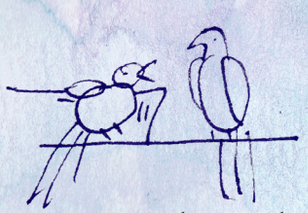


Everyone knows that the meaning of words can change. But how does this happen? This thesis is premised on the idea that it mainly happens in interaction, when language is at its most flexible. In a series of studies, we use formal and computational methods to investigate semantic variation and change at the sites of linguistic interaction, from conversations to community.

Bill Noble is a computational linguist working on semantic change and dialogue. He lives and works in Gothenburg, Sweden.

ISBN: 978-91-8069-205-2 (PRINT)
ISBN: 978-91-8069-206-9 (PDF)

SEMANTIC CHANGE IN INTERACTION | Bill Noble 2023

Semantic change in interaction

Studies on the dynamics of lexical meaning

Bill Noble

DEPARTMENT OF PHILOSOPHY,
LINGUISTICS AND THEORY OF SCIENCE

GÖTEBORGS UNIVERSITET

Doctoral thesis in computational linguistics

Semantic change in interaction

Studies on the dynamics of lexical meaning

Bill Noble

April, 2023

UNIVERSITY OF GOTHENBURG

Semantic change in interaction:
Studies on the dynamics of lexical meaning
Supervisors: Staffan Larsson and Asad Sayeed

The author is supported by grant 2014-39 from the Swedish Research Council (VR) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg

Cover by Noah Mease
Printed in Sweden by Stema Specialtryck AB
Publisher: University of Gothenburg (Dissertations)

Distribution:
Department of Philosophy, Linguistics and Theory of Science,
University of Gothenburg
Box 100, SE-405 30 Gothenburg

ISBN:
978-91-8069-205-2 (print)
978-91-8069-206-9 (PDF)

Part I is available online at
http://hdl.handle.net/2077/74969.

for my parents

Abstract

This compilation thesis investigates how word meanings change. In particular, it's concerned semantic change at the levels of *interaction* and the *speech community*. To this end, the compiled studies employ methods from both formal and computational semantics.

The first study presents a model for, and companion annotation study of, *word meaning negotiation*, a conversational routine in which the meaning of a word becomes an explicit topic of conversation. The next two studies introduce and apply *classification systems*, a model of communal conceptual resources for ordering and talking about a particular domain. We use a formalization thereof to model how *genus-differentia definitions* can be used in interaction to update lexical knowledge of perceptual categories. The next study considers a related phenomenon, *perceptual category description*, but this time from a computational perspective. By modeling a short interaction between two neural networks, we investigate how different ways of representing perceptual categories affect linguistic grounding. Following that, we turn to the dynamics of social meaning, particularly the meaning of implicit conversational assumptions called *topoi*, with a focus on situations of involving uncertainty about the speaker's social identity. The final two studies of the thesis shift the focus from particular interactions to the level of the community. First, we investigate linguistic variation using *community conditioned language models* to learn vector representations for a collection of online communities. These language-based representations are found to correlate with community representations based on community membership alone. Finally, we use diachronic distributional word vectors to study *short-term semantic shift* in online communities. We find that semantic change has a significant yet nuanced relationship with the social structure of the community.

Altogether, the compilation offers two main insights. First, semantic plasticity is directly related to the complexity of the lexical semantic system. Words exhibit both perceptual and inferential meaning potential, each of which play a role in conveying and learning new meanings. Monolithic representations of word meaning belie a structured flexibility that guides how words can be used, while providing opportunities for innovation. It is this flexibility that is often the site of new conventionalized meanings. Second, semantic change is rooted in the interactive practices of the community. Communities sustain the communicative norms that govern how linguistic interaction takes place. These norms also provide a framework for negotiating meaning, and comprise the social and semiotic context that supports semantic innovation and change.

Sammanfattning

Denna sammanläggningsavhandling undersöker hur ordbetydelser förändras. Mer specifikt handlar den om semantisk förändring på *interaktionsnivå* och på *språkgemenskapsnivå*. I detta syfte använder de i avhandlingen ingående studierna metoder från formell och komputationell semantik.

Den första studien presenterar en modell för, och en tillhörande annoteringsstudie av, *ordbetydelseförhandling*, ett samtalsmönster där ett ords betydelse blir det explicita samtalsämnet. De följande två studierna introducerar och tillämpar *klassificeringssystem*, en modell av gemensamma begreppsliga resurser som används för att organisera och tala om en viss domän. Vi använder en formalisering av klassificeringssystem för att modellera hur *genus-differentiae-definitioner* kan användas i interaktion för att uppdatera lexikal kunskap om perceptuella kategorier. Nästa studie behandlar ett besläktat fenomen, *perceptuella kategoribeskrivningar*, men denna gång från ett komputationellt perspektiv. Genom att modellera en kort interaktion mellan två neurala nätverk undersöker vi hur olika sätt att representera perceptuella kategorier påverkar språkligt delande av information. Därefter vänder vi oss till den sociala betydelsekomponentens dynamik och då särskilt med avseende på betydelsen hos underförstådda antaganden, så kallade *topoi*, och med fokus på situationer där det finns en osäkerhet om talarens sociala identitet. De två sista studierna i denna avhandling skiftar fokus från specifika interaktioner till språkgemenenskapsnivån. Först undersöker vi språklig variation med hjälp av *gemenskapsvillkorade språkmodeller* som lär sig vektorrepresentationer för grupper av onlinegemenskaper. Dessa språkbaserade representationer visar sig korrelera med gemenskapsrepresentationer som enbart grundas i gemenskapstillhörighet. Slutligen använder vi oss av diakroniska distributionella ordvektorer för att studera *kortsiktig semantisk förändring* i onlinegemenskaper. Vi finner att semantisk förändring har signifikanta men nyanserade samband med gemenskapens sociala struktur.

Sammantaget ger sammanställningsavhandlingen två huvudsakliga insikter. För det första är semantisk plasticitet direkt kopplad till det lexikala semantiska systemets komplexitet. Ord har både perceptuella och inferentiella betydelsepotentialer, och båda dessa aspekter spelar en roll i överförandet och inlärandet av nya betydelser. Monolitiska representationer av ordbetydelse bortser från en strukturerad flexibilitet som vägleder hur ord kan användas samtidigt som de erbjuder möjligheter till språklig innovation och förändring. Det är denna flexibilitet som ofta ligger till grund för uppkomsten av nya konventionaliserade betydelser. För det andra är semantisk förändring rotad i språkgemenskapens interaktiva praktiker. Språkliga gemenskaper upprätthåller de kommunikativa normer som styr den språklig interaktionen. Normerna erbjuder också ett ramverk för förhandladet av betydelser, och utgör den sociala och semiotiska kontext som möjliggör semantisk innovation och förändring.

Contents

Acknowledgements

It has been a long journey. My language and my world have changed a lot along the way. And there are so many people responsible for making that journey and this thesis possible.

First, I would like to acknowledge the substantial contributions of my PhD advisors, Staffan Larsson and Asad Sayeed. I spent so many hours talking through the ideas of this thesis with Staffan. How many times did I obstinately take another route, only to come back to his initial suggestion? Certainly more than once. Staffan also diligently helped me to push the thesis to completion and deserves much of the credit for making sure that the defense will, in fact, happen. Asad has had a major influence, not only on the thesis, but also on how I have come to view academic life and the project of research. Staffan and Asad were instrumental in helping to figure out how to frame the compilation and gave very important comments on multiple drafts of the kappa.

In addition to my advisors, I was fortunate to have many other co-authors on papers in the compilation. Each and every one of them have been a joy to work with and the compilation is at least as much theirs, collectively, as it is mine. Robin Cooper has, been a generous and careful mentor. I greatly admire the way he deftly draws out and nurtures the ideas of others. It was Raquel Fernández who first introduced me to computational linguistics research, and it was such a thrill to collaborate with her again, along with Staffan and Asad. I have had countless conversations with Ellen Breitholtz about social and pragmatic meaning in interaction. Her influence on the direction of the thesis goes well beyond our collaboration on Chapter 11. Jean-Philippe Bernardy spent many hours talking through algebraic and probabilistic modeling with me. Working with him has had a substantial impact on how I think about computation and formalization in linguistics. Kate Viloria's contribution to Chapter 7 was made as part of an independent study course for the Master's of Language Technology, in which she was an excellent student and collaborator. A special thanks goes to Nikolai Ilinykh for working with me to prepare a new and improved version of Chapter 10 in time for inclusion in the thesis, much to the chagrin of the poor people at Victor's Cafe, who were forced to, yet again, listen to us go on and on about birds.

I would also like to thank my other research collaborators at GU and beyond: Vladislav Maraev, Adam Ek, Julain Grove, Ben Clarke, Fahima Ayub Khan, Eleni Gregoromichaelaki, Christine Howes, Chiara Mazzocconi, Simon Dobnik, and Vidya Somashekarappa. I have been fortunate to work with so many people on a wide range of topics in computational linguistics. These collaborations have been hugely influential in shaping what would ultimately become the thesis, even when our work was outside the narrow

scope of the compilation.

Andy Lücking was the "green reader" for the thesis. His insightful comments were instrumental in shaping the final version of the kappa and will undoubtedly influence any future work I do on this topic. Nikolai Ilinykh and Noah Mease also gave comments on drafts of the kappa that helped a lot to improve its readability and cohesiveness.

I want to thank the members of CLASP, especially Sharid Loáiciga, Stergios Chatzikyriakidis, and Shalom Lappin for fostering a rich, nurturing research environment. I'd also like to acknowledge dialogue reading group, which has served as my academic "home base" during my PhD. Many of the ideas developed in the thesis were first incubated there.

The administration at the Department of Philosophy, Linguistics and Theory of Science (FLoV) supports our research in ways that I only begin to understand. I would especially like to thank Susanna Myyry, Hannna Edblom, and Iines Turunen. Christopher Kullenberg and Johan Söderberg also helped me a great deal with administrative questions leading up to the defense.

My fellow PhD students in FLoV have made our department a welcoming place that I was excited to go to each day. Thank you for your companionship.

Thank you to the Wannerskog family Maria, Lasse, Anna-Sofia, Lars, Miranda, Svante, Olaf, Kalle, Åsa, Axel, and Johanna. From the day I first landed in Göteborg you have made me feel like I belong. I can't begin to express my gratitude. Tack så jätte mycket.

Thank you to my friends and family back home for always being there for me. For countless hours on the phone, for board games over the internet, for visiting me from all the way across the Atlantic Ocean, and for keeping a place for me to come home to. Mom and Dad, Ryan, Brendan, Jack, Claire, Todd, Khanh-Anh, Dustin, and Ray, thank you.

And, of course, to Noah, thank you. How could I be so lucky?

Preface

This is a compilation thesis, meaning that the main scientific contributions come mostly in the form of studies that have previously been published as conference papers.[1] All these papers are, in one way or another, computational studies of variation and change in natural language. Those papers are reproduced in Part II of the print version of this thesis. Summaries of the studies, with links to the archival version of the papers can be found in Section 6.1.

Part I of the thesis, is what is colloquially referred to as the *kappa*, (English: *coat* or *cover* neé *hat*). Chapters 1 to 4 set up the theoretical framework that underpins the work in Part II of the thesis and Chapter 5 discusses the methodologies that are used. Finally, Chapter 6 (in addition to the summaries) offers some concluding remarks.

Naturally, some background and methodological exposition can be found in the introductory sections of the individual papers, but it is brought together a way that motivates the overall research outlook of the thesis. I hope, too, that the kappa makes the work available to a broader audience. Conference papers tend to be written with attendees (and especially reviewers) of the specific conference mind. Given the space constraints of a typical conference paper, this often means that a lot of theoretical background is assumed or introduced in a more cursory way than it would be for a more general audienceeven perhaps an audience familiar with the field of computational linguistics more broadly.

Computational linguistics is a highly collaborative field. All of the work in this compilation was all carried out in close collaboration with my PhD supervisors and other members of the computational linguistics community in Gothenburg, particularly at the Centre for Linguistics and Studies in Probability (CLASP) where I have been fortunate to be employed as a PhD student. I use the word *we* a lot in the thesis. Often times that's because the work being described was very much a joint effort. In other places, I'm just hoping to include you, the reader, in this adventure we're about to embark on.

Part II is not included in the PDF version of the thesis. However, the papers that comprise the chapters of Part II are all freely available online in their original format. Throughout Part I, the studies included in the compilation are referred to by their chapter numbers in Part II. Both the original citation for each of the chapters and links to the online versions can be found below.

[1] With the exception of Chapter 10, which has been published as a pre-print on ArXiv.

Chapter 7 Noble, B., Viloria, K., Larsson, S., & Sayeed, A. (2021). What do you mean by negotiation? Annotating social media discussions about word meaning. *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*

http://semdial.org/anthology/papers/Z/Z21/Z21-3016/

Chapter 8 Noble, B., Larsson, S., & Cooper, R. (2022a). Classification Systems: Combining taxonomical and perceptual lexical meaning. *Proceedings of the 3rd Natural Logic Meets Machine Learning Workshop (NALOMA III)*, 11–16

https://aclanthology.org/2022.naloma-1.2

Chapter 9 Noble, B., Larsson, S., & Cooper, R. (2022b). Coordinating taxonomical and observational meaning: The case of genus-differentia definitions. *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*

http://semdial.org/anthology/papers/Z/Z22/Z22-3020/

Chapter 10 Noble, B., & Ilinykh, N. (2023). Describe me an Aucklet: Generating Grounded Perceptual Category Descriptions. https://doi.org/10.48550/arXiv.2303.04053

https://arxiv.org/abs/2303.04053

Chapter 11 Noble, B., Breitholtz, E., & Cooper, R. (2020). Personae under uncertainty: The case of topoi. *Proceedings of the Probability and Meaning Conference (PaM 2020)*, 8–16

https://aclanthology.org/2020.pam-1.2/

Chapter 12 Noble, B., & Bernardy, J.-P. (2022). Conditional Language Models for Community-Level Linguistic Variation. *Proceedings of the 5th Workshop on NLP+CSS at EMNLP 2022*, 59–78

https://aclanthology.org/2022.nlpcss-1.9/

Chapter 13 Noble, B., Sayeed, A., Fernández, R., & Larsson, S. (2021). Semantic shift in social networks. *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, 26–37. https://doi.org/10.18653/v1/2021.starsem-1.3

https://aclanthology.org/2021.starsem-1.3

Part I.

Kappa

1. Introduction

> [...] the whole theory of language can be reduced to one question: what is the relationship between prevailing usage and the speech of an individual? How is the speech of an individual determined by prevailing usage in the community, and how in turn does the individual's speech affect prevailing usage?
>
> Herman Paul (1886)
> trans. Peter Auer (2015)

We know that words change. In the early 20th century the word *gay* meant *happy* or *joyous* in English. Now it almost always refers to sexuality. *Awesome* used to mean something awe-inspiring, frightening, even. Now it can also be used to mean *really very good*. Historical changes like these are well-documented, occurring in every language and at every time in history. We also know that speakers can coordinate a special vocabulary of word-meaning parings for collaborating on a project (Brennan & Clark, 1996), or even as a way of building and expressing intimacy (Hopper et al., 1981). But what do these two kinds of semantic plasticity have to do with one another? Is there a connection between these changes that take place on a historic time-scale and across whole languages and the *ad hoc* conventions that we develop for a particular communicative context? It would seem that there must be. After all, what makes a language if not the all of its individual occasions of use? And yet it is difficult to observe the transition from one to the other. **How does semantic coordination at the level of interaction relate to lexical change on the community level?**

On the other hand, when we consider variation across communities, we have something of a paradox. One view of language is that it is a channel for communication — a *code* in which one person can transmit information about the world to another person. Such a code functions best, one might assume, if its symbols and their meanings are perfectly aligned between the two speakers. To maximize efficiency the code should be stable be shared among as many speakers as possible. And yet this is not the situation in which we find ourselves. We have *many* different languages — not only among what are sometimes called the *macrolanguages* of the world. We also see a great deal of variation in the way that different communities communicate *within* these traditionally defined languages and dialects. There is a special *lingo*, *slang*, *jargon*, etc. associated with just about every vaguely communal activity you can think of. **Why does word meaning change across time and context?**

Unfortunately, this thesis will not answer either of these grand questions directly.

But they will nevertheless serve as two guiding stars as we set forth. The questions we do address in this thesis are small steps towards a better understanding of the dynamics of lexical meaning.

We start on the level of interaction, specifically interactions involving explicit talk about meaning to investigate the following questions:

1. What interactive resources are drawn upon when word meaning becomes an explicit topic of conversation? (Chapter 7)

2. When someone defines a word for us, how do we incorporate that meaning into our existing conceptual structure? (Chapters 8 and 9)

3. How does the way that perceptual categories are represented affect descriptions of those categories? (Chapter 10)

Continuing on theme of the dynamics of interaction, we consider the plasticity of certain *social signals*:

4. How does the meaning of a social signal change depending on what we learn about a speaker's social identity or ideology? (Chapter 11)

Finally, we shift our focus to the community level, addressing questions about how the character of communities relates to linguistic variation and change:

5. How do the linguistic particularities of a community correlate with its social makeup? (Chapter 12)

6. How does the social structure of a community affect the rate at which its word meanings change? (Chapter 13)

Along the way, we will come face to face with some of the most challenging questions in lexical semantics. Why is word meaning so flexible? How can it be that words have multiple senses? How do we synthesize seemingly incompatible aspects of word meaning? Does it matter to linguistics how words are associated with their meanings on a cognitive level? In Chapter 2, we set the stage to deal with these issues as they arrive. Chapter 3 situates lexical meaning in its natural habitat — in communities and in interaction. Finally, Chapter 4 gives some background on semantic variation and change, which is necessary to contextualize the contributions of the thesis.

The process of *lexicalization* — when new meanings become conventional — is difficult to observe directly. Because of this, the studies in Part II employ a variety of methods, from formal semantics to machine learning to investigate (1) interactions that *might* result in a speaker updating their understanding of the language of the community and (2) analysis of variation and change in small communities and across short

time periods. It's important to keep in mind how all these methods can be used to answer questions — what their limitations are and how insights gleaned using a certain methodology should be synthesized with the rest of the work. Chapter 5 introduces the methods used in the compilation.

Finally, Chapter 6 summarizes the studies presented in Part II and offers a synthesis of the conclusions that emphasizes the importance of lexical complexity and community-level interactive practice in understanding semantic change.

2. Lexical meaning

Why is a raven like a writing-desk?

Alice's Adventures in Wonderland
Lewis Carroll

In order to understand how words change in meaning, it's important to have a bit of background on what the field of linguistics understands meaning to *be*. In contemporary linguistics, *semantics* (the study of meaning) encompasses two fairly distinct sub-fields: **lexical semantics**, which studies the meaning of words (or, as we will discuss shortly, *lexical items*), and **compositional semantics**, which studies the meaning of larger linguistic units formed according to the language's *syntax*. Although we can never stray far from questions of compositional meaning, this thesis is primarily concerned with the dynamics of lexical meaning.

Lexical meaning implies a *lexicon*, i.e., a *book of words*, which suggests a certain model of natural language semantics in which the compositional and lexical aspects of a language are distinct modules. That model is reflected in the disciplinary division mentioned above, and is implicit in the **principle of compositionality**, which says that the meaning of a natural language expression is a *function* of *the meaning of its parts* (which come from the lexicon) and *how they are combined* (which comes from the syntax). This means that two expressions that use the same words can have different meanings:

(1) a. The man points at a dog.

b. The dog points at a man.

In English, this particular grammatical construction assigns different semantic roles (usually called *agent* and *patient*) to the two syntactic positions (*subject* and *object*). The meaning of (1a) differs from that of (1b) as a result of the compositional semantics, not because of the meaning of the words involved. But, of course, the meaning of the words does also matter:

(2) The man points at an apple.

To the extent that (2) differs from (1a) the principle of compositionality says that this difference must be explained by differences in the meaning of the words *dog* and *apple*.

The division of semantic labor implied by the principle of compositionality points to a *dictionary and grammar book* view of linguistic competency (Taylor, 2012), which says that with these two (more or less distinct) sources of knowledge, we have everything that is needed to understand and produce meaningful expressions in a given language. When studying variation and change of word meaning, it's helpful to consider how this "dictionary", which linguists call the **lexicon**, is structured. As a first attempt, we might consider the structure of a literal dictionary. Consider this entry for the word *point*:[1]

Point

noun (*plural*: **points**)

1. A discrete division of something.
    a) An individual element of a larger whole; a particular detail, thought, or quality. *The Congress debated the finer **points** of the bill.*
    b) A particular moment in an event or occurrence; a juncture. *There comes a **point** in a marathon when some people give up.*
    c) A focus of conversation or consideration; the main idea. *The **point** is that we should stay together, whatever happens.*
    d) A purpose or objective, which makes something meaningful. *Since the decision has already been made, I see little **point** in further discussion.*
2. A sharp extremity.
    a) The sharp tip of an object. *Cut the skin with the **point** of the knife.*
    b) An object which has a sharp or tapering tip. *His cowboy belt was studded with **points**.*
    c) A peninsula or promontory.
    d) [*falconry*] The perpendicular rising of a hawk over the place where its prey has gone into cover.

verb (*third-person singular simple present*: **points**:, *present participle*: **pointing**, *simple past and past participle*: **pointed**)

3. (*intransitive*) To extend the index finger in the direction of something in order to show where it is or draw attention to it. *It's rude to **point** at other people.*
4. (*intransitive*) To draw attention to something or indicate a direction *The arrow of a compass **points** north. The skis were **pointing** uphill.*
5. (*transitive, sometimes figuratively*) To direct towards an object; to aim *to **point** a gun at a wolf, or a cannon at a fort*
6. [*nautical*] (*intransitive*) To sail close to the wind *Bear off a little, we're **pointing**.*

The remainder of this chapter will use the dictionary model of the lexicon as a starting point for introducing lexical semantics topics that relate to variation and change. In Section 2.1, organizational structure of the dictionary which is, at its heart, a list of words where each item has its own distinct entry. Section 2.2 will talk about the

[1] This example is abridged and edited from the English-language Wiktionary entry for *point*, https://en.wiktionary.org/w/index.php?title=point, accessed December 1, 2022.

structure of the entry, which is itself a list of *senses*. Since words by their very nature exhibit variation in their contexts of use, how do we decide when, if at all, to make a distinction between different senses of the same word? When we do distinguish between senses, how do they relate to each other? This brings us to the question of semantic *generality*, which is discussed in Section 2.3. A word like *point* can apply to broad range of different real-world situations. How do we know what is included in its extension? Finally, Section 2.4 discusses two different *kinds of meaning* that arise from expectations we have of competent speakers.

2.1. Lexicality

Linguists disagree about the degree to which lexical and grammatical information can be considered distinct. But there does seem to be something to the idea of a flexible store of discrete lexical information that can slot in to a relatively stable compositional semantics.

To borrow an example from Larsson (2021), suppose you've never heard of a *wax jambu* (perhaps you haven't). And suppose I say to you *a wax jambu is pear-shaped fruit with a texture similar to that of an apple.* Your knowledge of wax jambus is still pretty incomplete, but you probably already have a pretty good idea of how to use the word in a sentence. Your knowledge of English morphology lets you construct and recognize the (admittedly awkward) plural *wax jambus*. You can even understand sentences like this one:

(3) The man points at a wax jambu.

You may not be able to call up a perfectly vivid image of a situation described by (3), but you can be assured that the difference in meaning between (3) and (2) is a function of the difference between apples and wax jambus, not because of how the sentence is formed or because of something about the meaning of the word *point*.

Lexicality is the dual of the principle of compositionality. Together, lexicality and the principle of compositionality give us a situation where lexical knowledge of specific words can change over time or vary across communities while the rest of the language remains stable. That the lexicon is so mutable means that we can study lexical change over relatively short periods of time and variations across relatively small communities of speakers, whereas grammatical changes are slower to manifest.

