.

Evans, C. and Yuan, D. (2017). A Large Corpus for Supervised Word-Sense Disambiguation.

Falkum, I. L. (2011). *The Semantics and Pragmatics of Polysemy: A Relevance-theoretic Account*. Doctoral, UCL (University College London).

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

Falkum, I. L. and Vicente, A. (2015). Polysemy: Current perspectives and approaches. *Lingua*, 157:1–16.

Ferreira, F. and Henderson, J. M. (1990). Use of verb information in syntactic parsing: Evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(4):555.

Foraker, S. and Murphy, G. L. (2012). Polysemy in Sentence Comprehension: Effects of Meaning Dominance. *Journal of memory and language*, 67(4):407–425.

Frisson, S. (2015). About bound and scary books: The processing of book polysemies. *Lingua*, 157:17–35.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., and Levy, R. (2019). Neural Language Models as Psycholinguistic Subjects: Representations of Syntactic State. *arXiv:1903.03260 [cs]*.

Glucksberg, S. and Keysar, B. (1993). How metaphors work. *Metaphor and thought*, 2:401–424.

Goldberg, A. E. (2019). *Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions*. Princeton University Press, Princeton, New Jersey.

Grice, H. P. (1991). *Studies in the Way of Words*. Harvard University Press.

Hale, J. (2001). A Probabilistic Earley Parser As a Psycholinguistic Model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational*

Linguistics on Language Technologies, NAACL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

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

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Hoffman, P., McClelland, J. L., and Lambon Ralph, M. A. (2018). Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychological Review*, 125(3):293–328.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, 81(10):3088–3092.

Hopfield, J. J. and Tank, D. W. (1985). "Neural" computation of decisions in optimization problems. *Biological cybernetics*, 52(3):141–152.

Jamieson, R. K., Avery, J. E., Johns, B. T., and Jones, M. N. (2018). An Instance Theory of Semantic Memory. *Computational Brain & Behavior*.

Jones, M. N. (2019). When does abstraction occur in semantic memory: Insights from distributional models. *Language, Cognition and Neuroscience*, 34(10):1338–1346.

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. and Baum, S. R. (2007). Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. *Journal of Neurolinguistics*, 20(1):1–24.

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.

Kocmi, T. and Bojar, O. (2017). An Exploration of Word Embedding Initialization in Deep-Learning Tasks. *arXiv:1711.09160 [cs]*.

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232.

Kripke, S. A. (1980). *Naming and Necessity*. Harvard University Press.

Lakoff, G. and Johnson, M. (2008). *Metaphors We Live By*. University of Chicago Press.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar: Theoretical Prerequisites*, volume 1. Stanford university press.

Langacker, R. W. (2008). *Cognitive Grammar: A Basic Introduction*. Oxford University Press, USA.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4(1):151–171.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5):358–368.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

McClelland, J. L., John, M. S., and Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes*, 4(3-4):SI287–SI335.

McRae, K. (2004). Semantic memory: Some insights from feature-based connectionist attractor networks. *The psychology of learning and motivation: Advances in research and theory*, 45:41–86.

McRae, K., Spivey-Knowlton, M. J., and Tanenhaus, M. K. (1998). Modeling the Influence of Thematic Fit (and Other Constraints) in On-line Sentence Comprehension. *Journal of Memory and Language*, 38(3):283–312.

Merity, S., Keskar, N. S., and Socher, R. (2017). Regularizing and Optimizing LSTM Language Models. *arXiv:1708.02182 [cs]*.

Merity, S., Keskar, N. S., and Socher, R. (2018). An Analysis of Neural Language Modeling at Multiple Scales. page 10.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2016). Pointer Sentinel Mixture Models. *arXiv:1609.07843 [cs]*.

Mihalcea, R. (1998). Semcor semantically tagged corpus. *Unpublished manuscript*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). 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.

Nunberg, G. (1995). Transfers of Meaning. *Journal of Semantics*, 12(2):109–132.

Passonneau, R. J., Baker, C., Fellbaum, C., and Ide, N. (2012). The MASC word sense sentence corpus. In *Proceedings of LREC*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., and Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037.