Example (3) brings up another issue which we should address now, and which is central to the notion of lexicality We've said that the topic of this thesis is how *word meanings* differ across communities and change over time. But this isn't quite true, depending what is meant by *word*. In deciding what to focus on as the unit of analysis, it is useful to look ahead to the phenomena we're interested in. Namely, we want to investigate differences in the *conventional association* between linguistic symbols and

their semantic meaning. These conventional associations are what is captured by the lexicon.

Orthographically *wax jambu* is two words (there is a space in between). Grammatically, *wax jambu* is an English adjective-noun construction. But the meaning of this expression cannot (at least not completely) be understood as a function of the meanings of *wax* and *jambu* and the grammar of adjective-noun constructions — otherwise we might think that a wax jambu is a jambu made out of wax (it isn't). Since the meaning can't be derived compositionally, *wax jambu* must have its own *lexical entry*. What we are really interested in in this thesis is not words in the grammatical sense, but the meaning of **lexical items** — any expression that has meaning which can't entirely be understood as composed of smaller parts. Informally, though, lexical items will still be referred to as *words* where it does not cause confusion.

Expressions like *laser printer*, *paperback book*, and *cell phone* might also be considered lexical items. "Phrasal verbs" like *let on*, *look up*, or *break down* might also be considered lexical items. There are also idioms like *cut corners* or *easy as pie* that encode some conventional meaning that can't be derived from the compositional semantics.[2] We might also like to consider expressions that you wouldn't find in a traditional dictionary. Think of sounds like *uh-huh*, which are often used as *backchannels* during another speakers turn to indicate understanding or agreement. Are such sounds words? They certainly play a meaningful role in conversation. The same can be said of laughter, including different qualities of laughter, which can have a variety of different communicative functions (Mazzocconi et al., 2022). Different morphological inflections of the same stem might be considered different words — for example, *jump*, *jumps* and *jumping*. But they derive their meaning from a single lexical item, which interacts with English morphology in a predictable way.

Non-compositional meaning can also accrue to longer expressions, such as idioms. *Kick the bucket* has a meaning that can't be discovered through compositional analysis. Should idioms be considered lexical items? By the criteria we have established they should, though including idioms in the lexicon poses a threat to our intuition that lexical items are word-like. Some linguists go so far as to take the view that *constructions*, which includes words as well as longer phrases and even syntactic patterns, are the fundamental building blocks of meaning (Croft, 2001). A related project is the *Generative Lexicon* (Pustejovsky, 1995), which redistributes the work of compositional meaning to the lexical level.

The question of what to count as a lexical item becomes operationally important when we want to conduct a corpus-based study of semantic variation or change. Of particular relevance are the pre-processing steps of **tokenization** (breaking up text into

[2] As with *wax jambu* it may be possible to get a rough idea of what some of these expressions mean by understanding the meaning of the words that make them up and the rules of English noun phrase composition. (Moon, 2015) argues that there is continuum in the degree to which such phrases can be considered together as multi-word expressions (see also Bücking (2010)). The important thing as far as we are concerned for the moment is that there is *some* aspect of the meaning which is conventional, not derivable from the composition of smaller units or general principles of communication (i.e., pragmatics; see Section 3.4.1).

discrete units of analysis) and lemmatization (normalizing different morphological inflections of the same lexical item). Section 5.2 discusses these issues in more detail, but for the most Part II proceeds with the assumption that lexical meaning *mostly* resides at the word level.

2.2. Polysemy

A word that has multiple *senses* is said to be **polysemous**. In the dictionary entry for *point*, each of the listed items represents a different sense. Polysemy is ubiquitous in natural language — it is hard to think of a word in English that *isn't* polysemous, at least to some degree. Different senses differ in meaning, and can also have different syntactic types. *Point*, for example, has both noun and verb senses.

Polysemy is important background because one of the main ways that words change in meaning is to gain or lose senses. Likewise, when a word is used differently in some community, it is often because it has an additional sense with a meaning specific to that community. In our example, senses 2d and 6 are special senses of *point* specific to falconry and nautical settings. Chapter 4 will discuss the relationship between polysemy and semantic variation and change in more detail, but for now we maintain a synchronic perspective.

When the meaning of an expression is undetermined with respect to two or more alternative interpretations, it is said to be **ambiguous**. Polysemy, then, is *lexical ambiguity*. When a polysemous word is used in a context that does not make clear which sense is meant, it can result in an ambiguous compositional expression, as in this example:

(4) When asked about his favorite 19th century American author, Jack pointed to the works of Louise May Alcott.

Since *works* itself is ambiguous in this context, the speaker could mean that Jack literally *pointed*$_3$ to a physical collection of books or it could mean that he figuratively *pointed*$_4$ to the abstract collection of literature that is the works of Louise May Alcott.[3]

Polysemy is sometimes distinguished from **homonymy** in that polysemous senses are *semantically related*, whereas homonymous senses (sometimes considered to be separate lexical items) are not. *Bank*, for example, has the *financial institution* sense and the *river bank* sense. The relation between these two senses is usually considered to be one of homonymy, since they are less semantically related than, for example, senses 1 and 2 of *point*. It is worth pointing out, however, that it is difficult to make a hard distinction between polysemy and homonymy (Murphy, 2003).

[3] In context, the gesture described by the first interpretation has a pragmatic meaning similar to the second interpretation since gestural pointing is used to draw attention to something, and a physical book can metonymously stand in for its contents. Insofar as the pragmatic meaning is what is important, the ambiguity of the sentence may not need to be resolved.

But *how* are polysemous senses related? The ways in which senses can be related can be split into two kinds: **regular polysemy**, which follows certain regular patters that can be found across many lexical items and **idiosyncratic polysemy**, where the semantic relation between senses does not follow an established route of connection.[4]

As we will discuss in Section 3.4, the semantic relations that hold between senses that are related by regular polysemy (certain kinds of metaphor, for example) often make it possible to use words in ways that are not lexicalized as senses.

One proposed test is to see whether two "senses" can be joined with a coordinating conjunction. If they can, they may simply be different contextual meanings of the same sense (Deane, 1988). Consider:

(5) The captain and the ship were both pointing.

Does this sentence admit an interpretation where the captain is pointing with his finger (sense 1a) and the ship is sailing close to the wind (sense 2)? If not, this might be evidence that they really are two distinct senses of the word.

It is not always so clear cut, however, whether a sense distinction is present. It would be more difficult to find a similar example that distinguishes between $point_{1c}$ and $point_{1d}$. In the process of writing a dictionary, lexicographers have to make decisions about when to *lump* different uses of a word into one sense or *split* them into multiple senses. This has lead some linguists to prefer a *monosemous* approach in which words are, in general, assumed to have a single meaning (Ruhl, 1989).

2.3. Generality and vagueness

Polysemy is one source of lexical flexibility, but it is far from the only one. Suppose we modify (4) as follows:

(6) When asked about his favorite 19th century American author, Jack pointed *dramatically* to the works of Louise May Alcott *sitting on his desk.*

This sentence no langer ambiguous in the way that (4) was since the context makes it clear that the literal sense of pointing is meant. But many of the details of the situation are still **underspecified**. In truth-theoretic terms, there are aspects of the situation that could change without affecting whether (6) is true. For example, we don't know, the color of the desk. We don't know how old Jack is. It's not specified whether he's pointing to a single *collected works* anthology, or if it's a pile of books. Particularly relevant to the word *point*, we still don't know the exact realization of the gesture Jack made. This is because of the **generality** of the meaning of $point_3$ with respect to some

[4] Deane (1988) calls idiosyncratic polysemy *lexical polysemy*, but here we follow the terminology of A. Blank (2003), who points out that this is somewhat confusing since both kinds of polysemy can be considered lexical.

aspects the gesture. Perhaps *point* implies an extended arm and index finger, but there certainly isn't any conventional (i.e., lexicalized) specification regarding whether it's with the left or the right (or how for the arm is extended or any number of aspects of the motion that gestures may depend on).

That lexical items create this kind of uncertainty may seem like a weakness of language, but in another way of thinking, underspecification *is the meaning* of the word. $Point_3$ picks out situations in which *pointing* is happening. That it doesn't on its own discriminate *between* those situations is exactly what brings them together and gives the word its meaning. Relatedly, lexical underspecification contributes to the flexibility of natural language. Broad lexical interpretations allow a finite vocabulary of words to apply to a wide variety of situations.[5]

Vagueness is a special kind of lexical underspecification in which the borders of the category are not precisely specified by the interpretation. A classic example is gradable adjectives like the word *tall* — there is no conventional height cutoff for when someone (or something) is considered tall. Following the implications of vagueness to their logical conclusions can lead to a paradox known as the Sorites paradox. Since *tall* has vague boarders, it isn't sensitive to very small differences in height — someone who is imperceptibly (say one millimeter) shorter than someone who is tall is still tall. This is what is known as the *tolerance principle* (Wright, 1975). But then we could imagine a long line of people, each one millimeter shorter than the last and by the previous reasoning, we would be committed to saying that someone who is clearly not tall is in fact tall.

There are various ways of dealing with this paradox, many of which involve rejecting or weakening one of the premises (Cobreros et al., 2012). Another is to say that the interpretation of vague terms is probabilistic, in which case the meaning can be represented with a probability distribution that captures uncertainty in where the boundary lies or represents the likelihood that the speaker would use that term in a given situation (Fernandez & Larsson, 2014; Lassiter & Goodman, 2017; Sutton, 2015).[6]

Another aspect of vagueness is that its interpretation is sensitive to what is sometimes called the *comparison class*. What is tall *for a person* may be different from what is tall *for a basketball player*. What is tall for a basketball player is certainly shorter than what is tall *for a skyscraper*. We will discuss context sensitivity and its relationship with semantic variation and change more thoroughly in Section 3.4, but it is necessary to mention here because it turns out that it is a general property of vague terms that they also exhibit this kind of context sensitivity, suggesting that a proper treatment of vague predicate boundaries must somehow take into account the comparison class.

[5] Words can also be used *outside* their conventional interpretations in what is known as *semantic innovation*. This is discussed in more detail in Section 3.4.

[6] See Sutton (2018) for an overview of such approaches.

2.4. Perceptual meaning

Formal semantics is mainly concerned with compositional meaning — given a string of words (perhaps enriched with a certain syntactic structure) and given the meanings of those words, how is the meaning of the string computed? The framing of this question takes lexical meanings for granted. Formal semantic theories often assume that the meaning of predicate-denoting words like *yellow*, *point*, *square* and *democratic* can be modelled as a certain kind of mathematical object. Once the *kind* of mathematical object is decided the semanticist's job is to decide how those meanings interact with each other in compositional expressions, *given their content*. For example, if the formal semantic theory says that predicates denote sets of entities, the semanticist might decide that the meaning of an expression like *yellow square* is the intersection of the meaning of yellow and the meaning of square. It doesn't matter exactly what sets *yellow* and *square* denote, the semanticist only cares that the meaning of adjective-noun phrases of this sort are computed by set intersection.

The fact remains, though, that the meaning of at least certain words *is* closely related to perception.

(7) a. A: Please pick up the yellow square.

b. B: Okay.

If we want to give an account of how this interaction can be successful — how *the yellow square* can successfully refer — we need to explain how B's interpretation of (7a) relates to their perception. In other words, the denotations of *yellow* and *square* must be **grounded** in perception.

Insofar as formal semantics is interested in inference, perceptual meaning cannot be ignored. Certain relations between words can be encoded by **meaning postulates**. We can imagine for example, encoding that *all ravens are birds* by restricting the denotation of *raven* to be a subset of the denotation of *bird*. However, Marconi (1997) argues that a speaker who only had access to lexical meaning encoded in this way could not ever be considered fully *competent* in the meaning of those words. This is especially apparent in situations where *referential competence* is required, as in (7), but it also extends to *inferential competence* in certain cases.

Computational models of meaning have a similar problem, as they often rely on the **distributional hypothesis** (Firth, 1957; Harris, 1954), which says that the meaning of a word can be approximated by the distribution of linguistic contexts in which it appears. A model based on the distributional hypothesis may be able to recognize that *yellow* and *orange* are similar in certain ways (they both appear in proximity to words like *paint*, *pigment*, perhaps even *sunrise* or *flower*), but also have some differences (perhaps *yellow* appears with *canary* and *orange* does not). This sort of model has been criticized for not explaining how symbols in the language are **grounded** in the actual world (Bender et al., 2021; Harnad, 1990; Lücking et al., 2019); such a model

cannot learn meaning representations like the ones humans have since they only relate text to other text, not to perception. A strong version of this argument might claim that even words like *democracy* that don't obviously relate to perception can't, in principle, be truly understood by such a model since meaning in the language system is interconnected and *democracy* is grounded in relation to the rest of the system.

One way of grounding perceptual meaning is to say that the meaning of a predicate-denoting perceptual word like *yellow* is at least in part determined by a **perceptual classifier** — something that computes a function that takes perceptual data as input and produces a category judgment. This approach can be used to ground the meaning of words in computational models, using machine learning classifiers (Schlangen et al., 2016; Silberer et al., 2017). Furthermore, classifier-based word-level meaning representations are subject to compositional analysis, at least in the case of referring expressions (Kennington & Schlangen, 2015).

Such an approach can also be made compatible with formal semantics and information state update models of dialogue (Larsson, 2013, 2020), which we will discuss further in Section 5.1. In brief, the classifier takes the place of a set of entities in the more traditional version of predicate denotation. This better tracks intuitions about how predicate denotations work for actual speakers — it's not that we carry around a list of all the yellow things in the world, but rather than we have the ability to determine if something is yellow, should the need arise. Furthermore, this classifiers as a basis for predicate denotation opens up possibilities for semantic learning based on linguistic and perceptual feedback (Larsson & Bernardy, 2021; Larsson & Cooper, 2021), something particularly important if we are interested in modeling semantic change.

2.5. Cognitive approaches

Wittgenstein (2009) points out that the meaning of a word can almost be described with a complete set of necessary and sufficient conditions. The word *game* is an example. It's very hard to come up with a definition that would cover everything we call a game. We can think of some examples that are *typical* games, but insofar as anything else is a game, it is through a sort of *family resemblance* to the other things we call games. **Prototype theory** holds that cognitive categories themselves are defined not by a set of features but rather in reference to certain ideal *prototypes*. Membership in the class, then, is judged in reference to the prototype (Rosch, 1975).

Some linguists, in turn, have argued that most of what is thought of as polysemy can be explained without making sense distinctions — that for the most part, there are just more and less prototypical realizations of the categories that words refer to (Ruhl, 1989). However, there are reasons to think that sense distinctions are real.

But then how do we explain that a situation described by $point_3$ seems more prototypical of *pointing* than one described by $point_4$? It may be that these two senses are not actually distinct (though the ambiguity of (4) suggests that they are), but others

have suggested that prototypically effects can obtain on two levels — on the *conceptual level*, between instances, as well as on the *semantic level*, between senses (Kamp & Partee, 1995; Tyler & Evans, 2001). Regardless, it seems difficult to make a hard distinction between when two different meanings come from different senses of a word versus when they result from different interpretations of the same sense, draw out by different contexts.[7] This difficulty is reflected in the nested list format often adopted by lexicographers when enumerating senses, which offers readers different options for granularity at which sense distinctions might be made.

Related to prototype theory is **exemplar theory** (e.g., Medin & Schaffer, 1978; Nosofsky, 1984), which is, in some way, an even stronger version of the same idea. In exemplar theory, a concept is still defined in relation to an ideal, but an exemplar is not an abstract idealization, but rather an ideal *member* of the very category. Put another way, exemplars of *of the same kind* as the members of the category, whereas prototypes need not be. Category membership is determined, then, in relation to one or more exemplars of the category. There is some experimental evidence to suggest that both exemplar and prototype-based strategies are involved in classification (H. Blank & Bayer, 2022; Malt, 1989).

While most of this work is not explicitly linguistic, it is of interest to the study of language since there is presumably some connection (via lexical meaning) between the processes linguistic production and interpretation and the representation of conceptual categories. In Chapter 10 we investigate this question using a neural language generation model provided with exemplar and prototype theory-inspired representations of visual categories.

[7] Section Section 3.4 further discusses the use of context for situated meaning making.

3. Sources of meaning

> We die. That may be the meaning of life. But we do language. That may be the measure of our lives.
>
> Toni Morrison
> 1993 Nobel Prize ceremony

The idea of *the lexicon* suggest a big book of words arranged in a list. In the previous chapter, we suggested that the structure of lexical knowledge might be a little more complex than what can be reasonably represented by a list, and that interrelationships between lexical items might have implications for how semantic change happens. In this chapter we'll question the idea that the "book" itself is a monolithic thing. Instead, we have many different sources of lexical meaning, which we draw on in different interactive contexts. Not only that, but context can allow us to draw in sources of meaning from outside the lexicon, or extend the meaning of words beyond what they would normally reach. All of this is important for change because change happens in a particular communicative context, and the meanings that are available in that context are also the ones that have the potential to become part of some lexicon.

We can't point out a language in the material world, and no more can we put our hands on its lexicon. Yet we talk about them as if they are individual entities that have properties, that can come into and go out of existence, and so on. *Ken and Helen both know French*. *Modern English appeared in the 15th century. Latin is a dead language.* But this way of talking about languages belies much of the complexity at the heart of this thesis. Do Helen and Ken have *exactly* the same knowledge of the French language? Of course not! Is the English of the 15th century the same as what is spoken around the world today? No! — it was very different, as are the many varieties of English that are spoken contemporaneously. The reason we call Latin a *dead* language is not because it doesn't have any speakers (indeed, some people do still learn a version of Latin in school), but because it doesn't have a *community of speakers* who use it, who breath life into it, whose communicative needs it serves and with whom it changes.[1]

The perspective on language adopted throughout Part II is that it exists *through people* and *in communities*, available as a resource for interaction. In dialogue, speakers

[1] Latin *is* spoken for ceremonial purposes and is even used in official documents in the Vatican but isn't generally used in *interactive* communicative contexts which, as we will see, are particularly important for engendering linguistic change.

generally assume that the language they are speaking is *common knowledge* among the interlocutors — that the meanings of words, how to construct and interpret utterances, etc. are shared by everyone, that everyone *knows* that they are shared by everyone, that everyone *knows that everyone knows* they are shared, and so on. This *and so on* makes things tricky. Practically, how do we get to a point where we don't need to go through an infinite regress of social deduction just to be assured that communication is possible? **Common ground** (Lewis, 1969; Stalnaker, 2002) is a model of common knowledge that solves this problem. We say that something is common ground for a group of people if there is a **shared basis** that indicates that it is true. A basis b is shared for a group G when:

1. everyone in G has information that b holds, and

2. b indicates to everyone in G that (1) is the case.

Clark (1996) identifies two kinds of common ground. **Communal common ground** is shared based on joint membership in a community. At a geology conference, the fact that *limestone is a sedimentary rock* may be considered common ground, based on joint membership in a community of geologists. Similarly the meaning of those and other geological terms may be taken to be common ground: the meaning of *limestone* is part of the lexicon for the community of geologists because everyone in the community knows what *limestone* is and because being part of the geology community is generally understood to entail knowing the meaning of *limestone*.

Personal common ground is grounded on a *perceptual* or *actional* basis.[2] Such a basis is shared because of their joint attention on some event or *situation*. Imagine we are at a baseball game. We're both intently watching a crucial at-bat. The batter hits the ball. It's a home run! Now it's common ground among us that the batter has hit a home run. This is the case on the basis of the situation just described, since you and I both have perceptual access to the situation (we were both paying attention) and since our joint attention is itself evident in that very situation we both have access to.

It is important to emphasize that common ground is a subjective notion — it depends on what an individual *takes to be* common ground.[3] It is not uncommon, for example, for dialogue participants to find that their construal of what has been said in a conversation is misaligned and in need of repair. What someone takes to be common ground in an interaction depends on the requirements of the interaction. For example, speakers may at times be rather loose with assuming that certain lexical items are common ground, trusting that misunderstandings will be identified and repaired.

[2] The main difference between actional and perceptual bases is that actional bases are events that are brought about by the joint action of the participants, usually by way of exploiting pre-existing common ground. For Clark (1996), actional bases are key to explaining how dialogue works

[3] In some cases, we may wish to consider what *all speakers take to be* common ground as a corollary to the objective notion. In other cases it makes sense to take a more explicitly agent-centric notion.

As we will discuss in the following two sections, lexical meaning can be grounded in both both communal and personal common ground, a fact that is particularly relevant to lexical semantic change.

3.1. Communal lexicons

Any community can serve as a basis for communal common ground, which would suggest that any community of language users could have its own lexicon. Indeed, this is the idea behind the notion of what Gumperz (1972) terms the **speech-community**, which is

> any human aggregate characterized by regular and frequent interaction by means of a shared body of verbal signs and set off from similar aggregates by significant differences in language usage. (Giglioli, 1972, p. 219)

A *language*, then, can be defined as the accumulation of the linguistic norms and practices grounded in a particular speech community.

Of course, this is quite a different notion of language from the one that is in common usage. We don't usually think of geologists and fire fighters as speaking *different languages*. But in a sense they do, especially when speaking with one another about topics of special importance to their respective communities. But this also reveals that there is a hierarchical relationship between speech communities. While geologists may have special terminology in the domain of geology, they default to what we will call a *macro-language* (English, for example) where the norms of the geologist community have no special bearing. There may of course be intermediary communities as well. Perhaps there are conventions among natural scientists, a designation which includes geologists.

Even this picture is a bit simplistic. French geologists mainly speak French. The ways in which their speech differs from the macro-language may in some ways be similar to how English geologists' speech differs from English (perhaps based in joint membership in an international community of geologists) and may in some ways be particular to the community of French geologists.

We usually think of macro-languages as at the top of this hierarchy, but some macro-languages are at least partially mutually intelligible. In these cases, it's not necessarily that there is a community that encompasses both, but rather that the norms of the two respective communities are close enough that certain linguistic conventions can be considered common ground for the purposes of conversation.

3.2. Interpersonal lexicons

Communal lexicons are probably what we usually think of when we think of lexical knowledge, but lexical meaning can also come from personal common ground. These interpersonal lexicons generally don't constitute a whole language, but rather, as with in a specialized community lexicon, supplement another more general linguistic resource that the participants also share.

Special idioms, nicknames, expressions of affection and more are often shared between families, close friends and romantic partners (Hopper et al., 1981). Not only do interpersonal lexicons facilitate communication (including possibly covert communication), they serve to express solidarity and closeness among intimates (Bell & Healey, 1992).

Interpersonal lexical resources are not limited to novel words, though. When someone uses a word in a particular way, their dialogue partner might take note and expect that sort of usage in the future. This is particularly true if the intended meaning wasn't obvious at first and required extra reasoning or especially repair (G. J. Mills & Healey, 2006). In general, when speakers have to coordinate on the meaning of a lexical item, the coordinated meanings may carry over to future dialogues, i.e., as part of an interpersonal lexicon.

3.3. Semantic coordination

So where do these interpersonal lexical resources come from? How do we go from not sharing any partner-specific meanings with someone to having them? The answer is semantic coordination. To understand how that works, we first need to discuss how personal common ground is built up over the course of an interaction.

The Collaborative Model (Clark & Schaefer, 1986; Clark & Wilkes-Gibbs, 1986) is a theory of conversation that explains, from a psycholinguistic point of view, how speakers collaborate through **communicative grounding**. The model describes a hierarchy of grounding levels that dialogue participants must move through in order to reach (and maintain) mutual understanding. To coordinate effectively, participants must tailor their actions to what has been grounded so far while also providing evidence (positive and negative) of grounding to facilitate their interlocutors doing the same. When evidence of understanding is demonstrated, participants can consider what was said to be common ground.

Communicative grounding is subject to *opportunistic closure*, meaning that grounding at higher levels are taken as evidence of grounding at lower levels. Closure can also work compositionally — if B gives evidence they understood A's utterance, that can be taken as evidence that they understood the words it was composed of in the way they were meant.

Such an understanding can be achieved even when the speaker uses a word outside of

	Level	Speaker (A) and addressee (B) actions
1	contact	A and B pay attention to each other
2	perception	B perceives the signal produced by A
3	understanding	B understands what A intends to convey
4	uptake	B accepts/reacts to A's proposal

Table 3.1.: Levels of communicative grounding (Fernandez, 2014)

its normal semantic range or in a way that the addressee is not familiar with. We discuss the ways in which extra-linguistic context interacts with word meaning in more detail in Section 3.4. In these situations, the addressee may shift their understanding of the word for the purposes of the conversation, what we call **implicit semantic coordination**.

Implicit coordination has been studied in experimental settings where participants develop **lexical pacts** (Brennan & Clark, 1996). These temporary, flexible conventions emerge as a consequence of successful interaction and persist as a resource as the interaction continues (or even in future interactions). Such conventions are not limited to isolated lexical items. G. Mills and Healey (2008) observed that participants asked to perform a collaborative maze-solving task would create a **conceptual pact** — a unified semantic model of how to refer to locations in the maze.

On the other hand, if they cannot figure out the intended meaning or wish to raise a meta-linguistic objection to that use of the word, they may initiate a **word meaning negotiations** (WMN) (Myrendal, 2015). Here, *negotiation* is meant in the sense that a group of friends might negotiate paying for a restaurant bill — WMNs are collaborative on the level of interaction, with the implicit goal of reaching a mutually agreed upon result. WMNs, *can* of course be adversarial in terms of the outcome, but they need not be and, as in any dialogue, some level of cooperation is needed to coordinate the interaction.

It's difficult to characterize exactly how common WMNs are in every-day conversation, since they take many surface-level forms, making them difficult to search for exhaustively in a corpus. Myrendal (2015) studied word meaning negotiation in Swedish discussion forums, collecting a corpus of exchanges by searching for the phrases like *Vad menar du med // what do you mean by*. In Chapter 7 we used variations of the same phrase in English to find WMNs.

Corrective feedback is another form of **explicit semantic coordination**. Consider these examples of adult-child speech (Larsson & Cooper, 2009):

(8) a. A: That's a nice bear.

b. B: Yes it's a nice panda.

(9) a. A: Mommy, where my plate?

b. B: You mean your saucer

Corrective feedback can be seen as a special case of WMN where there is an implied epistemic inequality between the participants.

Both implicit and explicit coordination have the potential to affect affect lexical resources beyond the current dialogue. After an interaction speakers may remember an unusual way they used a word or remember a word meaning that was negotiated. These newly coordinated meanings can be made available for use in future dialogues based on interpersonal common ground. Under the right circumstances, a speaker may also take the dialogue as evidence that new lexical information holds for a particular community, resulting in community-level change (for the speaker).

3.4. Meaning in context

Chapter 2 introduced the idea that words have a certain amount of semantic flexibility. A single word often has multiple related senses (polysemy) and meanings can describe a variety of different situations (generality). Although there are different ideas about where these sources of flexibility reside and how they interact with each other, it is clear that in-context meaning is dramatically *less* underspecified than lexical meaning in the abstract.

Abstract or lexical meaning is sometimes called **meaning potential**, which is considered to be something of an entirely different kind from in-context meaning. Norén and Linell (2007) describes meaning potentials as semantic *affordances*. They are something that "afford language users with semantic potentialities to be exploited in situated use." In Gibson (1966)'s theory of perception, an affordance is what the environment provides as an interactive possibility. Similar to the way that a cup with a handle may offer *grasping* as an affordance.

Lexical meanings (or meaning potentials) combine with linguistic and extralinguistic context to produce **situated meanings**. One way this happens is that context can narrow the generality of interpretation of a word. Consider a word like *eat*. Eating a soup and eating a sandwich are quite different activities, yet the sentential context can make clear what is meant. Such narrowings can persist even outside of a disambiguating sentential context by relying on discourse context.

(10) a. A: I'm eating soup.

b. ...

c. A: Leave me alone, I'm still eating!

To see how extralinguistic context can affect situated meaning, consider the interpretation of definite referring expressions, for example.

(11) a. *That red car over there* is mine.

b. *The red one* is mine.

c. *The head of department* will be at lunch tomorrow.

d. I'm going to *the library* later.

Expressions like these often require some additional perceptual common ground to successfully refer. We might imagine 11a being used in a situation where the interlocutors have joint attention on some visual scene. If the visual and communicative context makes it clear that the speaker is referring to a car, the 11c might do the trick Communal common ground could be required to interpret 11c, and 11d might draw on some shared relevance ordering on libraries.

Different modes of interaction are available depending on the *genre*. Situated meaning can also be affected by the **genre** of communication (Bakhtin, 1987). Genre is something similar to Wittgenstein (2009)'s *language games*. They are different modes of interaction that serve as communicative resources in different situations. Consider, for example, a waiter at a diner speaking to a colleague. The waiter may metanonomyously refer to a patron who ordered a ham sandwich as *the ham sandwich*. But this same referring expression wouldn't be available to another patron or to the waiter talking to someone else who isn't working there (A. Blank, 2003).

3.4.1. Pragmatics

Like semantics, *pragmatics* refers to both a subfield of linguistics and the collection of linguistic phenomena that the field studies. A classic way to make the distinction between semantics and pragmatics is to say that pragmatics has to do with **speaker meaning**, which is different from meaning in the abstract.

Of course, this distinction rests on the potentially precarious assumption that there *is* something like meaning in the abstract. Nevertheless, it is the case that there are a number of kinds of communicative situations where the interpretation of the speaker meaning is somehow based on a prior interpretation of the "literal" semantic meaning. Consider this classic example:

(12) a. A: Can you pass the salt?

b. B: Sure thing. *[passes the salt]*

We could imagine situations where (12a) is uttered as a genuine question, but typically one would interpret it as a request for the salt. Searle (1975) refers to this kind of utterance as an indirect speech act because the primary intention (requesting) is performed by performing an action with a different *literal* interpretation. B confirms the indirect interpretation, with their reply (12b), which grounds A's utterance as a request by responding to it as such.

So how does B know that (12a) is a request, given that it is literally a question? One story goes like this: B first processes the utterance as a question, using their "normal" faculty of semantic interpretation. From this interpretation, B reasons about why A might have asked this question, given that there is no reason for uncertainty about B's salt-passing ability. B realizes that (1) their ability to pass the salt is a prerequisite for actually passing it, (2) it would be considered impolite for A to make their request as an imperative (i.e., *Pass the salt.*), and perhaps (3) they are in a situation where it would be normal for A to want B to pass the salt. From here, B concludes that A must have uttered (12a) as an indirect way of requesting that they pass the salt.

This follows the classical *Gricean* account of *conversational implicature*, wherein the pragmatic meaning is derivative of the **cooperative principle** of communication, which says that in general, cooperative speakers try to be informative, truthful, relevant, and clear (Grice, 1975). When it would seem that a speaker is in violation of or *flouting* one of these maxims, a cooperative listener will go searching for some alternative interpretation under which their interlocutor is adhering to them.

Now, this is a rather long and involved story to tell about what would seem to be a rather simple and (crucially) routine interaction in (12). Indeed, Grice (1975) might instead explain this example in terms of *conventional implicature*. On this account, the meaning may have at one point been calculated as previously described, but it has since become *conventionalized*, obviating the need to perform the pragmatic inference.

Herein we see a clear connection between pragmatics and semantic change. If the phrasing *can you pass...* (or more generally *can you...*) becomes conventionally associated with the act of requesting, that would mean by definition that some lexicalization of the requesting function has occurred i.e., that the meaning has changed. It could be the case, as Morgan (1978) argues, that speakers can make a distinction between literal and indirect uses of a linguistic unit, even when the indirect function is conventionalized. This further fuzzies the border between semantics and pragmatics. If studying semantic change means studying changes in what thing speakers can *use words to mean*, we really can't avoid pragmatics.

Pragmatic inference often seems to rely on some shared assumptions about what follows from what.[4] In argumentation theory, an *enthymeme* is an argument in which one or more of the premises are not explicitly stated. A **topos** is that which supplies the missing premise (J.-C. Anscombre, 1995). Put another way, it is a function from enthymemes to full arguments. J. C. Anscombre and Ducrot (1983) observes that even language that is not *prima facie* argumentative still has a structure that can be analyzed as argumentation. Often this means that there are implicit (enthymematic) steps in a discourse. Breitholtz (2020) develops a theory in which topoi are a resource that can be drawn on in in linguistic interaction — part of the common ground in the same way that lexical items are. Indeed, we discuss topoi as having something like lexical

[4] Although we do not discuss it in the thesis, relevance theory is a theory of pragmatics that centers ethememetic reasoning (Sperber & Wilson, 2001).

meaning in Chapter 11, but in that work we focus not so much on their role in pragmatic inference, but rather their role in implicitly communicating social information about the speaker — their *social meaning*.

3.4.2. Social meaning

Social meaning is similar to pragmatic meaning in that it goes beyond the literal interpretation of an utterance. It is also similar in that recovering the social meaning usually involves considering the communicative context. In the case of social meaning, however, the relevant communicative context is usually not so much on the level of interaction, but rather the *social context* in which the interaction takes place. Very often the social meaning of an utterance communicates something about how the speaker themself relates to the social context — for example, by revealing something about their social position or ideology.

Eckert (2019) classifies the progression of sociolinguistics as a field in three *waves*, each of which have a different take on how social meaning functions. Early sociolinguistic work largely sought to describe regional linguistic variation and variation among macro-social categories of speaker (i.e., based on age, gender, class, etc.). First-wave sociolinguistics already acknowledged the potential for variables to carry social meaning. Labov (1963), for example, acknowledged that phonetic changes in the speech of certain non-native residents of Martha's Vineyard may have something to do with a desire to be associated with the working-class resident population as opposed to the upper class summer visitors. However, he also noted that the diphthong centralization he observed was not consciously salient to speakers he interviewed, suggesting that the changes were largely a matter of subconscious identification with a particular ideology. The idea that speakers use sociolinguistic variables as a way to (more or less) intentionally *construct* a social identity is the hallmark of second-wave sociolinguistics. However, it is in third-wave sociolinguistics that the variable moves from being seen as a theoretical tool to something with a "social and cognitive reality" (Campbell-Kibler, 2010).

This social and cognitive reality can be seen in the concept of the **persona**, which is a sort of stereotypical *kind of person*, which is a common ground resource (in the sense of Clark (1996)) that speakers can draw on to construct a social identity. Personae are not real people, but they do have a kind of social reality, give their common ground status as social reference points. People can draw on personae to construct an identity by way of social signals that are indexically associated with the personae in what Eckert (2008) calls the *indexical field*.

Importantly, social signals are not exclusively linguistic. They can include all kinds of things, including dress, body language, and so on. Furthermore, linguistic social signals are not limited to *how* something is said (although this has tended to be the focus of sociolinguistics), but also *what* is said. In Chapter 11 for example, where we develop

a probabilistic model of social signalling based on Eckert (2008)'s indexical field, we use topoi as the case study. A topos is evoked by ethememetic speech, something which is more related to content than style.

4. Semantic variation and change

> Gretchen, stop trying to make fetch happen. It's not going to happen!
>
> Regina, *Mean Girls* (2004)

Linguistic variation is an important concept in sociolinguistics, which is mainly concerned with variation across social groups.[1] On a structural level, change is just variation over time. Many of the same corpus-based methods used to study change can also be used to study synchronic variation (see Section 5.2.2). But it is also worth considering variation and change separately since each have a role to play in explaining the other. Variation leads to change as one community of speakers adopts ways of speaking from another community.

Change leads to variation as communities diverge in their language. Variation and change also have different social implications and different relationships to lexical meaning. The next two sections will draw out some distinctions that are made in *types of variation* and *types of change* we might observe. Some of these categories apply to both variation and change, while others are specific to one or the other.

4.1. Types of variation

One of these distinctions is related to the different factors driving variation. Sometimes, variation stems from the fact that speakers draw on different bases for linguistic common ground in different situations. An expert in some field may speak differently when talking to peers in her community than she would talking to non-experts who nevertheless speak the same macro-language. A teenager writing a message in a video game forum can relay a story about something that happened in the game differently from how they would tell the same story to their parents at the dinner table. This kind of variation is an example of **code-switching**, which can also involve multiple macro-languages.

However, it is somewhat naive to think that sociolinguistic variation is always rooted in differences in common ground. A politician from Skåne (in southern Sweden)

[1] There are also, of course, differences in language use across individuals which do not present on any particular axis of social identity. These differences usually fall under of *linguistic style* and, while they have received some attention (e.g., Johnstone, 1996), are not a main focus of sociolinguistic inquiry.

may use different vowel articulation, speech patterns, and lexical items depending on whether they are speaking to rural constituents or meeting with business leaders in Malmö. This code-switching may have nothing to do with common ground *per se* — both registers would probably be just as well understood in both situations, but rather the choice of register is explained by how the politician wants to be perceived — what *persona* they want to *project* (see Section 3.4.2). Different ways of speaking carry different *social meaning*

That variation is not always a result of different common ground is also evident in the fact that the social categories along which sociolinguists study variation are not always speech communities. Indeed, classical sociolinguistics is much more often concerned with variation across macro-social categories like gender, race, class, and even sexual orientation (Labov, 1963; Podesva, 2007). Eckert (2008) attempts to introduce further nuance, arguing for an approach in which linguistic variants are not mere markers of social identity, but a collection of signs in a complex semiotic system through which individuals may project their social identity in relation to social archetypes or *personae*. The personae themselves stand in relation to local and macro-social categories, but are not wholly constituted by them.

These two sources of variation are, of course, not easily separable. Variants that are understood across communities and whose primary function is to mark community membership or project personae may, with time, evolve into something that requires a certain common ground to understand. Similarly, variants that are only understood within a certain community may come to be understood more broadly while retaining their status as a social signal.

Type 1 and type 2 variation Another distinction has to do with the perspective we take on variation as observers of the system. We have already spoken loosely of *variants* as the units along which variation is observed, but this needs to be made more precise. To identify something as a *variant* means that there is a difference with respect to some reference point that stays the same. There is the *variation* and there is the thing it is a *variation of*. So when we talk about linguistic variation, what is it that changes and what stays the same?

In classical linguistic theory, language is thought of as a hierarchical system of semiotic relations where signs on a lower layer signify meanings on a higher layer (Fig. 4.1). Phonemes are signified by particular speech sounds, morphemes by particular phonemes, words are made up of morphemes, syntactic structures are determined by the grammatical categories of a string of words, and the meaning of units of speech larger than words (sentences, for example) is determined in part by the meaning of words that make it up and the syntactic structure they produce.

Without getting too into the weeds about what the appropriate unit of analysis is at each of these levels (or indeed where the lines between levels should be drawn, if at all), the classic assumption is that units at one level of analysis (sometimes alone

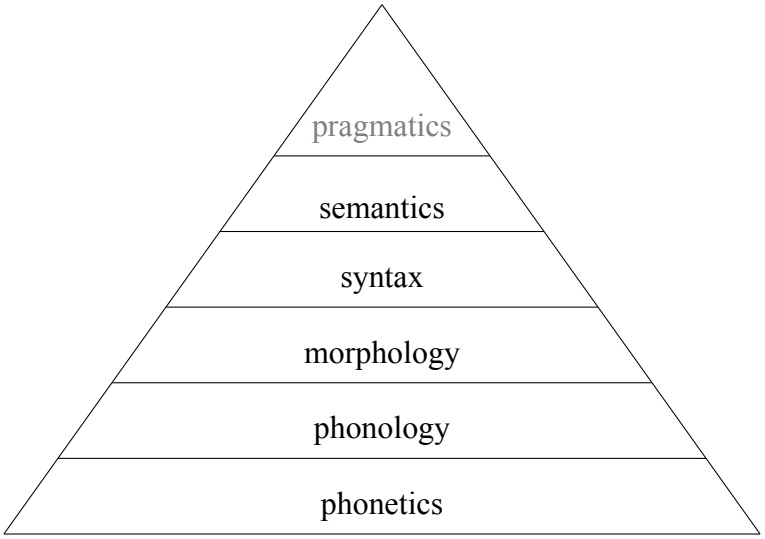


Figure 4.1.: The classical linguistic hierarchy. Forms at lower levels of the hierarchy combine to create meanings at higher levels. Pragmatics is somewhat different because pragmatic meaning depends on extra-linguistic context as well as semantic meaning, which might be why it is often left out of the classical linguistic hierarchy.

(a) Type 1 variation

(b) Type 2 variation

Figure 4.2.: Two types of linguistic variation. In type 1 variation, different forms variously signify the same meaning (for example, in communities A and B). In type 2 variation, the same form variously signifies different meanings.

and sometimes in combination with other units), which we will call *forms* stand in a signification relation with units at higher levels, which we will call *meanings*. Anttila (2004) points out that this leaves us with two kinds of variation in the relationship between forms and meanings (Fig. 4.2).[2] .

Sociolinguistics is almost always concerned with type 1 variation. Perhaps the most clear-cut example is sociophonetics, which studies variation in the relationship between speech sounds and phonemes. A *variant* in sociophonetics is a phoneme that can be signified by multiple different speech sounds. Think back (Section 3.4.2) to the centralized diphthongs that Labov (1963) observed in Martha's Vineyard. Those vowel sounds were phonetically different across groups, but interpreted phonologically

[2]This hierarchical model persists in spite of many arguments and counter-examples against it. In fact, there are interactions at many levels of the hierarchy, not only adjacent ones and not only in one direction (Cann et al., 2000). Indeed, pragmatics is not only about determining *why* something was said, but also *what* was said (Korta & Perry, 2008). If our linguistic theory computes meanings one level at a time from bottom to top, then it is going to run into problems.

to mean the same thing.

It can of course also be the case that the same speech sound is used to mean two different phonemes (in two different regional accents, for example). But this isn't really considered sociolinguistic variation. Why? It has to do with social meaning. When someone's way of saying M_1 is F_2 instead of F_1, that's a salient difference that we can ascribe meaning to. When someone's way of saying M_2 is F_1 (while someone else might use F_1 to mean M_1), that relationship is less salient because ambiguity is ubiquitous in language — just because the speaker used F_1 to mean M_2 doesn't mean they don't *also* use it to mean M_1.

This puts us in an awkward position if we want to investigate semantic change from a sociolinguistic perspective since semantics falls at the top of the linguistic hierarchy.[3] It's (relatively) easy to search for a particular word and see how its contexts of use vary, whether across time or some other factor. It's much more difficult to search for *contexts in which someone might want to express a particular meaning* and see what words they used. In fact, this is exactly what Hasan's (2009) work on semantic variation does. For example, in a corpus of child-directed speech, she investigates the different ways mothers have of issuing a command to their children (Hasan, 1989).

In lexicography, there is a parallel distinction between semasiological approaches (organized around form) and onomasiological approaches (organized around meaning). Historical linguists have adopted these terms: type 1 lexical variation over time is called **onomasiological** change, and type 2 lexical variation over time is called **semasiological** change. Historical linguistics mostly considers semasiological change for much the same reason (think of the classic examples like *awesome* and *gay* that we introduced in Chapter 1).

This thesis, likewise, mostly considers type 2 variation. Certainly in Chapter 13, when we use computational methods to quantify how much particular words change across corpora, we are working with type 2 variation. But type 1 variation is present in questions of semantic change as well. In Chapter 7 we study explicit conversations about word meaning (a phenomenon which we will introduce in Section 3.3). The word in question is very often a word that everyone already knows *some* meaning for, but the meaning in the particular context is unfamiliar (or even disagreeable) to someone.

4.2. Types of change

There are many different ways of categorizing lexical change, and different ideas about what should go into those taxonomies. Most change typologies take a semasiological perspective, but there are also onomasiological categories. Perhaps the most important of which is *lexical replacement*, which is when a word is replace by another word in a particular situation. We'll talk about two semasiological typologies here. The first

[3] See Hasan (1989) for discussion of how pragmatics fits (or doesn't) into this picture.

Change type	Meaning
Novel word	A new word form (and associated sense, possibly also novel)
Novel sense	A new sense for an existing word
Word death	A word form that goes out of use
Sense death	A sense of a word that goes out of use
Sense split	What was one sense of a word is now considered two senses
Sense join	Two senses are no longer distinguished

Table 4.1.: Sense-structure classification of lexical semantic change. Table adapted from Tahmasebi et al. (2021), which also includes an analysis of how these types are construed in the field of computational change detection.

is to classify according changes in the lexical **structure** — that is, with respect to the inventory of words, senses, and relations between the two (Table 4.1). Of course these distinctions assume a list-of-senses model of lexical meaning, which as we discussed in Section 2.2, is not without its problems.

Another way of classifying lexical change is in terms of the **explanations** for why the change happened or what might have made it possible (Table 4.2). Again, these explanations assume sense distinctions since they often are described in terms of a relationship between an old sense and a new sense. For example, the word *mouse* has a new (20th century) sense, which refers to a computer mouse. The new sense has a metaphorical relationship to the animal sense.

At the risk of just-so story, we can easily imagine that the old sense might have licensed innovative uses in contexts covered by the new sense before the change was lexicalized. For example, the change in the meaning of *mouse* (by adding a sense) came about by the conventionalization of metaphorical extensions of *mouse* to a new artifact (the computer input device) that played on visual similarity. These relations suggest a synergy between polysemy, lexical innovation, and change, although it's important to point out that there is no one-to-one correspondence between relations between senses and types of change (A. Blank, 2003).

Change type	Meaning
Metonymy	A new sense is related by metaphorical comparison
Metaphor	A new sense is related by metaphorical comparison
Co-hyponymous transfer	A new sense denotes something that shares a hypernym with the existing sense (the new sense and the old sense are co-hyponymous)
Semantic extension	The word now applies to more related situations (also called semantic *broadening*)
Semantic restriction	The word now applies to a more restricted set of situations (also called semantic *narrowing*)
Antiphrasis	A word gains a sense which is opposite to an existing sense (for example through innuendo or humorous innovation)

Table 4.2.: Explanatory classification of lexical semantic change. Abbreviated from A. Blank (2003), which includes a comparison with synchronic sense relations.

5. Methodology

> ... in that Empire, the Art of Cartography achieved such Perfection that an entire City was occupied with the Map of a Single Province, and a Province was required to display the Map of the whole Empire.
>
> In time, even these Vast Maps ceased to satisfy, so the Cartographers' Guilds unfurled a new Map of the Empire the Size of the Empire, each point overlaying exactly what it mapped.
>
> Later, Generations, less addicted to the Act of Mapping, understood that this Immodest Map was Useless — they irreverently surrendered it to the ravages of Sun and Snow. In the Western Deserts, tattered Ruins of the Map remain, home to animals and vagabonds; these are the Country's last vestiges of the Geographic Disciplines. (1658)
>
> ---
>
> *on Scientific Rigor*
> Jorge Luis Borges
> trans. Noah Mease

Part II uses a variety of different methods, including both formal and computational models of natural language semantics. The reasons for this methodological diversity are twofold. First, the nature of short-term semantic change places us in the liminal space between interaction, interpersonal relationships, and speech communities. We need to, on the one hand, investigate semantic plasticity in the context of concrete interactions and, on the other hand, investigate change over short time periods abstracted over communities of practice.

Second (and related to the first) is because of the scientific goals of the thesis. Some studies in the thesis test hypotheses about how change and variation takes place in communities (e.g., Chapters 12 and 13) or how metalinguistic communication can be implemented in neural models (e.g., Chapter 10). Other parts are more focused on developing theoretical models of semantic meaning and interaction that allow for change (e.g., Chapters 7 to 9 and 11).