Pelletier, F. J. (1975). Non-singular reference: Some preliminaries. *Philosophia*, 5(4):451–465.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pustejovsky, J. (1998). *The Generative Lexicon*. MIT Press.

Rabagliati, H., Gambi, C., and Pickering, M. J. (2016). Learning to predict or predicting to learn? *Language, Cognition and Neuroscience*, 31(1):94–105.

Rabovsky, M., Hansen, S. S., and McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, 2(9):693–705.

Recanati, F. (2004). *Literal Meaning*. Cambridge University Press.

Recanati, F. (2010). *Truth-Conditional Pragmatics*. OUP Oxford.

Recanati, F. (2017a). Contextualism and Polysemy. *Dialectica*, 71(3):379–397.

Recanati, F. (2017b). Local pragmatics: Reply to Mandy Simons. *Inquiry*, 60(5):493–508.

Recanati, F. (2019). Why Polysemy Supports Radical Contextualism. In Bella, G. and Bouquet, P., editors, *Modeling and Using Context*, Lecture Notes in Computer Science, pages 216–222, Cham. Springer International Publishing.

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. (2020). Settling Into Semantic Space: An Ambiguity-Focused Account of Word-Meaning Access. *Perspectives on Psychological Science*, page 1745691619885860.

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.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420.

Ruhl, C. (1989). *On Monosemy: A Study in Linguistic Semantics*. SUNY Press.

Sennet, A. (2016). Polysemy.

Sperber, D. and Wilson, D. (1996). *Relevance: Communication and Cognition*. Wiley-Blackwell, Oxford ; Cambridge, MA, second edition.

Stern, J. (2000). *Metaphor in Context*. MIT Press.

Stern, J. (2011). Metaphor and minimalism. *Philosophical Studies*, 153(2):273–298.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Wearing, C. (2006). Metaphor and What is Said. *Mind & Language*, 21(3):310–332.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

Jiangtian Li

Current position

PhD Candidate, Department of Philosophy, Western University

Areas of specialization

Philosophy of Language, Natural Language Semantics, Psycholinguistics, Computational Modeling of Cognition.

Areas of competence

Philosophy of Mind and Psychology, Natural Language Processing with Deep Learning

Education

2016-now PhD Candidate in Philosophy, Western University
Advisor: Robert Stainton.
Dissertation: A Theory of Polysemy

2015-2016 MA in Philosophy, Western University
Advisor: Robert Stainton.
Thesis: Names are Also Polysemous

2009-2013 BA in Philosophy, Minzu University of China
Advisor: Jixuan Zhang.
Thesis: Truth and Meaning —- a Defense of Propositional Deflationism

Grants, honours & awards

2010 Hong Kong Xinshan Scholarship
2013 Secondary Professional Scholarship
2014 Fellowships from Philosophy Summer School in China and Ranked 1st
2015 Chair's Entrance Scholarship

Teaching

2016 fall Teaching Assistent of *Understanding Science*
2017 winter Teaching Assistent of *Introduction to Philosophy of Language*
2017-2018 Teaching Assistent of *Critical Thinking*
2018 fall Teaching Assistent of *Understanding Science*
2019 winter Teaching Assistent of *Philosophy of Neuroscience*
2019 fall Co-instructor of *Introduction to Philosophy of Mind*

Academic activities

2015-2016 Assistant on the category of Philosophy of Language in PhilPapers
2016 Attending the North American Summer School on Logic, Language, and Information (NASSLLI)
2017 Commenter in UWO's Annual Graduate Conference in Philosophy of Mind, Language, and Cognitive Science (PhilMiLCog)
2018 Commenter in UWO's Annual Graduate Conference in Philosophy of Mind, Language, and Cognitive Science (PhilMiLCog)
2020 Graduate Research Assistant in Univeristy of Western Ontario

Languages and professional skills

Computational: Python, R, PyTorch, LaTex, JavaScript, Pychopy
Language: English (fluent), Chinese (native), Japanese (fluent), Latin (reading)

Last updated: September 3, 2020 •