Every study in the compilation uses some kind of *model*. But what *is* a model? Perhaps the most prototypical models are ones that mimic the workings of something they are *modeling*. Think of an environmental geologist who builds a physical model of a riverbed to asses the risk of flooding. They might try some experiment like building or removing a dam and extrapolate the effects observed in the model to what would happen if the analogous actions were carried out in on the real-life stream. A com-

putational model can play a similar role. Such a model would attempt to extract and quantify key aspects of the stream bed, many of which the physical model also captures (rates of flow, soil permeability, etc.). The model can then be used to make predictions as a function of those parameters.

But *models* in computational linguistics and NLP don't always have the same relationship to the *modeled*. In fact, it is not always very clear what, if anything, is being modeled. The mechanisms of language processing in a large language model like BERT are not at all clear, and there is no reason to think that, on a low level, it is doing anything at all analogous to what humans do when they process language. In this way, the relationship between such a model and human language processing is more like the relationship between a bird and a quadrocopter drone. Sure, they do *some* of the same things, but they do them by entirely different means. But that doesn't mean that models like BERT are useless for studying natural language. Just as one might use a drone to get closer to the habitat of birds, to measure the flow of wind over a cliff, or capture what a rabbit in a field looks like from high in the sky, machine learning models can, through careful analysis, be useful for investigating relationships in the linguistic environment in which human language use takes place.

Formal models, on the other hand, are usually *descriptive*, attempting to mirror real-world processes in a way that elucidates some theoretically important aspect of them. Computational models *can* be descriptive, but more often they are primarily designed to be *predictive*. If a computational model is good at predicting the outputs of a certain real-world process from its inputs, we might conclude that the real world process resembles the mechanism of the model in certain ways (for example, we might conclude something about the computational complexity of the real-world process or something more detailed by *probing* the model for relations between its internal mechanisms). Alternatively, we might just be interested in the practical applications of a computational model. If certain kind of model performs well enough over the long-term to be useful in applications, this might provide a different kind of evidence that it "gets something right" about how the real-world process works.

As the goals of the thesis would suggest, the models we use have a variety of different methodological roles to play. A useful question to ask is what *level of description* the model is targeting. The neural machine learning models described in Section 5.2.1 are *inspired by* biological neural networks, but no one would claim that neural language models process linguistic input in the same way humans do at the level of the neuron. If these models do mirror human language processing — and there is at least some evidence that they do, to some degree (Bhattasali & Resnik, 2021) — then the relationship between material parts of the brain and parametrized functions in the model is much more abstract than a one-to-one mapping between the two. Formal models mostly don't attempt to model psychological processes at all (although there are exceptions). Rather formal models tend to focus on structural phenomena that emerge from psychological processes and the causal factors (including mathematical, and log-

ical factors) that govern them.[1]

Formal modeling can have a symbiotic relationship with more empirical approaches. Formal models provide a language to pose hypotheses and can provide inspiration structuring computational and statistical models. Computational models, on the other hand, can provide a way to test hypotheses on large amounts of real-world data. This thesis has a small role to play in that relationship, but it is a great pleasure (and sometimes even scientifically productive) to play on both sides of it.

5.1. Formal methods

Understood broadly, formal methods are ways of making a theory precise, usually by the use of an abstract symbolic system borrowed from logic or mathematics. This is often what is meant to *formalize* a theory — the theory is translated into logic or math, which aids in precisely formulating (and often generating) hypotheses that can be tested empirically. Under this characterization, formal methods have long been employed in linguistics, though exactly what is meant by *formal* can vary by sub-discipline and across research traditions.

This section describes the formal methods adopted in the thesis. Not all of the work in this thesis uses formalization, but Chapters 7 to 9 and 11 all include at least some formalization in Type Theory with Records (TTR) and it's probabilistic counterpart ProbTTR.

To give context and motivate this choice of system, this chapter includes a broader introduction to the methodology of the formal semantics tradition sometimes termed *logical grammar*. Montague semantics, named after mathematician-turned-linguist Richard Montague. Montague's work on natural language semantics (1970, 1973), is certainly the most influential progenitor of the logical grammar approach. His paper *English as a formal language* 1970 defied the conventional wisdom that semantics was not a candidate for formalization. As Barbara Partee (1973) described the situation,

> Logicians seem to have felt that natural languages were too unsystematic, too full of vagueness and ambiguity, to be amenable to their rigorous methods, or if susceptible to formal treatment, only at great cost. Linguists on the other hand, emphasize their own concern for psychological reality, and the logicians' lack of it, in eschewing the logicians' approach. (p. 509 B. Partee, 1973)

These tensions are still present in semantics today, often manifesting as disagreement about what work counts as *formal*, what counts *natural* language, and what the goals of formalization in semantics should be. The overall trend, however, is to provide formal descriptions of more and more of those "unsystematic" features that motivated

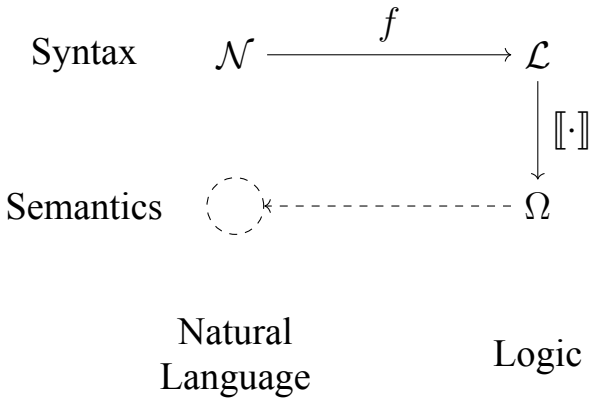
[1] See B. H. Partee (1979) and Teichman (n.d.) for further discussion.

the initial skepticism of logicians. This often means reaching for logical systems with more expressive power than what was used in Montague's initial work.

So what does it mean to treat a natural language *as a formal language*? And how does this result in a theory of meaning? A formal language is a mathematical object. It starts with a set of symbols, Σ. These symbols might be words or letters or even audio symbols like phonemes — anything at all really, as long as they can be reproduced, distinguished from one another, and put into sequence. A finite sequence of symbols from Σ (possibly including repeats) is called a Σ-**string**. A **formal language**, $\mathcal{L}$ *over* Σ is a particular subset of all the possible Σ-strings. A formal language is usually defined in terms of a rules that describe admissible strings — that is, strings that are part of $\mathcal{L}$.

Logics have a *syntax* and a *semantics*. The syntax is given by a formal language. The semantics lies in some other "realm" Ω, which is often *set theoretic* — made up of mathematical sets, including especially functions. A **logic** connects a syntax and semantics by means of an interpretation function, $[\![\cdot]\!] : \mathcal{L} \to \Omega$.

The final piece of the puzzle is a translation function, $f : \mathcal{N} \to \mathcal{L}$ which maps expressions in the natural language to strings of the formal language.[2] In practice, $\mathcal{N}$ is a **fragment** of some natural language, often defined in terms of a lexicon of words of various syntactic types and a set of rules for forming expressions. This means that elements of $\mathcal{N}$ are not just strings, but come along with a syntactic structure that can be used in the definition of f. Together, these three collections of objects and two functions that map between them give us the following (albeit simplistic) schema of a model theoretic logical grammar.



A potentially misleading aspect of this schema is that while there is a vertical divide between the syntactic and semantic of both logic and natural language, it is really the relationship between *all three* of these realms ($\mathcal{N}$, $\mathcal{L}$, and Ω) that comprises the semantic theory. By composing f and $[\![\cdot]\!]$, the theory proposes a mapping from the

[2] One complication to this picture worth noting already is that f is not always, strictly speaking, a function, since many theories account for *semantic ambiguities* — situations where a single syntactic form can have multiple semantic interpretations. This is the usual analysis given to sentences like *everyone pointed at someone*, where the two quantifiers create a *scope ambiguity*; is there one specific person who everyone pointed at or did everyone point at various (possibly different) people? Although the grammatical structure is the same, this sentence can be translated to to different predicate logic formulas which capture the two different readings.

meaning of natural language expressions to the semantic space of the logic. This means that f is just as much a part of the semantic theory as $[\![\cdot]\!]$; in practice, it is often the more consequential part.[3] Putting these mappings together, given a natural language expression $e \in \mathcal{N}$, there is an object $o \in \Omega$ such that $[\![f(e)]\!] = o$. The semanticist who proposes this theory asserts some modeling relationship (the dashed line) between the o and the *actual meaning* of e.

To make all of this a little more explicit let's consider a very simple example of a formal semantic theory for a fragment of English including the word *and*.

$\mathcal{N}$	$\mathcal{L}$
• *John is pointing* $\in \mathcal{N}$	• $p \in \mathcal{L}$
• *Kim is pointing* $\in \mathcal{N}$	• $q \in \mathcal{L}$
• $s_1, s_2 \in \mathcal{N} \Rightarrow [s_1 \textit{ and } s_2] \in \mathcal{N}$	• $\varphi, \psi \in \mathcal{L} \Rightarrow (\varphi \wedge \psi) \in \mathcal{L}$

f	$[\![\cdot]\!]$
• $f(\textit{John is pointing}) = p$	• $[\![p]\!] = 1$
• $f(\textit{Kim is pointing}) = p$	• $[\![q]\!] = 0$
• $f([s_1 \textit{ and } s_2]) = (f(s_1) \wedge f(s_2))$	• $[\![(\varphi \wedge \psi)]\!] = [\![\varphi]\!] \times [\![\psi]\!]$

In this theory, the objects in Ω (the semantic realm) are the boolean values 1 and 0, and functions from boolean values to boolean values. In particular, *and* is interpreted as boolean multiplication — a binary function that gives and 1 if both the arguments are 1 and 0 if one or both of them is 0.[4] Typically an analysis such as this one is presented as a **truth conditional** semantic theory, meaning that these boolean values are taken to correspond to the concepts of *truth* and *falsity*. In fact, the symbols T and F are often used for these values to emphasize that relationship, but here we use 1 and 0 to make the point that Ω is a realm of formal objects with only a meta-theoretical relation to the concepts of truth and falsity.

Notice that $\mathcal{N}$ is defined in such a way that expressions are endowed with syntactic structure, here conveyed with square brackets:

(13) a. [*John is pointing and Kim is pointing*]

b. [*Kim is pointing and John is pointing*]

c. [[*John is pointing and Kim is pointing*] *and John is pointing*]

[3] Indeed, some semantic theories see fit to do away with Ω all together, instead allowing a syntactically defined notion of logical consequence (i.e., a proof system) give content to the natural language expressions.

[4] In this presentation, *and* is interpreted *syncategorematically* since we don't give it a denotation directly, but rather provide a rule for interpreting expressions involving *and*. The categorematic approach would interpret the word *and* as the multiplication operation directly. In this case, the two approaches are essentially equivalent.

Each of the sentences of (Section 5.1) are part of $\mathcal{N}$. The brackets make it easy to define the translation function, f, which maps (Section 5.1) to $(p \wedge q)$, (Section 5.1) to $(q \wedge p)$, and (Section 5.1) to $((p \wedge q) \wedge p)$. Importantly, the semantics of the logic gives the same interpretation for all three of these sentences, due to the commutativity and associativity of boolean multiplication. This is an example of a prediction that makes the formal semantic theory falsifiable by empirical data like speaker judgments — if speakers don't think these three sentences have the same meaning, there might be a problem with the theory.[5]

Naturally this simplistic theory has quite a few shortcomings as a formal semantics. First of all, it fails to capture the compositional semantics internal to the expressions that are translated to p and q. What if we want to talk about *other* people pointing? Do we really need to enumerate every such sentence? Surely that's not analogous to how it works in the natural language — if we know who *Lisa* is, we can understand *Lisa is pointing* by analogy, even if we've never heard that exact sentence before.[6] Perhaps even more damning (insofar as this formalization is supposed to give a theory of the meaning of the word *and*), we can't seem to account for sentences like these:

(14) a. [[*John and Kim*] *are pointing*]

b. [*John is* [*pointing and laughing*]]

How does the word *and* behave when it is not joining propositions but expressions denoting entities or actions? Our theory has nothing to say here since there sentences aren't part of our fragment of English. We might expect a more complete theory to satisfy certain intuitions, such as that (Section 5.1) is given the same interpretation as (Section 5.1), or that (Section 5.1) entails *John is pointing*. Such an analysis would require a more expressive system than propositional logic, of which $\mathcal{L}$ and $[\![\cdot]\!]$ are a subset. More generally, a satisfactory theory ought to cover a more complete fragment of English.

Another problem is that this theory equates the meaning of an expression with its truth value. This means that *John is pointing* and [*Kim is pointing and John is pointing*] have the same meaning, 0 (or *False*). This is seems like an odd conclusion since intuitively each of these sentences have a different meaning, even if they both *happen* to be false. After all, isn't it possible for someone to *believe* the first sentence while *not believing* the second? If I utter one of these two sentences am I not conveying different *information*? This is what is known as the problem of **intensionality**. A good semantic theory should give interpretations that go beyond the contingent state of the world,

[5] There is some wiggle room for the theorist here if f and $f \circ [\![\cdot]\!]$ are taken to model two different kinds of semantic interpretation — note that each of the sentences in (Section 5.1) are translated to different propositional formulas by f even though the interpretation is the same.

[6] Implicit in this criticism is a particular modeling goal for the semantic theory — the meaning of a compositional expression should be computed in a way that is (at some level of description) analogous to the way that actual speakers compute that meaning.

since natural language functions in so-called *intensional contexts* such as when speaking hypothetically or talking about belief states. The success of Montague Semantics is due in part to the fact that its formal analysis extended to a very large fragment of English, including to many expressions that require dealing with intentionality.

Montague Semantics Montague's formal semantics (1970, 1973) uses a categorical grammar to capture a fragment of English and a combination of simply typed lambda calculus and higher order predicate logic for the semantics. Aside from giving a fully compositional treatment of a fragment of English, (*a/an*, *the*, *every*, *some*, etc.), Montague (1973) was principally interested in giving a compositional analysis of intensional contexts and quantified noun phrases (and especially quantified noun phrases *in* intensional contexts).[7]

In an **intensional context**, you cannot replace a constituent with a co-extensive expression without changing the meaning.

(15) a. Kim seeks the President of the United States

b. John is the President of the United States

c. ⁄∴Kim seeks John.

Here, we can tell that 15a includes an intensional context. Since substituting co-extensive terms (*John* for *the President of the United States* is not truth-preserving; it is possible for 15a to be true and 15c to be false if, f or example, Kim does not know that 15b is the case. Compare this, to the following:

(16) a. Kim points at the President of the United States

b. John is the President of the United States

c. ∴Kim points at John.

It would seem that this inference *does* go through. Note that the only difference between 15 and 16 is the verb. *Seeks* creates an intensional context where as *points* usually[8]does not. For this reason, Montague gives an analysis of intensionality where intensional contexts are created by particular lexical items.

Montague Semantics uses a context-free categorical grammar for the syntax of English and intensional logic (a combination of simply typed lambda calculus and model-theoretic higher-order predicate logic) for the semantics. The grammar defines the possible syntactic categories of expressions, CAT as follows:

[7]More detailed introductions to Montague Semantics can be found in Dowty et al. (1981) and Gamut (1991).

[8]There certainly are senses of point that do create an intensional context. Consider, for example, the sense where *point at* is used metaphorically to mean *to make an accusation*. Depending on the situation there might even be contexts where even literal pointing needs to be interpreted intensionally if, for example, the pointing is serving the purpose of making an accusation. This is just another example of how lexical meaning is deeply dependent on communicative context.

1. $CN, IV, S \in CAT$
2. If $A, B \in CAT$ then $A/B \in CAT$

The basic categories correspond to common nouns, intransitive verbs, and sentences. The slash categories can be thought of as something that gives you an A if you provide it with a B (this is specified more formally in the semantics). For example, *individual terms* like *John* are not a basic category, but are instead represented as S/IV — something that given an intransitive verb, will give you a sentence.[9]

Each syntactic category corresponds to a semantic type, which can be described as a formal language:

1. e is a type and t is a type
2. if σ and τ are types, then $(\sigma \to \tau)$ is type
3. if σ is a type, then $(s \to \sigma)$ is a type

The basic type e corresponds to entities and t corresponds to truth values. Combining two types forms a higher-order "function" type. The third "basic" type s (corresponding to possible worlds), is unlike the other two in that it cannot appear alone, but only as the antecedent to higher order types. Function types with s as the antecedent are what introduces intensionality into the system.

The correspondence between syntactic categories and semantic types is defined as follows:

1. $f(CN) = (e \to t)$
2. $f(IV) = (e \to t)$
3. $f(s) = t$
4. $f(A/B) = ((s \to f(B)) \to f(A)$

Finally, a lexicon assigns a grammatical category and an intensional logic formula (of the appropriate type) to each English lexical item. Intensional logic uses Kripke models (cite) for its semantics. Kripke models are set-theoretic constructions that include functions from possible worlds (corresponding to the semantic types of the form $(s \to \sigma)$, which allow the system to account for intensional phenomena.

It's not an exaggeration to say that contemporary formal semantics largely defined by variations and extensions of Montague Grammar. Some of these extensions use

[9] In fact, Montague's original presentation was a bit different, taking individual terms as a basic category and common nouns and intransitive verbs as derivative. Bennett (1976) came up with this version, which simplifies the grammar somewhat.

more powerful syntactic formalisms that can deal with phenomena like discontinuous constituents. Another line of work extends formal analysis to more semantic phenomena by using *rich type theories* in the semantics. We will briefly introduce this trend more generally before giving and overview of TTR (Section 5.1.1).

Type theoretic semantics Since Montague, the field of formal semantics has grown rapidly, extending formal analysis to fragments of to cover more and more semantic phenomena (including, importantly, phenomena that do not occur in English). A subfield of formal semantics has focused on applying new methods in type theory, a field which has independently seen a flourishing in recent decades, with applications in programming language theory and foundational mathematics as well as linguistics.[10] These type theories differ in character in a number of ways from the simply typed lambda calculus. For one, they are often **many-sorted**, meaning that the basic types are not a closed class as they are in Montague's intensional logic (limited to e, t and s), but are formally more akin to propositions in propositional logic or predicates in predicate logic in that the system could, in principle, include any number of them, depending on the lexicon in the fragment of natural language being modeled.

A many-sorted type theory might, for example, might have a type *Man* corresponding to the noun *man* and a word *Point* corresponding to the verb *to point*. We would write

$$j : \mathit{Point} \tag{5.1}$$

to mean that the object j (corresponding to *John*) is pointing — that is, John is the type of thing that is pointing. If the type judgment expressed by (Eq. (5.1)) holds, we would say that j is a **witness** for the type *Point*.

Another feature that makes rich type theories attractive for formal semantics is the **types as propositions** interpretation of types. Under this interpretation, a type stands for a proposition; namely the proposition that the type has a witness. In this interpretation *Point* would stand for the proposition that *someone is pointing* and $j : \mathit{Point}$ would constitute a proof of that proposition with j as the witness. For systems based on intensional logic, logically equivalent propositions are indistinguishable. But in **hyperintensional** contexts such as belief, it may be necessary for a formal semantics to distinguish between them.

Types as propositions has a convenient relationship with the Austinian notion of truth (Austin, 1950) in which propositions are not fundamentally true *simpliciter*, but rather truth *of* some part of the world. Barwise and Perry (1983) expand on this notion by developing a theory of *parts of the world*, which they call **situations** and types of situations, which correspond to propositions. Just as a proof may be a witness for a proposition, a situation may be a witness for a situation type. Cooper (2005) formalizes this relationship in Type Theory with Records (TTR), which we will present briefly in the next section.

[10] See Chatzikyriakidis and Cooper (2018) for an overview of type theory for natural language semantics.

5.1.1. Type Theory with Records

Type Theory with Records extends many-sorted dependent type theory with structured objects called **records**, defined as labeled sets of objects (including possibly records):[11]

$$r = \begin{bmatrix} \mathrm{k}_1 & = & \mathrm{a}_1 \\ \vdots & = & \vdots \\ \mathrm{k}_n & = & \mathrm{a}_n \end{bmatrix} \tag{5.2}$$

and corresponding structured types called **record types**, which are labeled sets of types (including possibly record types):

$$T = \begin{bmatrix} \mathrm{l}_1 & = & \mathrm{T}_1 \\ \vdots & = & \vdots \\ \mathrm{l}_m & = & \mathrm{T}_m \end{bmatrix} \tag{5.3}$$

Here, $\{k_1, ..., k_m\}$ and $\{l_1, ..., l_n\}$ are sets of *labels*, drawn from a special set of symbols reserved for labeling records and record types. We write $r.k$ to refer to the object r with corresponding to the label k. The record r is of type T (written $r : T$) just in case for each l_i, there is some k_j such that $r.k_j : T.l_i$.[12]

Type Theory with Records (TTR) is a logical system like the simply typed lambda calculus — on its own, it doesn't offer any theory of natural language semantics as such. However, formal semantic theories that use TTR do tend to have certain aspirations in common, which are supported by the expressive features of TTR. For one, these theories usually try to persevere much of the compositional analysis afforded by Montague semantics. This is made possible by the fact that TTR is a dependent type theory, meaning that it has all of the expressive power of the simply typed lambda calculus. TTR theories are often oriented towards going beyond sentence-based theories of meaning, focusing instead on interaction a starting point for natural language semantics. Modeling action (and the change that results from action) is core to these theories and type theory with records is particularly well-suited to model that kind of dynamics (Cooper, 2012).

This focus on action and interaction also gives TTR-based theories an agent-oriented outlook. Types and type judgments are often taken to be relative to a particular agent, whose *information state* is modeled with a record. Actions (for example an utterance of their own or by another agent) then result in updates to this information state, in what is called the **information state update** (ISU) approach to semantics (Larsson, 2002; Traum & Larsson, 2003). Related to this is the **dialogue game board**, which models the public component of dialogue participans' information state. Dialogue game board theories of dialogue seek to understand the structure of the common ground that is built

[11] A full formal description of TTR can be found in Cooper (2023). Cooper and Ginzburg (2015) also gives a brief introduction to TTR and describes a range of applications in semantics and dialogue.

[12] Technically the type judgment definition for record types also allows *re-labellings*, but we will ignore that detail in this presentation.

up during dialogue, and how speakers make use of it to facilitate communication. Insofar this thesis is interested in how interaction affects common-ground lexical semantic resources, is is important that we can connect the work in the thesis to a dialogue game board account (KoS is one such theory that uses TTR (Ginzburg, 2012)).

TTR-based semantics usually takes an Austinian notion of truth in which, a situation (or alternatively, an agent's *take on* a situation) is modeled by a record and propositions are modeled by a record type. Consider a type judgment like the following:[13]

$$\begin{bmatrix} \mathrm{x} & = & \mathrm{jack} \\ \mathrm{y} & = & \mathrm{helen} \\ \mathrm{c}_1 & = & \mathrm{s}_1 \\ \mathrm{c}_2 & = & \mathrm{s}_2 \end{bmatrix} : \begin{bmatrix} \mathrm{x} = \mathrm{jack} & : & \mathit{Ind} \\ \mathrm{y} & : & \mathit{Ind} \\ \mathrm{c}_1 & : & \mathrm{PointAt}(x, y) \end{bmatrix}$$

Here, the situation (on the left) is judged to be of the type of situation where Jack is pointing at someone. This implies that $\mathrm{s}_1 : \mathrm{PointAt}(\mathrm{jack}, \mathrm{a})$ for some individual a. The object s_1 can be thought of as a part or aspect of a situation. Note that if it's the case that $\mathrm{s}_2 : \mathrm{PointAt}(\mathrm{helen}, \mathrm{jack})$ then the record is *also* a situation of the type where Helen is pointing at Jack. The definition for record type judgments mirrors the intuition that situation types (infons in Barwise and Perry (1983)'s terminology) can involve some underspecification. This allows situation types to model underspecification in a word's lexical meaning.

The type of situation in which Jack points at Helen is a **subtype** of the type of situation where they are both pointing at each other, since something of the second type is always also of the first type:

$$\begin{bmatrix} \mathrm{x} = \mathrm{jack} & : & \mathit{Ind} \\ \mathrm{y} = \mathrm{helen} & : & \mathit{Ind} \\ \mathrm{c}_1 & : & \mathrm{PointAt}(x, y) \\ \mathrm{c}_2 & : & \mathrm{PointAt}(y, x) \end{bmatrix} \sqsubseteq \begin{bmatrix} \mathrm{x} = \mathrm{jack} & : & \mathit{Ind} \\ \mathrm{y} = \mathrm{helen} & : & \mathit{Ind} \\ \mathrm{c}_1 & : & \mathrm{PointAt}(x, y) \end{bmatrix}$$

This subtype relation can be verified by examining the structure of the two record types — for every label on the right-hand side there is a label on the left-hand side corresponding to a type that is either equal to– or a subtype of– (but in this case always equal to) the type on the right. This subtype relation also holds:

$$\begin{bmatrix} \mathrm{x} = \mathrm{jack} & : & \mathit{Ind} \\ \mathrm{y} = \mathrm{helen} & : & \mathit{Ind} \\ \mathrm{c}_1 & : & \mathrm{PointAt}(x, y) \end{bmatrix} \sqsubseteq \begin{bmatrix} \mathrm{x} = \mathrm{jack} & : & \mathit{Ind} \\ \mathrm{y} & : & \mathit{Ind} \\ \mathrm{c}_1 & : & \mathrm{PointAt}(x, y) \end{bmatrix}$$

To see that the type situation where Jack points at Helen is a subtype of the type of situation where Jack points at *someone*, we need to know that $T_{\mathrm{helen}} \sqsubseteq Ind$, which is true by the definition singleton types (see Footnote 13).

[13] The notation $\mathrm{x} = \mathrm{jack} : \mathit{Ind}$ is a *manifest field* (Coquand et al., 2003), which is shorthand for $x : \mathit{Ind}$ and $x : \mathit{Ind}_{\mathrm{jack}}$, where $\mathit{Ind}_{\mathrm{jack}}$ is a *singleton type*. In general, for any object $a : T$, $b : T_a$ if and only if $b : T$ and $b = a$.

5.1.2. Probabilistic Type Theory with Records

Probability theory provides a mathematical formalization of uncertainty. Insofar as natural language interpretation deals with uncertainty, probabilistic concepts can be useful in formal semantics. Probabilistic Type Theory with Records (Cooper et al., 2015) adapts TTR to the probabilistic setting by definition a probabilistic type judgment:

$$p(\mathrm{a} : T) = r,$$

there, r is a real number between 0 and 1. We can read this as saying that the probability that a is of type T is r. The value of p is given by a probability model.[14] Conditional type judgements can be expressed similarly:

$$p(a : T_2 \mid a : T_1) = r$$

In the probabilistic setting there are multiple candidate notions for subtype, but a minimal requirement for $T_1 \sqsubseteq T_2$ would be that whatever something is certainly of T_1 it is certainly of type T_2.[15] That is,

$$T_1 \sqsubseteq T_2 \Rightarrow p(a : T_2 \mid a : T_1) = 1.$$

5.1.3. Classifier-based meaning

In order to ground perceptual meaning in classification in TTR, we need to do two things: (1) we need to give an account of how, given a classification function, the semantics of the TTR types it is based on are determined, and (2), we need to encode the classifier in TTR in such a way that makes the classification function available. Another approach would be to forego (2) and instead use the classifier as a witness condition for some type corresponding to the meaning (in the place of a set theoretic model, for example). The problem with this approach is that if we want linguistic activity to serve as a basis for semantic learning, we need to make the parameters of the classifier, not just the classification function, available to our theory of interaction, which is stated in TTR (see Fernandez & Larsson, 2014; Larsson & Cooper, 2021).

[14] Cooper et al. (2015) first defines the model as a probability function over a set of possible worlds, following van Eijck and Lappin (2012). Doing so guarantees adherence to the standard Kolmogorov (1950) probability axioms, but at the cost of completeness and cognitive plausibility. They suggest that probabilities might alternatively be assigned to situation types, as this is analogous to the assumption that is commonly made in probabilistic AI in which the universe of *worlds* is not made up of maximally consistent sets of propositions, but rather a local set of alternative possible outcomes (Cooper & Ginzburg, 2015, §1.2). Another approach might be break from classical probability theory and use a model theory that assigns probability to type judgments directly in the style of de Finetti (see de Finetti, 1992). Indeed, this is essentially what we do in Chapter 8 when we use classifiers as witness conditions for certain types. More work is needed to ensure that this approach would yield a well-behaved probabilistic type system in the general case, however.

[15] A stricter requirement, for example might be that $p(a : T_1) \leq p(a : T_2)$ in every possible interpretation.

Larsson (2013) demonstrates how to encode a linear perception in TTR, using the example of providing grounded semantics for the terms *left* and *right*. In Chapter 8, we expand this treatment to multiclass classifiers. To do so, we define a categorical variable type $\mathbb{A}$, which ranges over a set of value types $\mathfrak{R}(\mathbb{A}) = (A_1, ..., A_n)$. Where each A_i is a record type.[16] A classifier $\kappa_{\mathbb{A}}$, for $\mathbb{A}$ is a function of the following type:

$$\Pi \rightarrow Sit_{\mathfrak{Y}} \rightarrow \{ \begin{bmatrix} \text{sit} & : & Sit_{\mathfrak{Y}} \\ \text{sit-type} & : & RecType_{A_i} \\ \text{prob} & : & [0, 1] \end{bmatrix} \mid A_i \in \mathfrak{R}(\mathbb{A}) \}.$$

Here, Π is the type of the parameters needed by by the classifier, $Sit_{\mathfrak{Y}}$ is the type of situations that yield perceptual input, and $RecType_{A_i}$ is the (singleton) type of records identical to A_i. We assume that for parameters $\pi : \Pi$ and input $x : Sit_{\mathfrak{Y}}$, we have $\sum_i \kappa_{\mathbb{A}}(\pi)(x)(A_i) = 1$.[17]

5.2. Computational methods

Formal methods seek to precisely state a theory of how meaning works in natural language. As we have seen in Section 5.1, the scope of such a theories has been extended (or perhaps shifted) in some cases to focus not only sentence meaning, but the meaning of a wide range of types of utterances situated in an interactive context. Computational semantics starts with a similar goal, to understand meaning in natural language. The methodology in computational semantics — its relationship it's relationship to *models*, *data*, and *hypotheses* tends to be much different, however.

Data and preprocessing Humans receive linguistic input as combination of audio and visual signals,[18] most typically perceived in the course of an interaction in which we ourselves take part. Most psycholinguists would agree that processing speech signals involves some degree of discretization by way of classifying the continuous input (sounds into phonemes, strings of phonemes into words, etc.). Nevertheless, the continuous signal remains available, for example, for use in communicative repair.

In computational linguistics, we rarely work with the raw speech signal,[19] applying some discretization before we even start modeling. For spoken data, this means working with a transcript, which is the result of a laborious human transcription process or a noisy speech-to-text system. Transcription comes with a lot of choices about what

[16] In Chapter 9 these correspond to the meaning of lexical item.

[17] In Chapter 9 this is ensured by the standard standard softmax function used in neural multiclass classification models.

[18] Why visual signals? In addition to the many signed languages of the world, gesture serves an important communicative function in-person spoken and signed dialogue. In the following *speech*, is used to refer to both verbal and gestural communication spoken and signed interaction.

[19] There are exceptions — computational phonology, for example. But this work generally stays on the level of phonology. It is rare for raw speech signal to be used as input to computational models that work further up the classical linguistic hierarchy.

aspects of the speech and how much of the interaction to capture. Speech-to-text processing is noisy and error-prone and captures an extremely impoverished record of the speech, especially in interactive settings.

For these reasons it is much more common to work with text data in computational linguistics. Digital text is already discretized into characters and hand-written text can be converted to digital text through automatic character recognition which, though not without errors, captures a more faithful representation of the original data than speech-to-text.

This is not where preprocessing ends, however. **Tokenization** is a key preprocessing step for most work in computational linguistics. Tokenization divides a sequence of characters into multi-character strings from a finite vocabulary. Often these tokens are meant to correspond to something like words or lexical items, though there is probably no way to do this perfectly in principle since, as we discussed in Section 2.1, there is no clean separation between lexical and compositional meaning. Sub-word tokenization strategies are also popular. They divide text into hopefully meaningful sub-word units, either by employing some morphological analysis, or with strategies that group together common sequences of characters. Once tokenized, text is represented as a sequence of tokens, each of which is drawn from a finite vocabulary of token *types*. One could, for example count up the tokens in a piece of text and compute a distribution of token types over the vocabulary. This sort of thing is the basis for modeling in computational linguistics, including for machine learning models.

Machine learning The *machine learning paradigm* is pervasive in computational semantics and computational linguistics more generally. In this paradigm, an abstract **task** is defined, which seeks to approximate some human-like competency involving language use. *Sentiment analysis* (judging if a piece of text expresses positive or negative sentiment), *natural language inference* (determining if a premise sentence entails a hypothesis, if they are contradictory, or if there is no relation), and *image captioning* are all examples of machine learning tasks that involve natural language semantics.

A **dataset** is the concrete manifestation of a task. A dataset consists of a set of pairs $D = \{\langle x, y, \rangle \mid x \in X, y \in Y\}$, where each x is some input (a sentence, a pair of sentences, or an image for example, respective to the above) and each y is a ground-truth *labels* (a sentiment score, entailment relation, or image caption), usually produced by a human annotator. A **model** is a function, $\varphi(\theta, x)$, that given some parameters, θ and an input, produces something of the same kind as the elements of Y.

Standard practice is to split into disjoint *train* and *test* sets.[20] A *loss function* is defined such that $\mathcal{L}(\hat{y}, y)$ measures the distance between a model prediction and ground-truth label. Then, a learning algorithm is used **train** the model — to find the parameters that minimize the loss over the training set; that is, to find:

[20] And often also a *validation* set, which is used to select hyperparameters, such as model size, and to check during training if the model is *overfit* to the train set.

$$\hat{\theta} = \arg\min_{\theta \in \Omega} \sum_{\langle x,y \rangle \in D_{\text{train}}} \mathcal{L}(\varphi(\theta, x), y), \tag{5.4}$$

where Ω is the space of all possible values of θ.

Generally speaking Ω can be extraordinarily high-dimensional. Together with a complex φ, this means that an analytic solution to Eq. (5.4) often doesn't exist or is completely intractable — you can't just *solve for* $\hat{\theta}$ as you would in algebra class. The core of the discipline of machine learning is to define a model, φ, loss function $\mathcal{L}$ and an algorithm for estimating $\hat{\theta}$ from D_{train} such that $\varphi(\hat{\theta}, x)$ has good *performance* on the test set according to one or more **performance metrics**, which measure how well the model approximates the ground truth output of the test set.[21] If the dataset is a faithful realization of the task, performance on the test set will indicate how well the model *generalizes* beyond the training data — that is, how good it is at performing the task *in general* without respect to the particular examples it "saw" during training.

Scientific knowledge is not always the goal of machine learning — a model that performs well on a particular task can have useful real-world applications. But it's worthwhile to consider how computational methods differ from formal methods when the aim is to discover something about natural language as an empirical phenomenon. What does it mean if a model performs well on a particular task? This depends somewhat on the dataset, but if the model exhibits non-trivial generalization to the test set, that means it has learned some patterns that connect the input and the output. If we take natural language inference (NLI) as an example, we could see a machine learning model that performs well to be a model of inferential meaning in natural language in the same way that formal models with the same goal are. Such a model could be taken as evidence that (1) the training data is sufficient to learn how inference works in general and (2) the model architecture has sufficient computational power to determine entailment relations, as well as to learn *how* to determine them from the training data. In practice, there are reasons to be skeptical of claims that NLI models, even ones that perform very well, really capture inferential meaning the way that human semantic interpretation does. For one thing, the NLI datasets available capture a certain only a certain *kind* of inference which is a bit different from what is assumed by formal theories and doesn't generalize to all contexts (Bernardy & Chatzikyriakidis, 2019). Furthermore, some models can perform almost as well when they are trained using the premise sentence alone, suggesting that the model is taking a "shortcut", using certain correlations between the premise sentences and the entailment relation without considering the hypothesis sentence at all (Gururangan et al., 2018).

[21] The loss function is often different from the performance metrics are used in testing. This is counter-intuitive; you might think that maximizing the same performance metrics during training would be the best way to maximize them during testing. However, the loss function must be chosen carefully in conjunction with the learning algorithm. For example, most strategies for training neural networks (Section 5.2.1) require that a gradient of the loss function can be computed with respect to the model's current parameters. Doing so requires a differentiable loss function, which is not generally the case for performance metrics.

Language modeling is an especially important task that has a complicated relationship with both practical applications and linguistic theory. In the strict sense, a **language model** is a function that estimates a probability distribution over strings of tokens. This is usually done by training the model to perform *next token prediction*, since a model that computes the probability of a token given its preceding context can be used to compute the probability of the string:

$$P(w_1, ...w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, ..., w_{i-1})$$

Of course, it is also impossible to compute the conditional probability on the right, since in principle the context string, $w_1, ..., w_{i-1}$, could be anything. A language model must estimate this probability, for example by substituting it for $P(w_i \mid w_{i-n}, ..., w_{i-1})$ for a small value of n, as in a *n-gram* model.

One especially salient feature of language modeling as a task is the fact that it doesn't require any annotation. Since the training objective is to predict the next word in a sequence, tokenized text can serve as its own labelled data. This makes language modeling a **self-supervised** learning paradigm.

Since language models estimate a probability distribution over possible strings, they can be seen as the statistical corollary to a formal language (Section 5.1) It has been argued (Lau et al., 2017) that this means they can be interpreted as capturing grammatical competence in a particular language. But they are also enormously useful in downstream tasks. For example, in machine translation, a language model can serve as a prior distribution of strings in the target language, meaning that the translation model need only estimate a probability of source language strings *given* a string in the target language. This is what is known as a *noisy channel* translation model. The intermediary representations learned by a language model can also be useful. This is especially relevant for neural language models, as we will discuss in the next section.

5.2.1. Neural network models

A neural network is a kind of machine learning model inspired by the way synaptic signals pass through biological brains. Most neural network models are arranged in layers, with the output of the previous layer supplying the input for the next layer. A network with multiple layers is called **deep**. Intermediary layers are called **hidden layers** and their outputs are **hidden states**.

Each layer consists of a collection of "neurons" (the gray circles in Fig. 5.1), each of which compute a real-number value based on the outputs of the neurons of the previous layer and some trainable parameters. An *activation function* is usually applied as a final step in computing the neuron's output value. Activation functions are often non-linear, amplifying the output value if it reaches a certain (soft) threshold. The activation function is supposed to mimic the behaviour of biological neurons whose

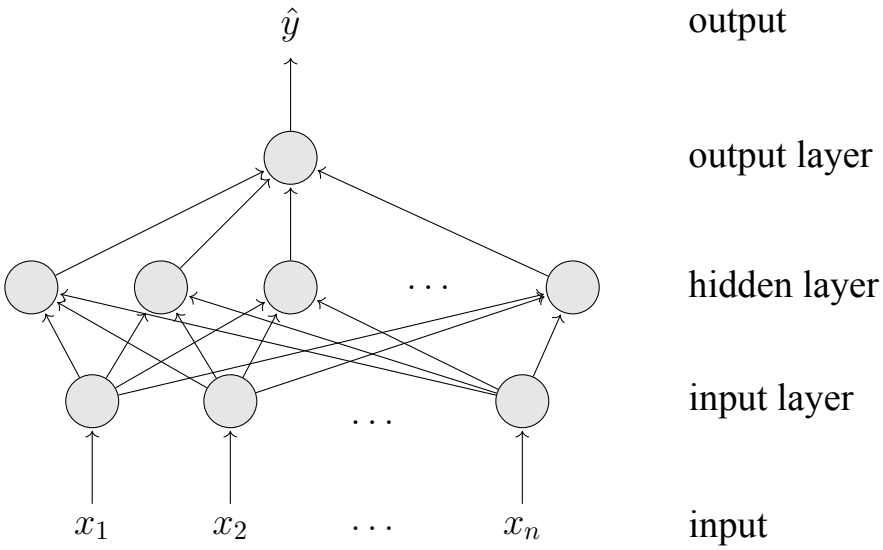


Figure 5.1.: Simple neural network with 3 fully-connected layers and a one-dimensional output.

synapses *fire* given a certain amount of stimulus. In practice, the non-linearity of the activation function is what lets deep neural networks learn functions where the desired output can not be computed as a linear combination of the input.

Figure 5.1 may look fancy, but in fact it is just a series of matrix multiplications with a bias term added:

$$\begin{aligned} \hat{y} &= \sigma_2(\mathbf{W_2} \cdot \mathbf{h_2} + \mathbf{b_2}), && \text{where} \\ \mathbf{h_2} &= \sigma_1(\mathbf{W_1} \cdot \mathbf{h_1} + \mathbf{b_1}) && \text{and} \\ \mathbf{h_1} &= \sigma_0(\mathbf{W_0} \cdot \mathbf{x} + \mathbf{b_0}). \end{aligned} \tag{5.5}$$

Here, for example, the ith neuron of the hidden layer is parametrized by the ith row of $\mathbf{W_1}$ and the ith element of $\mathbf{b_1}$, and σ_1 is the activation function for that layer.

Different neural network architectures have different patterns of connections between neurons. Convolutional layers, for example, compute outputs based on a sliding context window over the input. Recurrent layers are made up of a sequence of *recurrent units*, where each item in the sequential output is a function of the previous item and a (possibly) sequential input.

Modern neural networks are commonly trained with the aid of *back-propagation*, an algorithm that estimates the gradient of the model's parameters with respect to the loss (Rumelhart et al., 1986) This gradient is used by optimization functions like *gradient descent* to incrementally adjust the parameters to minimize the loss by moving the parameters in the direction indicated by the gradient. The *learning rate* controls how big of a step in the direction of the gradient the optimizer takes. Some optimizers also include a momentum hyperparameter, which biases the change in parameters to

continue on in the direction it went in previous steps. The Adam optimizer (Kingma et al., 2015), which is used to train the neural networks used in Chapters 10 and 12, is such an optimizer.

Neural networks dealing with text data usually make use of an **embedding** layer, which converts tokens into vectors of a particular dimensionality or *size*, n. It's called an embedding because, given the right learning objective, it *embeds* the vocabulary in the $\mathbb{R}^n$ vector space. More precisely, given a vocabulary of tokens V, and a embedding size n, a embedding layer is a function: $Emb : \mathbf{W} \rightarrow (V \rightarrow \mathbb{R}^n)$, with parameter matrix $\mathbf{W} \in |V| \times n$ and where

$$Emb(v_i) = \mathbf{W}_i.$$

Converting discrete tokens into continuous-valued vectors makes them available as signals for later layers of the network. Like the parameters of other layers, $\mathbf{W}$ is learned in training. When the tokens are roughly word-level units of text, the rows of $\mathbf{W}$ are called **word vectors**. But embeddings are suitable for representing any kind of discrete input, not just text tokens — especially if the "vocabulary" items have relationships between each other that can be learned during training. In Chapter 12 we use an embedding to represent a discrete set of communities. Each community was represented by a vector of size $n = 64$ and the model learned to represent similar communities with similar vectors.

Neural networks are highly *modular*, meaning that it is relatively easy to swap out layers or extend a model with additional layers. This is the principle behind **pre-training** in which a model is trained on one task (language modeling is particularly popular, since it is self-supervised) and one or more of the initial layers along with their learned parameters are joined with additional untrained layers to perform a different *target task*. The new combined model is then trained on on data for the target task. The parameters of the new layers are trained while the parameters of the pre-trained layers are either kept *frozen* (meaning the layers act as a constant function of the input) or *fine-tuned* — also trained, but often with a slower learning rate.

Until recently, the typical way to make use of pre-training in NLP was to use pre-trained word embeddings. A common practice, for example, was to train word embeddings using a model like skipgram (Mikolov et al., 2013). And then use those word embeddings as input for recurrent neural model like an LSTM (Hochreiter & Schmidhuber, 1997) to perform sequence-level tasks. The skipgram model learns word embeddings by trying to predict context words drawn from a context window of a certain size around the input word.

Now it has become much more common to use multiple pre-trained layers of a deep neural network. These models produce representations that are a function of a whole sequence of tokens, meaning that they can, in principle, capture not just lexical but also some degree of compositional meaning. BERT (Devlin et al., 2019) is one such model. It is trained by *masked token prediction*, where it tries to guess the identity of

one or more tokens that have been masked out, as in a cloze task.[22]

5.2.2. Semantic change detection

There is a still relatively young but growing subfield of computational semantics that uses distributional methods to study semantic change.[23] Following the machine learning paradigm, this has been construed as a task, **lexical change detection** (LCD), where the objective is to automatically determine which words in a diachronic corpus have changed in meaning (or at least change in *usage*) over a certain time period.[24]

There are wide range of approaches to mode design for LCD. There are also a variety of ways the task can be realized, and any realization of the task needs to make certain assumptions about the nature of semantic change. In some variations the task involves detecting the *kind of change* that takes place for a certain word over the period of interest. The DURel corpus (Schlechtweg et al., 2018), for example, distinguishes between *innovative meaning change* and *reductive meaning change*, a distinction that the authors justify by demonstrating good inter-annotator agreement in the dataset.

The same can be said of different *methods* of semantic change detection. For example, one methodology involves **word sense induction** (WSI) or **word sense disambiguation** (WSD). WSD is the task of assigning a sense from a pre-determined sense inventory to each instance of a certain set of vocabulary items in a corpus. WSI is the same, but no pre-determined sense inventory is provided. On the assumption that semantic change typically involves adding or removing senses from a word's sense inventory, a WSI or WSD model can then be used for semantic change detection by measuring how sense distributions change over a certain time period (see, for example Mitra et al., 2015; Tahmasebi et al., 2013).

Another very popular methodology, is to use **diachronic word vectors**. This method involves training separate sets of word vectors across multiple time periods and ensuring the vocabularies are embedded in a shared vector space across time.[25] In Chapter 13 we used a *diachronic skip-gram* model (Kim et al., 2014) with noise-rectified change scores (Dubossarsky et al., 2017) to compare semantic change across a collection of online communities. Diachronic word vectors were preferable to a methodology using WSD/WSI for this study because it took relatively little training data and because we

[22] Both skipgram and BERT are sometimes referred to as language models, but it's important to point out that the are not language models in the sense introduced previously since they don't perform next-token prediction, and so are not trained to estimate a probability distribution over strings. They are similar, however in that the training objective allows for self-supervised learning.

[23] See Tahmasebi et al. (2021) and Kutuzov et al. (2018) for recent surveys. The former is a comprehensive survey including a diversity of methodologies and the later focuses on diachronic word vectors.

[24] A recent example of an LCD is the 2020 SemEval shared task in unsupervised lexical semantic change detection (Schlechtweg et al., 2020). A *shared task* is an event where different teams concurrently design and train models to perform a certain task using the same dataset.

[25] This is achieved by either by sharing some model parameters across time periods or by post-hoc alignment (see Hamilton et al., 2016, for detalis).

could measure change for a given word as a scalar value, comparable to other items in the vocabulary and across communities.

Diachronic word vectors, as most LSC methodologies, rely heavily on the distributional hypothesis. It is important to interpret the results of any study involving LCD with this in mind. As we've discussed, meaning representations based on the distributional hypothesis really measure change in *usage*, which may or may not correspond to a change in lexicalized meaning *potential* with respect to a certain community. This can have surprising implications for the aspects of meaning change that are captured. In Chapter 13 we measured meaning change in a collection of online communities. We found that one word consistently appeared among the words that changed most in each community: *2016*. As it happens, the two time time periods in our corpus were from 2015 and 2017, so *2016* went from referring to a year in the *future* to referring to a year in the *past*. One could argue that it didn't really change in *meaning* since the denotation was the same, but the contexts in which the word appeared nevertheless did change.

5.3. Statistical modeling

Statistical techniques can be applied to almost any kind of data. In general, they are useful when there is some underlying stochasticity, perhaps due to features that aren't observed as variables in the dataset. In this situation, we might want to know if some relationship between variables is "real", or if it can be explained by chance alone. With this in mind, statistical modeling can have essentially there different goals: (1) to **test** a hypothesis about a relationship between variables, (2) to **explore** relationships in the data, and (3) to **predict** some unknown or not-yet-realized values based on known values. While all of these are perfectly valid goals conflating them can lead to the appearance of statistical significance where there is none. For example, if one does some exploratory analysis to find relationships between variables, confidence metrics like the p-value are no longer a good indicator of whether the relationship is statistically significant since the p-value assumes the researcher is testing an *a priori* hypothesis.[26] Chapter 13, wherein we do perform exploratory analysis of the relationship between community structure and semantic change, is the only study in the thesis with sophisticated statistical modeling, but there are instances of statistical testing elsewhere in the thesis.

Agreement statistics In Chapter 7 we developed a new annotation scheme for word meaning negotiation. We wanted to test how much our annotators agreed on their annotations, since high agreement means that the results of the annotation can be considered **reliable** — suitable to use as the basis for further analysis and modeling.

[26] See McGill (2013) for further discussion.

High agreement can also seen be seen as validation that the annotation schema captures categories that correspond to "real" categories, although as we discuss in Chapter 7 this can be controversial since some real-world phenomena are inherently subjective.

Agreement statistics test whether annotators agree more than one would expect by random chance. Suppose we can always frame an annotation task as a collection of items, I, where annotators select one of a set of labels, L for each item. For annotator A, let $L_A : I \to L$ represent their annotations — i.e., let $L_A(i) \in L$ be the label that annotator A assigns to item $i \in I$. The naive agreement statistic, A_0, measures what proportion of items the two annotators agree on. Assuming we have two annotators A and B,[27]

$$A_0 = \frac{|\{i \in I \mid L_A(i) = L_B(i)\}|}{|I|} \tag{5.6}$$

The problem with A_0 as a metric is that it doesn't account for the possibility that some annotations will agree by chance, and that this is more likely to happen for labels that are more common. This makes A_0 incomparable between label sets and difficult to interpret in general. Agreement statistics thus try to adjust the agreement score based on the prior distribution the labels. This is done by computing the expected chance-level agreement, A_e, then computing a ratio which tells us what proportion of agreement beyond chance-level was actually observed:

$$\frac{A_0 - A_e}{1 - A_e} \tag{5.7}$$

The difficulty is, since we don't have access to an objective ground truth, we don't know what the prior distribution is and therefore have no objective way of computing A_e. In Chapter 7 we use two different statistics, which make different assumptions about how A_e should be estimated. For **Scott's pi**, the chance-level agreement, A_π is estimated from the data by assuming that each label has a different prior, which does not depend on annotator. **Cohen's kappa**, on the other hand, estimates A_κ by assuming that labels have annotator-specific prior distributions.[28] The agreement statistics π and κ are computed by plugging A_π and A_κ respectively in to Eq. (5.7). One reason to compute both statistics is that getting very different results for π and κ would indicate that annotators have different priors for the labels.

Confidence in embeddings There are two occasions in the thesis where we use statistics to measure the significance of measurements taken on *embeddings* (as described in Section 5.2.1). In Chapter 12 we have two embeddings S and L, which each represent the same set of n communities, but are computed in different ways. We

[27] All of these metrics can be generalized to n annotators. For the annotation study in Chapter 7 we had four annotators total, but each item was annotated by two people.

[28] See Artstein and Poesio (2008) for precise definitions of the agreement statistics and an extensive analysis of their use in computational linguistics.

want to test if there is a correlation between them. To do this, we align the coordinate systems, performing *orthogonal Procrustes* by *singular value decomposition*.[29] As a correlation metric, we compute

$$d(L, S) = n - Tr(\Sigma), \tag{5.8}$$

where Σ is the square matrix computed by singular value decomposition and Tr is the sum of its diagonal entries (that is, the sum of the *singular values*). As explained in Chapter 12, if $Tr(\Sigma)$ is equal to n, this would correspond to a perfect correlation between the two matrices, so Eq. (5.8) provides a normalized correlation metric between embeddings.

The problem is that singular value decomposition will *always* find some correlation between embeddings, even if they are completely random. To establish the significance of the correlations we measured in the paper, we wanted to compare the measured correlation to what one would expect to measure by chance. Since this expectation is difficult to compute analytically, we took 10 random embeddings L'_i and measured $d(S, L'_i)$ for each of the random embeddings. This gave us a mean, $\bar{x}_d$, and Bessel's corrected standard deviation,[30] s_d. From there we computed

$$\frac{d(S, L) - \bar{x}_d}{s_d}, \tag{5.9}$$

the *number of standard deviations* between the similarity computed for the real embedding and the mean of the similarities computed for each of the random embeddings. This was observed to be between 60 and 70 for each of the versions of L we tested, meaning that we could be very confident in concluding that the observed correlations were not by chance.

In Chapter 13 the situation was a bit different. We again had two aligned embeddings, but this time they were diachronic word embeddings (see Section 5.2.2), and we wanted to measure word-level change. In general, we can measure change for a word as the cosine distance between its two vector representations:

$$\Delta^{\cos}(\vec{w}_0, \vec{w}_1) = \frac{\cos^{-1}(\cos\,\text{sim}(\vec{w}_0, \vec{w}_1))}{\pi} \tag{5.10}$$

where

$$\cos\,\text{sim}(\vec{w}_0, \vec{w}_1) = \frac{\vec{w}_0 \cdot \vec{w}_1}{\|\vec{w}_0\|\|\vec{w}_1\|}. \tag{5.11}$$

[29] In general, two coordinate systems can encode the same information in different ways. Consider, for example CMYK color coding versus RGB, or a faucet that control water temperature and pressure with two knobs versus one. All the same results are achievable in both cases, but coordinate (or control) systems represent them differently. Orthogonal Procrustes is the problem of finding a transformation that aligns two vector spaces. Singular value decomposition is a kind of matrix factorization that can be used to solve the orthogonal Procrustes problem (Schönemann, 1966).

[30] Bessel's correction is a way of estimating a true prior standard deviation from the standard deviation of a sample.

The problem is that over a very large vocabulary, some amount of change is bound to observed by chance. This is related to the fact that word vectors are meaning representations based on the distributional hypothesis. Even if a word has not changed in meaning at all, there might, just by chance, be statistical regularities in the differences in contexts that it appears in across time periods. This is *especially* true for words that appear in highly variable contexts (for example, because of polysemy or lexical flexibility).

We corrected for this problem using a method described by Dubossarsky et al. (2017). First, we constructed 10 pseudo-diachronic corpora by shuffling the data from the two time periods and splitting them in half again. This resulted in 10 corpora with the same structure as the genuinely diachronic corpus, but where the actual dates of the texts were evenly distributed across time periods. Then, we trained the two embeddings again on 10 pseudo-diachronic corpora, resulting in a pseudo-diachronic embedding pair $\langle w'_{i,0}, w'_{i,1} \rangle$ for each word w and random trial i. In theory we would expect to measure $\Delta^{\cos}(\vec{w}'_{i,0}, \vec{w}'_{i,1}) = 0$ everywhere since the dates are roughly uniformly distributed over the corpora, so no genuine change can be measured. Of course, due to the reasons stated above, these values will be positive and tend to be larger for words with higher contextual variability. Similar to what was done in Chapter 12, we measured the mean, $\bar{x}_w$, and Bessel's-corrected standard deviation, s_w of $\Delta^{\cos}(\vec{w}'_{i,0}, \vec{w}'_{i,1})$ for each word across the pseudo-diachronic embeddings. We then defined the *rectified change score* as the t-statistic:

$$\Delta^*(\vec{w}_0, \vec{w}_1) = \frac{\Delta^{\cos}(\vec{w}_0, \vec{w}_1) - \bar{x}_w}{s_w \sqrt{1 + 1/10}} \tag{5.12}$$

Note that this is very similar to the metric computed in Eq. (5.9), but the t-statistic is slightly more interpretable — on the assumption that the $\Delta^{\cos}(\vec{w}'_{i,0}, \vec{w}'_{i,1})$ scores are normally distributed on a word level (we checked that they roughly are), the t-statistic can be used to compute confidence intervals. For example, if we observe $\Delta^*(\vec{w}_0, \vec{w}_1) = 4.74$, we can be sure with 95% confidence that continuing to sample $\bar{x}_w$ in the long-run will still show change for w above what can be explained by random noise.

Generalized mixed-effects modeling A generalized linear model is a statistical model that attempts to predict a response variable, based some number of *fixed effect* predictors. The model is called *generalized* because the response variable is not assumed to be normally distributed. A generalized linear *mixed effects* model (GLMM) also includes some number of *random effects* as predictors, which split the data points into groups. The response variable is modeled as sampled from an exponential distribution, which is parametrized by a linear combination of all the predictors — both fixed and random effects. In statistical modeling, these parameters are called a *design matrix*. To *fit* the model is to find the design matrix that explains the maximal amount of variance in the response variable, similar to how a machine learning model is trained

to find parameters that optimize the loss function with respect to the data. Statistical software like the lme4 package for R (Bates et al., 2015) includes various algorithm for fitting the model to the data.

Random effects are used to capture effects that correlate with a predictor, but aren't necessarily a function of its value. In Chapter 13, for example, we use GLMMs to model word-level semantic change (i.e., with change as the response variable) in 45 different online communities. There, we used community ID as a random effect, since we assumed there would likely be idiosyncratic community-level factors affecting the rate semantic change that wouldn't be captured by the other community-level fixed effects (like community size) that we included.

GLMMs, like all linear models, fit the model as a linear combination of the feature variables. But some of the variance in the response variable may also be explained by non-linear combinations of the features. To account for this, it is common to include **interaction features**, which are typically computed as products of two or more of the fixed effect features.

Fitting the model results in coefficients and standard errors (from the design matrix) for each of the included features. Unlike with neural network parameters, these coefficients are nicely interpretable, since they define the linear combination that explains the maximal variance in the response variable. For example, a positive coefficient means that there is a positive correlation between the corresponding feature and the response variable. The standard error can be used to compute a p-value, which helps to asses whether the relationship is statistically significant.

However, in order to ensure that the results are interpretable as described above, it can be necessary to test for **multi-collinearity** among the fixed effects. If one of the predictors can itself be reliably predicted as a linear combination of the other features, then it would be dubious to use the model coefficients to infer effects among the predictors, since they may be acting as proxies for each other in ways that can't be easily identified. For that reason, it is good practice to do some multi-collinearity detection before fitting a GLMM. This can be done by calculating the *variance inflation factor* (VIF) on a simple linear regression model (Fox & Monette, 1992). The VIF is used to find a set of predictors where the overall multi-collinearity of the model is low enough that the results of the GLMM will be reliably interpretable.

In Chapter 13, after eliminating predictors to reduce multi-collinearity, we performed our exploratory analysis by backwards model selection. We started with six fixed effects and all interactions between the three community-level and the three word-level features. Then, we removed features one-by-one and compared the overall predictive power of the model with and without those features. This allowed us to asses which features had a significant effect on the response variable, per-community word-level semantic change.

5.4. Social network modeling

In social network theory, social networks are modeled by graphs. Social networks analysis is a collection of methodologies used in various social sciences, especially sociology, political science, and economics, but it has also been used in linguistics, especially in sociolongusitics. In general, social network modeling attempts to capture the structure of communities and answer questions about how social structure affects the flow of information and ideas, material resources, and even contagious diseases. Graph models are a very good way of capturing social structure,[31] A graph is a set theoretic object consisting of two components,

$$G = \langle V, E \rangle, \tag{5.13}$$

in which V is a set of *nodes* (also called *vertices*) and $E \subseteq V \times V$ is a set of *edges* that connect the vertices. The nodes (usually) represent individuals and an edge between two nodes $\langle v_1, v_2 \rangle \in E$ represents a (directed) relationship between v_1 and v_2. A graph can also be represented as an *adjacency matrix*, $M : \{0, 1\}^{|V| \times |V|}$ where

$$M_{i,j} = \begin{cases} 1 & \text{if } \langle v_i, v_j \rangle \in E \\ 0 & \text{otherwise} \end{cases} \tag{5.14}$$

Various extensions of the graph-based model are possible. One might like to define multiple sets of edges for different types of relations, for example. Some relations might be directed (like *boss of*) while others are symmetric (like *coworker of*). Edges can also be *weighted*; that is, where $E : (V \times V) \to \mathbb{R}$.

Social network models like the one we use in Chapter 13 have not seen very extensive use in sociolinguistics because it is, in general, difficult to get complete information on all the relationships in a given community.[32] Certain types of social media data make this a possibility, however. While one can never be sure that there aren't interactions going on between members of the community in a different venue, a forum-style social network like Reddit allows the researcher to compile all the interactions that take place on a given forum.

Once one has a graph model of a social network, there are various node- and network-level metrics that can be computed on the graph. *Centrality metrics*, for example, are ways of measuring the "importance" or centrality of a given node in the community. For example, *betweenness centrality* is the proportion of all of the shortest paths between pairs of nodes that go through a given node. *Eigenvector centrality* uses measures how connected a node is to other highly central nodes.[33]

[31] See Jackson (2010) for an introduction to graph-based social network modeling and its applications.

[32] Sharma and Dodsworth (2020) gives survey of social network theory in sociolinguistics, including a detailed explanation of the different types of models.

[33] Eigenvector centrality is the basis for Google's PageRank algorithm.

In Chapter 13 we are interested in comparing different social networks to each other. In particular, we want to measure the effect of network cohesion on the pace of lexical change. As a measure of cohesion we use the **clustering coefficient**, which is defined as follows: First, for a given node v_i, let the *neighborhood* of v_i be the set of nodes connected to i:

$$N(v_i) = \{v_j \in V \mid \langle v_i, v_j \rangle \in E\}. \tag{5.15}$$

The clustering coefficient for a node v_i is defined as the proportion of a nodes neighbors that are also connected to each other

$$C(v_i) = \frac{|\{\langle v_j, v_k \rangle \in E \mid j, k \in N(i)\}|}{|N(i)| \cdot (|N(i)| - 1)}. \tag{5.16}$$

We use this metric to define the community-level metric as the average clustering coefficient across all its nodes.

6. Exposition

you are a participant in the future of language

Ocean Vuong
from *On Being with Krista Tippett*

With both theoretical and methodological background out of the way, we can now turn to the contributions of the thesis. Broadly speaking, the studies can be thought of in two categories. Chapters 7 to 11 are geared towards interaction. With the exception of Chapter 10, which uses neural language models, all of theses studies employ some formal interaction modeling. Chapters 12 and 13 investigate community-level variation and change using machine learning models trained on social media corpora. In contrast to Chapter 10, the neural networks in the final two chapters do not act as models *of agents*, but rather as models of the community-level linguistic norms, aggregating over the data in the corpora.

6.1. Part II summaries

This section contains summaries of each of the studies included in the thesis, with an eye towards how they fit together to tell a cohesive story. In the final part of this chapter we make some concluding remarks that draw insights from across the studies in Part II.

Chapter 7: What do you mean by negotiation?

> Noble, B., Viloria, K., Larsson, S., & Sayeed, A. (2021). What do you mean by negotiation? Annotating social media discussions about word meaning. *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*

Word meaning negotiation (WMN) is a conversational routine in which speakers explicitly discuss the meaning of a word or phrase — the so-called *trigger word* (because it triggered the discussion). This study has two parts. In the first part, we develop a model of WMN as a formal interaction game. In the second part, we use that model to

develop an annotation protocol and report on the results of an annotation study of 150 WMNs collected from Twitter.

The goal of the WMN interaction game model is to describe the structure that these interactions take, what moves are possible at different game states, and effect of different moves on the dialogue state, especially as it pertains to word meaning. The model we describe is built on previous work on WMNs, especially by Myrendal (2015) and Larsson and Myrendal (2017). Additional background on WMN can be found in Section 3.3 of the thesis.

Our model starts with the observation that WMNs involve setting up certain other reference points, which we refer to as semantic **anchors**. These reference points may be introduced with another lexical item, a description of a type of situation, or even a particular individual or situation which is either in the environment or commonly known to the participants. Participants then use these anchors to triangulate the meaning of the trigger word by relating the anchor to the trigger word with semantic **relations** (*X is an example of Y* or *X is a partial definition for Y*), which can then be grounded or rejected by other participants. As the WMN progresses, participants may even draw relations between non-trigger anchors in an attempt to find common ground.

In summary, the game state is represented by a graph structure that includes a set of anchors and a set of relations between those anchors. Relations are decorated with labels that indicate which speakers have committed to the relation (or its negation). The game state defines what future actions are possible (e.g., it is possible to introduce a new relation between anchors that have already been introduced; grounding existing relation is possible if the relation has been proposed). We can also read off **semantic updates** from the game state. The update is computed recursively on the sub-graph of relations that all speakers have committed to. The update works by minimally accommodating the grounded relations — e.g., if for two anchors A and B, it is grounded that B is an example of A, then A is updated such that its interpretation includes B. It should be noted that this constitutes a very conservative update. In Chapter 9 we explore semantic update from a certain type of definition in more detail.

In the second part, we report on the results of an annotation study of 150 WMN interactions from Twitter. The annotation protocol, which was developed by carrying out a series of pilot studies, suggests a sequence of steps for annotating the WMN: (1) identify the trigger word, (2) find text spans that introduce or refer to anchors and determine the relation they describe, (3) connect co-referring anchors, and (4) find explicit statements of commitment or grounding. The annotation protocol results in annotations that can be used to recover game states, as described in the formal model. We also annotated whether the interaction overall was one originating in non-understanding or disagreement. We found good agreement on token-level relation type (example or definition) and polarity (positive or negative), but poor agreement on statements of grounding. We found only moderate agreement on non-understanding or disagreement. Our error analysis found that most annotator disagreements about text spans indicating relations between anchors was disagreement about the extent of the span, or whether it refers to

two anchors or one. We also noticed that a number of disagreements result from different interpretation by the annotators due to different background knowledge about the topic of the Twitter interaction. This highlights the fact that WMNs and the meanings they negotiate can be highly specific to the context of a particular speech community.

Author contributions I developed the initial interaction model in close consultation with Staffan Larsson and Asad Sayeed. Kate Viloria and I conducted the pilot annotations and developed the annotation guide in consultation with Staffan Larsson, which also resulted in adjustments to the interaction model. All the authors performed annotations for the annotation study. I performed the agreement analysis of the results and Staffan Larsson, Asad Sayeed and I conducted the error analysis together. All authors read and approved the final manuscript.

Chapter 8: Classification systems

Noble, B., Larsson, S., & Cooper, R. (2022a). Classification Systems: Combining taxonomical and perceptual lexical meaning. *Proceedings of the 3rd Natural Logic Meets Machine Learning Workshop (NALOMA III)*, 11–16

As we discussed in Section 2.4 of the thesis, lexical meaning seems to have both referential and inferential aspects, though there is no clear separation between the two. We consider the domain of **classification systems** as a case study for unifying these two aspects of meaning. A classification system, as we conceive of it, is a common ground resource for a particular community of practice, which sets out a conceptual structure and methods for classifying entities within that structure for a particular domain. Having these classification systems as common ground facilitates teaching and learning the how to identify new classes within the community. Some examples of classification systems might include the way that a community of mushroom foragers identifies mushrooms, how a group of birders distinguish between local bird species, and the system by which professional astronomers categorize celestial objects. The goal of this paper is to develop a model of lexical meaning that synthesizes referential and inferential aspects of meaning in the context of classification systems. Ideally this model should be compatible with a Montague-style account of compositional semantics.

To do this, we use ProbTTR, which is introduced in Section 5.1.1 of the thesis. Our account starts with two components, a **folk taxonomy**, which represents the structural relations between concepts, and a set of multiclass **perceptual classifiers**, which give content to the concepts. A folk taxonomy is represented by a particular kind of tree structure, where each node can support multiple sets of branches. It can equivalently be represented as a set of **distinctions**, which, consist of a pair including a base concept

and a set of sub-concepts that partition the base concept. With these two ingredients, we define a ProbTTR type system with types representing concepts in the folk taxonomy. We use perceptual classifiers as **witness conditions** for auxiliary types, which are then combined with structural witness conditions that ensure the types in the classification system respect the inferential relationships specified by the taxonomy.

In the end, we have a type system in which probabilistic type judgements can be used to classify where objects belong in the taxonomy. In a small experiment using simulated data, we compare a classification system defined in this way two other methods for classifying in a hierarchical label set and find that, using the same underlying classifier architecture, the classification system outperforms those methods in both precision and recall.

Author contributions I originated the idea of combining perceptual classifiers and taxonomies in classification systems. The type theoretic model was developed in close collaboration by all the authors. I was responsible for the empirical comparison. All authors read and approved the final manuscript.

Chapter 9: Genus-differentia definitions

> Noble, B., Larsson, S., & Cooper, R. (2022b). Coordinating taxonomical and observational meaning: The case of genus-differentia definitions. *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*

Classically, a genus-differentia definition has two parts: First, it gives a **genus**, a super-concept of which the definiendum is a part. Second, it gives a method for differentiating the definiendum from other species of the same genus, the **differentia**. Although genus-differentia definitions are known from their role in the Aristotelian philosophical tradition, many real-world examples from dialogue can be analyzed as genus-differentia definitions, including utterances occurring as moves in a WMN, and corrective feedback, as in in child-directed speech. This chapter builds directly on Chapter 8 to formalize the semantic update incurred by grounding a genus-differentia definition of previously unknown concept in the context of a classification system. Throughout the paper, we use the utterance, *a raven is a large black corvid*, as our canonical example.

The goal of our account is to, given an existing classification system, define a new type, *Raven*, that (1) is a subtype of *Corvid* and, (2) is such that the properties described by the differentia (i.e., being *Large* and *Black*) are *taken as evidence* of that something is of type *Raven*, given that it is of type *Corvid*.

To do this, we first must define record types that correspond to multi-class classifiers. In contrast to Chapter 8 where classifiers were used as witness conditions for

basic types, we need to represent classifiers in the type system so that the type of the definiendum can be defined. We assume that, in addition to the **distinction classifiers** that we previously postulated as part of a classification system, there may be certain **feature classifiers** corresponding to features like *Large* and *Black* that don't directly define types in the taxonomy, but which may be used in conjunction with each other to define those types (for example in a naive Bayes classifier). Some features like *large* may require a comparison class for their interpretation. These we represent as dependent types which, given a certain context type, result in a classifier type (Fernandez & Larsson, 2014).

We first attempt a constructive definition which simply defines *Raven* as something that is a corvid, and large-for-a-corvid, and black (i.e., $Raven = Corvid \wedge (Large(Corvid) \wedge Black)$), but this definition results in the subtype relations $Raven \sqsubseteq (Large(Corvid))$ and $Raven \sqsubseteq Black$), which is undesirable since the subtype relation is intensional and we can at least *imagine* contexts where an individual of the type of the definiendum is nevertheless not a witness of one of the differentia types (an albino raven or a raven chick, for example). Instead, we argue that the definiendum should be represented as an **underspecified type** — a type with no explicit witness conditions, but where certain relationships with other types are specified as constraints on the type system as a whole. We show that an underspecified type can meet our previously stated desiderata under the following conditions: (1) all ravens are corvids ($Raven \sqsubseteq Corvid$), and (2) all else equal, something that is a raven is assumed to be black and large-for-a-corvid ($p(Large(Corvid) \wedge Black \| Raven) = 1$).

Author contributions I conceived of the general approach and made some initial attempts at formalization. Staffan Larsson was responsible for defining multiclass classifiers as a ProbTTR type. The remaining parts of the paper were developed in close collaboration with all the authors. All authors read and approved the final manuscript.

Chapter 10: Describe me an Aucklet

> Noble, B., & Ilinykh, N. (2023). Describe me an Aucklet: Generating Grounded Perceptual Category Descriptions. https://doi.org/10.48550/arXiv.2303.04053

There are many language and vision tasks in machine learning that require some degree of perceptual grounding. Image captioning and visual question answering are two examples of such tasks. But both of these setups put forth a particular image as the focus of each trial in the task (i.e., the image that is being captioned or that the questions are about). When humans use language, though, we can talk about perceptual experience at a level of concepts. Moreover, we argue that the *grounding* can't be

abstracted from a particular *communicative context*. How can you tell if language use is grounded if you don't know what was supposed to be communicated, or what the norms are under which the communication is taking place? The best contexts in which to investigate perceptual grounding in machine learning models are contexts that center communication.

We propose a task that we call *perceptual category description* for this purpose. The scenario is very much like the one described in Chapter 9. A teacher model, which has knowledge of a large set of perceptual classes must describe one or more the classes to a student model. The student model then uses those descriptions of classes they didn't previously know about to classify among all the new and previously known classes. The role of the student model is to perform *zero-shot classification*, which is not in itself a novel task. What we hope to contribute with this is the idea of using the classification performance of the student model as a way of measuring communicative success and obliquely evaluate the generation model.

In this study, we investigate how well different cognitively-inspired neural network architectures perform in the task of perceptual category description. In particular, we investigate generation models that use prototype-based representations, models that use exemplar representations, and hybrid models that use both. Both the generation and interpretation models have two modules that are trained jointly: a classifier module and a grounded language module. The interpretation model is trained to take text descriptions of categories and produce a vector representation close to the representation learned by the classifier. The generation model is trained to take class representations and use them to generate descriptions of the corresponding class.

For the **prototype** models, we simply use the representation learned by the classifier as the class representation. For **exemplar** models, we let the model use its classifier to select the highest-scoring training image for each class and used that as the class representation. A third model used **both** of these representations by concatenating them together.

The results showed that our models were able to achieve modest communicative success, but that the interpretation model still performed better when using the ground-truth descriptions of the unknown classes (written by human annotators). In general, exemplar models achieved the highest communicative success, which suggests that the other models aren't learning to abstract visual information to the class level. In essence the exemplar model converts the task back into one that can be solved by referring to a particular image. Finally, we found that certain generation strategies resulted in poor communicative success despite generating descriptions that were more statically discriminative among the classes. This could have to do with the way those descriptions expressed the information. Perceptual grounding is not only about packing perceptual information into text, but also doing so in a way that will be understood (for example in a particular speech community or by a model trained on a particular dataset).

Author contributions Nikolai Ilinykh trained and evaluated the generation models. I trained and evaluated the interpretation models. The task of perceptual category description was developed in close collaboration by both authors. Both authors read and approved the final manuscript.

Chapter 11: Personae under uncertainty

> Noble, B., Breitholtz, E., & Cooper, R. (2020). Personae under uncertainty: The case of topoi. *Proceedings of the Probability and Meaning Conference (PaM 2020)*, 8–16

A **topos** is an unstated assumption, which is necessary for interpreting certain *enthymematic* arguments or utterances in dialogue (J.-C. Anscombre (1995); also see Section 3.4.1 of this thesis). When someone makes an utterance that requires a certain topos to be interpreted, we say that the shared topos is **evoked**, since, although it is unstated, the listener must use the topos to bridge a certain chain of reasoning that is required to understand the meaning of the utterance. Topoi operate as background assumptions and, as such, may be associated with (or even constitutive of) certain ideologies.

A **persona** is a commonly recognized archetypical *kind of person*, which, in third-wave sociolinguistics, has an important interpretation as a source of social meaning (see Section 3.4.2). When someone speaks (or dresses, or acts, etc.) in such a way that indicates their idealogical alignment with a certain persona, we say that they are **projecting** that persona. In third-wave sociolinguistics, the **indexical field** of a social signal is the "constellation of ideologically related meanings" (Eckert, 2008) that arises in virtue of the variable's relationship with one or more personae. When someone projects a persona, it is understood that this is not the *only* persona they associate with. Instead, people construct a multifarious social identity as a **bricolage** of aspects of different personae with different idealogical associations.

Topoi are particularly interesting as social signals because, whereas many sociolinguistic analyses make a clean distinction between *what is said* and *how it is said*, topoi are, first of all, not *said* at all, but rather evoked by omission. Furthermore, the social meaning is not cleanly separable from the inferential meaning, since there are situations where the evoked topos is ambiguous and the listener must rely on what they know about the speaker's social identity to infer which topos they meant to evoke.

Chapter 11 has two main goals: (1) to develop a probabilistic model of social meaning based on the indexical field, and (2) to account for the social meaning of topoi in terms of updates to the speaker's perceived bricolage of personae. We proceed by introducing two probabilistic models. Both models consider a situation where a listener, Self, updates their representation of the social identity of a speaker, Other. Both

models also associate each persona, π, with a prior distribution, φ_π, over topoi, which captures the idealogical associations of the persona.

The *first-order* model represents social identity as a categorical distribution, θ, over personae. When a speaker evokes a topos, this distribution is updated by Bayesian update based on the likelihood (computed from φ_π, with θ as a prior) that a certain persona would project that topos. This model is nice because of it simplicity, but it doesn't achieve all of our modeling goals. We can interpret θ as either Self's uncertainty about Other's (singular) persona, or a representation of Other's personae bricolage (without uncertainty), but it can't represent both without conflating the two.

In the *second-order model* we seek to address this limitation by representing Self's understanding of Other's social identity as a Dirichlet distribution, α, *over* categorical distributions of personae. Given an utterance that evokes a topos, τ, we compute the *projected persona* as the persona that maximizes the probability of τ, give the prior α and the likelihood of each π_i resulting from φ_{π_i}. We again update α by Bayesian update, this time relying on the fact that the Dirichlet distribution is a conjugate prior for the categorical distribution.

This model of interpreting of social signals in the presence of social uncertainty about the speaker can be characterized as a kind of **category adjustment effect**, something which has been observed in the interpretation and recollection of perceptual stimuli. Essentially, the effect results in stimuli being biased towards the mean of the perceptual category in which they fall. Something similar goes on in our second-order model — the social meaning we assign to a certain topos is biased based on our priors about the speaker and the persona (or personae) it is associated with.

Finally, we show how to incorporate the parameters of the second-order social meaning model in a *dialogue gameboard* (see Section 5.1.1), with the aim of modeling social meaning as resulting in incremental updates in an idealogical context. To this end, we define an information state update (based on the Bayesian update defined in the second-order model), which is licensed by the evoked topos and the projected persona. The information state update is implemented as an asymmetric merge of record types, resulting in a new dialogue game board.

Author contributions I developed the probabilistic models of social signalling and conducted the signaling game simulations. Robin Cooper and Ellen Breitholtz created the dialogue game board interpretation representation of the model and defined the information state update function. All authors read and approved the final manuscript.

Chapter 12: Conditional language models for community-level linugistic variation

Noble, B., & Bernardy, J.-P. (2022). Conditional Language Models for Community-Level Linguistic Variation. *Proceedings of the 5th Workshop on NLP+CSS at EMNLP 2022*, 59–78

Language models make use of left-to-right text context to predict the next word in a sequence. But they can make use of additional extra-linguistic context as well (consider, for example an image captioning model, which is is trained to generate text conditioned on an image). In this study we introduce **community-conditioned language models** (CCLMs) as a technique for investigating community-level linguistic variation. We experiment on a dataset of social media posts from 510 different Reddit communities.

Experiments were carried out on LSTM and Transformer language models with a word embedding layer and three stacked sequence-to-sequence layers before the final prediction layer. The CCLMs also include a community embedding layer, which is concatenated to the hidden state of the language model at 4 different layer depths (directly to the word embedding and between each of the sequence-to-sequence layers). We compare the CCLMs to vanilla language model without community information. The models are assessed according to their **perplexity**, which measures performance on the language modeling task, and **information gain**, which measures the reduction in entropy of the CCLM over its un-conditioned counterpart.

We found that almost all models benefit from community-level information, but the distribution of average information gain for messages across different communities was highly skewed right. That is, the model benefits a little from community information for the majority of communities, but a lot for a small minority of communities.

Since the conditioned language models are trained with a community embedding, we also "incidentally" learn a vector representation of communities, similar to how a neural language model with word embeddings learns a vector representation of words as a consequence of optimizing for the next-word prediction task. Since the community embeddings are optimized for the same goal, we refer to them as *linguistic embeddings*. We compare these embeddings to another embedding which is trained based on user-community co-occurrence, with no linguistic information whatsoever. We refer to this embedding as the *social embedding*.

As an initial analysis, we examine pairs of communities that are similar (with respect to the cosine similarity of their vector representations) in the linguistic embeddings *and* in the social embeddings, and pairs of communities that are similar in one but not the other. We find that we can identify pairs of communities in all three conditions: socially and linguistically similar, socially similar but linguistically different,

and linguistically similar but socially different.[1] This suggests that the while the two types of embedding *do* capture something different about the communities, what they capture is nevertheless highly correlated.

Although this initial analysis is encouraging, we are limited to comparing *pairs of communities* across embeddings, since the two vector spaces represent the embeddings differently in their axes. To solve this problem, we use orthogonal Procrustes by singular value decomposition to align the axes. We find that all of the linguistic embeddings are correlated with the social embedding to a high degree of confidence (see Section 5.3 for details).

The main results of this study are that (1) information abut which community a message came from is useful for the next-word prediction task in almost all communities and language model architectures we tested, and (2) socially similar communities are also linguistically similar, which provides further evidence for the *homophilic* hypothesis from sociolinguistics. We also make a number of qualitative observations, perhaps the most evident of which is that our models make the most use of community information for messages from communities with highly routinized patterns of interaction (such as communities centered around organizing trades of different kinds). This provides support for the idea (discussed in sec Section 3.1) that the *community of practice* is the site of linguistic convention.

Author contributions I was responsible for training the models. Jean-Philippe Bernardy developed the method for testing correlations between embeddings. Both authors were responsible for the analysis and the remaining aspects of the research. Both authors read and approved the final manuscript.

Chapter 13: Semantic shift in social networks

> Noble, B., Sayeed, A., Fernández, R., & Larsson, S. (2021). Semantic shift in social networks. *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, 26–37. https://doi.org/10.18653/v1/2021.starsem-1.3

Most work on language change, both in historical linguistics and in computational linguistics, has focused on change at the level of the macro-language and on a time scale of decades or even centuries. In this study, we turn our focus to short-term lexical change in relatively small online communities. As in the previous study, we use a corpus of Reddit comments. This time we limit our focus to 45 randomly selected sub-forums and use a diachronic corpus split into two time periods (2015 and 2017 with a one-year gap in between).

[1] Of course the vast majority of pairs of communities are dissimilar in both types of embedding.

To measure semantic change, we use a diachronic skip-gram model (Kim et al. (2014); also see Section 5.2.1 of this volume) and compute **rectified change** scores to account for the possibility that words appearing in more variable contexts will have inflated cosine change scores (Dubossarsky et al. (2017); also see Section 5.3 of this volume). We observed that naive (un-rectified) cosine change assigned high scores to discourse connectives and other words with a distinctly rhetorical function like *possibly*, *however*, and *;* (semicolon). This is consistent with the hypothesis that this metric over-estimates change for words that appear in highly variable contexts. The words recording the highest rectified change scores were much more varied across community and tended more towards nouns and verbs. We also observed that while there is an (apparently) strong (albeit non-linear) relationship between naive change and log word frequency, that relationship is not present for rectified change.

In addition to the community-level change scores, we also measured semantic change on a larger collection of Reddit comments (not restricted to any forum) over the same time period. This **generic change** score is intended to help distinguish between change that originates at the community level and change that is happening on a broader scale but reflected in the community.

We also considered word **frequency** and **change in frequency** as factors that might predict lexical change.

Next, we induced a social network graph on each of the communities in the dataset by drawing edges between users that interacted at least once in 2015. We then computed the mean **clustering coefficient** for each community (see Section 5.4 for details). We also defined a number of other community-level metrics that we thought might be correlated with semantic change, including community **size**, **stability** (overlap in active members between 2015 and 2017), and **mean posts** per member.

Finally, we performed an exploratory analysis by backward model selection on generalized linear mixed effects models to investigate the relationship between rectified change (as the response variable) and the community- and word-level features as predictors. We found a significant positive effect between change in frequency and community-level change and also between generic change and community-level change. Word frequency had a small but also significant negative effect. Among the community-level features, we found that there is a significant three-way interaction between community size, stability and clustering coefficient. In particular, in loosely-connected communities (those with low clustering), more stability among the members is correlated with more semantic change. For more densely connected communities (with average or high clustering), the positive relationship between stability and change only holds in smaller communities. For large and dense communities, the relationship between stability and change actually trends negative.

Author contributions I was responsible for training the models and computing the community-level metrics. Asad Sayeed was responsible for the GLMM exploratory

analysis. The research questions as well as the qualitative analysis and conclusions were developed in close collaboration with all the authors. All authors read and approved the final manuscript.

6.2. Conclusions

Using a variety of methodologies can make it difficult to draw direct connections from one project to the next, but it does afford us the benefit of multiple perspectives from which to make sweeping conclusions. To distinct patterns of insights emerge from the compilation.

Lexical complexity supports semantic plasticity. Words rarely, if ever, have a monolithic meaning. Whether or not it is correct to make sharp sense distinctions, it is clear that all words have a range of situations in which they can be used and that a word can carry a different meaning depending on the situation. This non-uniformity of meaning creates opportunities for lexical innovation. Words also have a range of communicative affordances. They have inferential as well as referential potential. They can be associated with other words, situations, or feelings by connotation. They can carry social meaning. Innovative uses draw on these different affordances to extend a word's range. To understand how that happens and what it means when innovations are lexicalized, (or when someone is a explicitly taught a completely new sense of a word), we can't avoid getting into the messy details of lexical structure.

- In **Chapter 7**, we saw many examples of WMNs about common words where it was clear that both participants *knew* the word, but didn't understand how it was being used in the current situation or as part of a particular construction. This could be either because the use was an innovative or because it was conventional in some community that the WMN initiator wasn't familiar with. This means that (1) other senses of the word can be used as a resource to help negotiate the meaning and (2) if the new meaning is grounded, it may only apply in situations like the one that initiated the WMN. While these dynamics were evident in the annotation study, our interaction model would have to be extended to fully accommodate them.

- We used monolithic vector representations of words in **Chapter 13**. While this allowed us to easily quantify change from one time period to the next. It did mean we were limited in what we could understand about *how* a word was changing. Similar to the triggers in Chapter 7, many of the most-changed words by community were words already present in the community's vocabulary (although increase in frequency *was* highly correlated with change). Relatedly, the distributional approach can't tell us whether changes in word representation are

a result merely of changes in the *distribution of use*, or if those changes reflect (or engender) underlying changes in the word's *meaning potential*.

- In **Chapter 8**, we showed how referential and inferential aspects of meaning can be synthesized by broadening our perspective from considering lexical items one-by-one to considering a classification system as a lexical resource from which individual lexical meanings can be derived. In **Chapter 9**, we used that structure to give an account of how genus-differentia definitions can be interpreted to create a new lexical entry that carries both inferential and referential meaning.

- In **Chapter 10**, we showed that the cognitive structure of perceptual concepts matters for how they can be described. Our models performed best when they generated textual descriptions from exemplar *instances* of a perceptual class, rather than aggregated class representations. More work is needed to understand whether the same might be true of humans, or whether there are machine learning architectures that would better model the way people represent perceptual classes.

- It is not only words that carry meaning. In **Chapter 11**, we assigned a prior over topoi to each persona, imbuing the personae with idealogical content, but also giving social meaning to the topoi by Bayesian inference. Of course, there are all sorts of indexical relationships in the world that we wouldn't necessarily want to consider as part of the lexicon (smoke *means* fire, for example). But topoi point to the fact that it is not always easy to make a distinction between the lexical and the non-lexical. This suggests lexical change is related to the more general cognitive phenomena of inference and uncertainty that govern how indexical relationships are established.

Methodologically, we can get access to new ways of understanding the process of semantic change by starting with frameworks that acknowledge the complexity of lexical structure and its implications for both compositional meaning and interaction. On the formal side, systems like TTR make it possible to represent structured lexical information. There may also be benefits to adopting a construction grammar approach to meaning, since it seems that multi-word constructions are often the site of coordination and change. Finally, modeling lexical meaning at the level of cognition can give us a more fine-grained understanding of what happens when a word's meaning potential changes in the mind of a speaker.

Community-level change stems from the interactive practices of the community. All language use takes place in a communicative context. When that context includes a particular community, lexicalization is possible. Some interactive practices (WMN, for example) are explicitly oriented towards lexicalization. In other cases,

semantic coordination is more implicit, and whether or not the coordinated meaning "sticks" (or is propagated to the community level) may depend on multitude of factors, including the communicative utility of the innovation, and whether it is compatible with existing community norms.

- Our WMN interaction model from **Chapter 7** relied on on the concept of semantic *anchors*, which highlight the importance of existing common ground when negotiating new meanings. As an interaction game, WMN itself relies on community norms about how the game proceeds, what moves are possible at different times, and how different moves should be interpreted to maintain a shared understanding of the state of the joint activity.

- In **Chapter 9** we gave an account of how an agent might update their lexicon based on a genus-differentia definition. Importantly, this account relied on community-level norms about how to classify entities in a particular domain. Sharing a classification system is a way, not only to classify for oneself, but to make it possible to teach and learn new concepts among the community.

- Similarly in **Chapter 10**, our generation model was able to successfully describe novel perceptual categories to the interpretation model. This success depended not only on an existing set of shared perceptual categories, but also on norms (implicit in the training data) about how how a bird should be described to maximize class-level discriminativity of the description.

- In **Chapter 12**, the interactive practices of the community appeared to be related to how linguistically idiosyncratic the community was. Communities with highly formulaic patterns of interaction tended to be more informative to the language model, whereas communities where interactions tended towards general conversation were less informative. Although this is *prima facie* a synchronic observation about linguistic variation, it suggests that a certain task-orientedness can serve as motivation for innovation and conventionalization.

- In **Chapter 13**, we saw that in more loosely-connected communities, stability of membership was always correlated with more change, but the same was not true for densely connected communities, especially large, dense communities. Anecdotally, it seemed that densely connected communities tend to have more extended interactions involving multiple parties. It could be that in these more intensely interactive environments, changes are more easily propagated to the community level, relying less on the pairwise common ground that is preserved by a more stable membership.

Again, these insights suggest certain methodological recommendations. An approach that centers interaction can yield a lot of new insights about language change.

If we are interested in *why* change takes place, we must go to the site of change — the particular communicative context or interaction. This is where the rubber meets the road: where we try out new semantic innovations, accommodate unfamiliar language, and learn from each other. A flexible mutable language is what gives linguistic interaction its distinctly human character. And it is in interaction that we make our mark on the language.

Bibliography

Anscombre, J. C., & Ducrot, O. (1983). *L'argumentation dans la langue*. Editions Mardaga.

Anscombre, J.-C. (1995). La théorie des Topoï : sémantique ou rhétorique ? *Hermès*, (15), 185. https://doi.org/10.4267/2042/15167

Anttila, A. (2004). Variation and Phonological Theory. In *The Handbook of Language Variation and Change* (pp. 206–243). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470756591.ch8

Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, *34*(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Auer, P. (2015). Reflections on Hermann Paul As a Usage-Based Grammarian. In P. Auer & R. W. Murray (Eds.), *Hermann Paul's Principles of Language history revisited: Translations and reflections*. De Gruyter.

Austin, J. L. (1950). Truth. *Aristotelian Society Supp*, *24*(1), 111–29.

Bakhtin, M. M. (1987). *Speech Genres and Other Late Essays* (C. Emerson & M. Holquist, Eds.; V. W. McGee, Trans.; 2nd Edition). University of Texas Press.

Barwise, J., & Perry, J. (1983). *Situations and Attitudes*. MIT Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bell, R. A., & Healey, J. G. (1992). Idiomatic Communication and Interpersonal Solidarity in Friends' Relational Cultures. *Human Communication Research*, *18*(3), 307–335. https://doi.org/10.1111/j.1468-2958.1992.tb00555.x

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bennett, M. (1976). A Variation and Extension of a Montague Fragment of English. In *Montague Grammar* (pp. 119–163). Elsevier. https://doi.org/10.1016/B978-0-12-545850-4.50010-8

Bernardy, J.-P., & Chatzikyriakidis, S. (2019). What Kind of Natural Language Inference are NLP Systems Learning: Is this Enough?: *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*, 919–931. https://doi.org/10.5220/0007683509190931

Bhattasali, S., & Resnik, P. (2021). Using surprisal and fMRI to map the neural bases of broad and local contextual prediction during natural language comprehension. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3786–3798. https://doi.org/10.18653/v1/2021.findings-acl.332

Blank, A. (2003). Polysemy in the lexicon and in discourse. In B. Nerlich, Todd, V. Herman, & C. David D. (Eds.), *Polysemy: Flexible Patterns of Meaning in Mind and Language* (pp. 267–293). Mouton de Gruyter.

Blank, H., & Bayer, J. (2022). Functional imaging analyses reveal prototype and exemplar representations in a perceptual single-category task. *Communications Biology*, *5*(1), 1–13. https://doi.org/10.1038/s42003-022-03858-z

Breitholtz, E. (2020). *Enthymemes and Topoi in Dialogue: The Use of Common Sense Reasoning in Conversation*. Brill.

Brennan, S. E., & Clark, H. H. (1996). Conceptual Pacts and Lexical Choice in Conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482.

Bücking, S. (2010). German Nominal Compounds as Underspecifiesd Names for Kinds. *Linguistische Berichte. Sonderheft*, (17), 253–281.

Campbell-Kibler, K. (2010). The sociolinguistic variant as a carrier of social meaning. *Language Variation and Change*, *22*(3), 423–441. https://doi.org/10.1017/S0954394510000177

Cann, R., Grover, C., & Miller, P. H. (Eds.). (2000). *Grammatical interfaces in HPSG*. CSLI Publications.

Chatzikyriakidis, S., & Cooper, R. (2018). Type Theory for Natural Language Semantics. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.329

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Clark, H. H., & Schaefer, E. F. (1986). Collaborating on contributions to conversations. *Language and Cognitive Processes*, *2*(1), 19–41.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Cobreros, P., Egré, P., Ripley, D., & van Rooij, R. (2012). Tolerant, Classical, Strict. *Journal of Philosophical Logic*, *41*(2), 347–385. https://doi.org/10.1007/s10992-010-9165-z

Cooper, R. (2005). Austinian Truth, Attitudes and Type Theory. *Research on Language and Computation*, *3*(2), 333–362. https://doi.org/10.1007/s11168-006-0002-z

Cooper, R. (2012). Type Theory and Semantics in Flux. In R. Kempson, T. Fernando, & N. Asher (Eds.), *Philosophy of Linguistics* (pp. 271–323). North-Holland. https://doi.org/10.1016/B978-0-444-51747-0.50009-3

Cooper, R. (2023). *From Perception to Communication: A Theory of Types for Action and Meaning*. Oxford University Press.

Cooper, R., Dobnik, S., Lappin, S., & Larsson, S. (2015). Probabilistic Type Theory and Natural Language Semantics. *Linguistic Issues in Language Technology, Volume 10, 2015*.

Cooper, R., & Ginzburg, J. (2015). Type Theory with Records for Natural Language Semantics*. In *The Handbook of Contemporary Semantic Theory* (pp. 375–407). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118882139.ch12

Coquand, T., Pollack, R., & Takeyama, M. (2003). A Logical Framework with Dependently Typed Records. In M. Hofmann (Ed.), *Typed Lambda Calculi and Applications* (pp. 105–119). Springer. https://doi.org/10.1007/3-540-44904-3_8

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

de Finetti, B. (1992). Foresight: Its Logical Laws, Its Subjective Sources (H. E. Kyburg Jr., Trans.). In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics: Foundations and Basic Theory* (pp. 134–174). Springer. https://doi.org/10.1007/978-1-4612-0919-5_10

Deane, P. D. (1988). Polysemy and cognition. *Lingua*, *75*(4), 325–361. https://doi.org/10.1016/0024-3841(88)90009-5

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dowty, D. R., Wall, R. E., & Peters, S. (1981). *Introduction to Montague semantics*. D. Reidel Pub. Co. ; sold and distributed in the U.S.A. and Canada by Kluwer Boston Inc.

Dubossarsky, H., Weinshall, D., & Grossman, E. (2017). Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1136–1145. https://doi.org/10.18653/v1/D17-1118

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, *12*(4), 453–476. https://doi.org/10.1111/j.1467-9841.2008.00374.x

Eckert, P. (2019). The limits of meaning: Social indexicality, variation, and the cline of interiority. *Language*, *95*(4), 751–776. https://doi.org/10.1353/lan.2019.0072

Fernandez, R. (2014). Dialogue. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics 2nd edition* (Second). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199573691.001.0001

Fernandez, R., & Larsson, S. (2014). Vagueness and Learning: A Type-Theoretic Approach. *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, 151–159. https://doi.org/10.3115/v1/S14-1019

Firth, J. (1957). A synopsis of linguistic theory 1930–1955. *Studies in Linguistic Analysis (special volume of the Philological Society)*, *1952–59*, 1–32.

Fox, J., & Monette, G. (1992). Generalized Collinearity Diagnostics. *Journal of the American Statistical Association*, *87*(417), 178–183. https://doi.org/10.2307/2290467

Gamut, L. T. F. (1991). *Logic Language and Meaning, Volume 2: Intensional Logic and Logical Grammar*. University of Chicago Press.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Mifflin OCLC: 1222716492.

Giglioli, P. P. (1972). *Language and social context: Selected readings*. Harmondsworth : Penguin.

Ginzburg, J. (2012). *The Interactive Stance*. Oxford University Press.

Grice, H. P. (1975). Logic and Conversation. In P. Cole (Ed.), *Speech acts* (5. print, pp. 44–55). Acad. Pr.

Gumperz, J. (1972). The Speech Community. In P. P. Giglioli (Ed.), *Language and social context: Selected readings*. Harmondsworth : Penguin.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., & Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 107–112. https://doi.org/10.18653/v1/N18-2017

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1501. https://doi.org/10.18653/v1/P16-1141

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Harris, Z. S. (1954). Distributional Structure. *Word*, *10*(2-3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hasan, R. (1989). Semantic variation and sociolinguistics. *Australian Journal of Linguistics*, *9*(2), 221–275. https://doi.org/10.1080/07268608908599422

Hasan, R. (2009). *Collected works of Ruqaiya Hasan. Vol. 2, Semantic variation: Meaning in society and in sociolinguistics*. Equinox.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hopper, R., Knapp, M. L., & Scott, L. (1981). Couples' Personal Idioms: Exploring Intimate Talk. *Journal of Communication*, *31*(1), 23–33. https://doi.org/10.1111/j.1460-2466.1981.tb01201.x

Jackson, M. O. (2010). *Social and Economic Networks* (Illustrated edition). Princeton University Press.

Johnstone, B. (1996). *The Linguistic Individual: Self-Expression in Language and Linguistics*. Oxford University Press.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*(2), 129–191. https://doi.org/10.1016/0010-0277(94)00659-9

Kennington, C., & Schlangen, D. (2015). Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 292–301. https://doi.org/10.3115/v1/P15-1029

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., & Petrov, S. (2014). Temporal Analysis of Language through Neural Language Models. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 61–65. https://doi.org/10.3115/v1/W14-2517

Kingma, D. P., Ba, J., & Amsterdam Machine Learning lab (IVI, FNWI). (2015). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*.

Kolmogorov, A. N. (1950). *Foundations of the theory of probability*. New York: Chelsea Pub. Co.

Korta, K., & Perry, J. (2008). The pragmatic circle. *Synthese*, *165*(3), 347–357. https://doi.org/10.1007/s11229-007-9188-3

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. *Proceedings of the 27th International Conference on Computational Linguistics*, 1384–1397.

Labov, W. (1963). The Social Motivation of a Sound Change. *WORD*, *19*(3), 273–309. https://doi.org/10.1080/00437956.1963.11659799

Larsson, S. (2002). *Issue-based Dialogue Management* (Doctoral dissertation). University of Gothenburg. Gothenburg, Sweden.

Larsson, S. (2013). Formal semantics for perceptual classification. *Journal of Logic and Computation*, *25*(2), 335–369. https://doi.org/10.1093/logcom/ext059

Larsson, S. (2020). Discrete and Probabilistic Classifier-based Semantics. *Proceedings of the Probability and Meaning Conference (PaM 2020)*, 62–68.

Larsson, S. (2021). The role of definitions in coordinating on perceptual meanings. *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.

Larsson, S., & Bernardy, J.-P. (2021). Semantic Classification and Learning Using a Linear Tranformation Model in a Probabilistic Type Theory with Records. *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, 14–22.

Larsson, S., & Cooper, R. (2009). Towards a formal view of corrective feedback. *Proceedings of the EACL 2009 Workshop on Cognitive Aspects of Computational Language Acquisition - CACLA '09*, 1–9. https://doi.org/10.3115/1572461.1572464

Larsson, S., & Cooper, R. (2021). Bayesian Classification and Inference in a Probabilistic Type Theory with Records. *Proceedings of the 1st and 2nd Workshops on Natural Logic Meets Machine Learning (NALOMA)*, 51–59.

Larsson, S., & Myrendal, J. (2017). Dialogue Acts and Updates for Semantic Coordination. *SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*, 52–59. https://doi.org/10.21437/SemDial.2017-6

Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, *194*(10), 3801–3836. https://doi.org/10.1007/s11229-015-0786-1

Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, *41*(5), 1202–1241. https://doi.org/10.1111/cogs.12414

Lewis, D. K. (1969). *Convention: A Philosophical Study*. Wiley-Blackwell.

Lücking, A., Cooper, R., Larsson, S., & Ginzburg, J. (2019). Distribution is not enough: Going Firther. *Proceedings of the Sixth Workshop on Natural Language and Computer Science*, 1–10. https://doi.org/10.18653/v1/W19-1101

Malt, B. C. (1989). An on-line investigation of prototype and exemplar strategies in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 539–555. https://doi.org/10.1037/0278-7393.15.4.539

Marconi, D. (1997). *Lexical competence*. MIT Press.

Mazzocconi, C., Tian, Y., & Ginzburg, J. (2022). What's Your Laughter Doing There? A Taxonomy of the Pragmatic Functions of Laughter. *IEEE Transactions on Affective Computing*, *13*(3), 1302–1321. https://doi.org/10.1109/TAFFC.2020.2994533

McGill, B. (2013). In praise of exploratory statistics.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *NIPS Proceedings*, 9.

Mills, G., & Healey, P. (2008). Semantic negotiation in dialogue: The mechanisms of alignment. *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 46–53.

Mills, G. J., & Healey, P. (2006). Clarifying spatial descriptions: Local and global effects on semantic co-ordination. *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.

Mitra, S., Mitra, R., Maity, S. K., Riedl, M., Biemann, C., Goyal, P., & Mukherjee, A. (2015). An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, *21*(5), 773–798. https://doi.org/10.1017/S135132491500011X

Montague, R. (1970). English as a Formal Language. In B. Visentini (Ed.), *Linguaggi nella societa e nella tecnica* (pp. 188–221). Edizioni di Communita.

Montague, R. (1973). The Proper Treatment of Quantification in Ordinary English. In P. Suppes, J. Moravcsik, & J. Hintikka (Eds.), *Approaches to Natural Language* (pp. 221–242). Dordrecht.

Moon, R. (2015). Multi-word Items. In J. R. Taylor (Ed.), *The Oxford Handbook of the Word* (pp. 120–140). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199641604.013.031

Morgan, J. L. (1978). Two Types of Convention in Indirect Speech Acts. In P. Cole (Ed.), *Pragmatics* (pp. 261–280). BRILL. https://doi.org/10.1163/9789004368873_010

Murphy, M. L. (2003). *Semantic Relations and the Lexicon: Antonymy, Synonymy and other Paradigms*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486494

Myrendal, J. (2015). *Word Meaning Negotiation in Online Discussion Forum Communication* (Doctoral dissertation). University of Gothenburg. University of Gothenburg.

Noble, B., & Bernardy, J.-P. (2022). Conditional Language Models for Community-Level Linguistic Variation. *Proceedings of the 5th Workshop on NLP+CSS at EMNLP 2022*, 59–78.

Noble, B., Breitholtz, E., & Cooper, R. (2020). Personae under uncertainty: The case of topoi. *Proceedings of the Probability and Meaning Conference (PaM 2020)*, 8–16.

Noble, B., & Ilinykh, N. (2023). Describe me an Aucklet: Generating Grounded Perceptual Category Descriptions. https://doi.org/10.48550/arXiv.2303.04053

Noble, B., Larsson, S., & Cooper, R. (2022a). Classification Systems: Combining taxonomical and perceptual lexical meaning. *Proceedings of the 3rd Natural Logic Meets Machine Learning Workshop (NALOMA III)*, 11–16.

Noble, B., Larsson, S., & Cooper, R. (2022b). Coordinating taxonomical and observational meaning: The case of genus-differentia definitions. *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.

Noble, B., Sayeed, A., Fernández, R., & Larsson, S. (2021). Semantic shift in social networks. *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, 26–37. https://doi.org/10.18653/v1/2021.starsem-1.3

Noble, B., Viloria, K., Larsson, S., & Sayeed, A. (2021). What do you mean by negotiation? Annotating social media discussions about word meaning. *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.

Norén, K., & Linell, P. (2007). Meaning potentials and the interaction between lexis and contexts: An empirical substantiation. *Pragmatics*, *17*(3), 387–416. https://doi.org/10.1075/prag.17.3.03nor

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *10*(1), 104–114. https://doi.org/10.1037//0278-7393.10.1.104

Partee, B. (1973). Some transformational extensions of Montague grammar. *Journal of Philosophical Logic*, *2*(4), 509–534. https://doi.org/10.1007/BF00262953

Partee, B. H. (1979). Semantics–Mathematics or Psychology? In R. Bäuerle, U. Egli, & A. von Stechow (Eds.), *Semantics From Different Points of View* (pp. 1–14). Springer Verlag.

Paul, H. (1886). *Prinzipien der Sprachgeschichte*. Max Niemeyer.

Podesva, R. J. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona1. *Journal of Sociolinguistics*, *11*(4), 478–504. https://doi.org/10.1111/j.1467-9841.2007.00334.x

Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, *7*(4), 532–547. https://doi.org/10.1016/0010-0285(75)90021-3

Ruhl, C. (1989). *On monosemy: A study in linguistic semantics*. State University of New York Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Schlangen, D., Zarrieß, S., & Kennington, C. (2016). Resolving References to Objects in Photographs using the Words-As-Classifiers Model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1213–1223. https://doi.org/10.18653/v1/P16-1115

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1–23.

Schlechtweg, D., Schulte im Walde, S., & Eckmann, S. (2018). Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 169–174. https://doi.org/10.18653/v1/N18-2027

Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, *31*(1), 1–10. https://doi.org/10.1007/BF02289451

Searle, J. R. (1975). *Indirect Speech Acts*. Brill. https://doi.org/10.1163/9789004368811_004

Sharma, D., & Dodsworth, R. (2020). Language Variation and Social Networks. *Annual Review of Linguistics*, *6*(1), 341–361. https://doi.org/10.1146/annurev-linguistics-011619-030524

Silberer, C., Ferrari, V., & Lapata, M. (2017). Visually Grounded Meaning Representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2284–2297. https://doi.org/10.1109/TPAMI.2016.2635138

Sperber, D., & Wilson, D. (2001). *Relevance: Communication and cognition* (2nd ed). Blackwell Publishers.

Stalnaker, R. (2002). Common Ground. *Linguistics and Philosophy*, *25*(5-6), 701–721.

Sutton, P. R. (2015). Towards a Probabilistic Semantics for Vague Adjectives. In H. Zeevat & H.-C. Schmitz (Eds.), *Bayesian Natural Language Semantics and Pragmatics* (pp. 221–246). Springer International Publishing. https://doi.org/10.1007/978-3-319-17064-0_10

Sutton, P. R. (2018). Probabilistic Approaches to Vagueness and Semantic Competency. *Erkenntnis*, *83*(4), 711–740. https://doi.org/10.1007/s10670-017-9910-6

Tahmasebi, N., Borin, L., & Jatowt, A. (2021). Survey of computational approaches to lexical semantic change detection. Zenodo. https://doi.org/10.5281/ZENODO.5040302

Tahmasebi, N., Niklas, K., Zenz, G., & Risse, T. (2013). On the applicability of word sense discrimination on 201 years of modern english. *International Journal on Digital Libraries*, *13*(3), 135–153. https://doi.org/10.1007/s00799-013-0105-8

Taylor, J. R. (2012). The dictionary and the grammar book: The generative model of linguistic knowledge. In J. R. Taylor (Ed.), *The Mental Corpus: How language is represented in the mind* (pp. 19–43). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199290802.003.0002

Teichman, M. (n.d.). Greg Kobele discusses mathematical linguistics https://elucidations.vercel.app/posts/transcript-episode-111/.

Traum, D. R., & Larsson, S. (2003). The Information State Approach to Dialogue Management. In J. van Kuppevelt & R. W. Smith (Eds.), *Current and New Directions in Discourse and Dialogue* (pp. 325–353). Springer Netherlands. https://doi.org/10.1007/978-94-010-0019-2_15

Tyler, A., & Evans, V. (2001). Reconsidering Prepositional Polysemy Networks: The Case of Over. *Language*, *77*(4), 724–765. https://doi.org/10.1353/lan.2001.0250

van Eijck, J., & Lappin, S. (2012). Probabilistic Semantics for Natural Language. *Logic and Interactive Rationality (LIRA)*, *2*, 11–35.

Wittgenstein, L. (2009). *Philosophische Untersuchungen =: Philosophical investigations* (G. E. M. Anscombe, P. M. S. Hacker, & J. Schulte, Trans.; Rev. 4th ed). Wiley-Blackwell.

Wright, C. (1975). On the Coherence of Vague Predicates. *Synthese*, *30*(3/4), 325–365.

Part II.

Compilation

7. What do you mean by *negotiation*? Annotating social media discussions about word meaning

Bill Noble, Kate Viloria, Staffan Larsson, and Asad Sayeed

Abstract We present a formalisation and annotation protocol for *word meaning negotiation* (WMN), a conversational routine in which speakers explicitly discuss the meaning of a word or phrase. WMN is formalised as an interaction game with a shared game board and rules for subsequent contributions, as well as a semantic update function based on the state of the game board. We develop an annotation schema based on this formalisation and present the results of annotating 150 Twitter conversations as WMNs.

7.1. Introduction

Meaningful dialogue requires some degree of alignment between participants' lexico-semantic resources. When misalignments are discovered, participants may choose to explicitly engage with the discrepancy in a metalinguistic discussion where the meaning of a misaligned word or phrase is at issue. These discussions—termed *word meaning negotiations* (WMN)—exhibit a certain structure, which we attempt to characterise and put to use by annotating WMNs collected from Twitter.

The opportunity for a WMN arises whenever a dialogue participant finds that they disagree with—or do not understand—what another speaker meant by a certain *trigger word* or phrase. They may ignore the discrepancy or silently deal with on their own (Larsson, 2010), or they may *indicate* it to their interlocutor (perhaps in the form of a clarification request). If the interlocutor responds to the indicator, a WMN has been initiated. As the WMN progresses, participants may propose, accept, reject, or raise the question of particular semantic relations between the word that triggered the WMN and other entities, which we refer to as *anchors*.

We start by discussing previous work on WMN that underpins this contribution (Section 7.2). Then, we develop the formal model of WMN, including a semantic update

rule that can be integrated in a game board model of dialogue (Section 7.3). After that, we introduce an annotation schema, based on our WMN model, and present the results of an annotation study using that schema (Section 7.4). Finally, we discuss insights into the phenomenon of WMN resulting from the annotation study and suggest avenues for future work (Section 7.5).

7.2. Background and Related Work

There is surprisingly little work on word meaning negotiation as such. WMNs, in the form of corrective feedback, have been studied as an aspect of first language acquisition (E. V. Clark, 2007). There has also been work that teaches artificial agents the meaning of novel terms based on defintions and grounded perceptual examples (Krause et al., 2014; Mohan et al., 2012). WMNs have also been studied in conversations between non-native language learners (Long, 1996; Varonis & Gass, 1985). Myrendal (2015, 2019) has taken a more in-depth look at WMNs between adult speakers, focusing on conversations in Swedish online discussion forums.

The model and annotation scheme we develop in this work builds on the structural model of Varonis and Gass (1985) and the classificatory schemas of Myrendal (2015, 2019). The semantic update fuction we define in Section 7.3.4 extends the dialogue acts proposed by Larsson and Myrendal (2017). We discuss this foundation in more depth below.

TIR model In the Trigger-Indicator-Response model, when an interlocutor recognizes a non-understanding and chooses to address it overtly, the discourse enters a "subroutine" in which participants attempt to repair the non-understanding and align their semantic common ground. These subroutines are embedded in the regular linear flow of dialogue in such a way that the current line of conversation is suspended Furthermore, WMNs may be nested if, in the course of resolving one non-understanding, another non-understanding occurs and is indicated by one of the participants.

A WMN has three key elements:

Trigger – an utterance by a speaker, S_1, that contains a lexical item resulting in non-understanding by another participant, S_2.

Indicator – an utterance in which S_2 explicitly indicates their non-understanding of the trigger.

Response – an utterance in which S_1 overtly acknowledges the non-understanding.

A trigger can occur at any point in a dialogue (e.g., in a question *or* in a response). The non-understanding is only made part of the common ground once it has been indicated

by S_2—thus, the trigger can only be identified retrospectively, with respect to its indicator. Likewise, the response refers back to the indicator: it may attempt to rectify the non-understanding, or merely acknowledge that a discrepancy was indicated.

Although the T-I-R model was developed for WMNs in a language learning context, Myrendal (2015) found it to be a good model for the initiation of WMNs in discussion forums as well.

Non-understanding vs. disagreement Myrendal (2015) categorises WMNs as those resulting from *misunderstanding* (NON), when one dialogue participant doesn't understand the meaning of a word uttered by another participant, in the context in which it was used, or *disagreement* (DIN), when a participant disagrees with how someone else used a word,(Myrendal, 2015). NONs are generally initiated with a *metalinguistic clarification request*, whereas DINs are initiated with a *metalinguistic objection*.

WMN dialogue acts Myrendal (2019) inventories types of WMN contributions, including *generic* and *specific explicifications*[1] (which we refer to as *partial definitions*), *exemplification*, *contrasting*, *metalinguistic objections* (which can be used in an ongoing WMN, as well as to initiate one), and *endorsement* (of a using a particular word in a given context).

Larsson and Myrendal (2017) propose dialogue acts based on these contribution types, and propose semantic update functions for exemplification partial definition and contrasting, that apply to the meaning of the trigger word, in the event that the dialogue act is grounded. In this paper, we expand on that work by using the act-level update functions to define an update that takes the entire WMN into account.

7.3. Formal model

The model presented in this section has a dual purpose. First, it affords the precise formulation of hypotheses about WMNs (in general or in a particular domain) that can be tested in terms of the model. Second, the model itself implies a certain structure to the phenomenon of WMNs which may, to a greater or lesser degree, capture what is observed. As is often the case, these two roles are not entirely separable: What is expressible in the model affects the hypotheses that can be tested; How well the model aligns empirically with the phenomenon it seeks to describe affects the reliability of the conclusions one can draw.

In addition to the descriptive goal, we want the model to support a semantic *update function* that computes the change in shared lexical resources resulting from a WMN (section 7.3.4). The rule we define builds on the work of Larsson and Myrendal, 2017,

[1] (see also Ludlow, 2014).

taking their dialogue act-specific rules and extending them to operate over a whole WMN.

Our model of word meaning negotiation depends on the notion of semantic *anchors* and speaker commitments to semantic *relations* between those anchors. This is motivated by the intuition that when speakers discuss the meaning of a word, they do so by triangulating it in reference to other points (or regions) of semantic space. In a successful WMN, the meaning of the word in question is "anchored" by the participants as a result of joint commitment to relations between the word and reference points (i.e., *anchors*) that *are* grounded.

When the project of aligning on meaning has started, it is not uncommon to discover that further discrepancies exist; that is, it can be that some of the anchors introduced to negotiate the meaning of the trigger word are themselves lacking semantic common ground (as in Varonis & Gass, 1985). This shouldn't be surprising: First of all, once a WMN has begun, discrepancies that might have gone unnoticed or un-remarked-upon are suddenly difficult to ignore. Furthermore, new anchors are introduced precisely *because* one of the participants thinks they have an elucidating relation to the trigger. Where one semantic misalignment exists, misalignment on related terms may be lying in wait. What makes something eligible as an anchor is not that its *meaning* is common ground and fully specified, but that it can be grounded as a shared *discourse referent*, available for participants invoke anaphorically (or by name or description) and put in relation to other anchors as well as to the trigger.

We represent a word meaning negotiation, between a set of speakers S taking place over N turns, as sequence of tuples:

$$\mathsf{WMN} = \langle s_i, A_i, R_i \rangle_{i \leq N} \tag{7.1}$$

where s_i is the speaker at turn i, A_i is the set of anchors introduced in that turn (we let $t \in A_0$ be the trigger), and R_i is the set of relations between anchors that s_i publicly commits (Asher & Lascarides, 2008) to during that turn.

7.3.1. Anchors

Once introduced, anchors are available for the remainder of the WMN, accessible by co-referring expressions, including anaphora. Thus, the set of common ground anchors at turn i is defined as the union of anchors introduced so far:

$$\boldsymbol{A}_i = \bigcup_{j \leq i} A_i \tag{7.2}$$

We let $[\![a]\!]$ denote the meaning of a, given the context of the dialogue and the semantic common ground of the speakers, without yet considering any updates resulting from the WMN.[2]

[2]Note that this interpretation, as with the negotiated meaning defined in Section 7.3.4, may be different for different

7.3.2. Semantic relations

Word meaning negotiation depends on a commonly understood set of possible semantic *relation types* between anchors, $\mathcal{R}$. In the remainder of the formalisation and in the annotation study (Section 7.4), we assume two semantic relations, *example* and *partial definition*:

$$\mathcal{R} = \{\mathsf{Exa}, \mathsf{Def}\} \tag{7.3}$$

We also make use of a set of *polarities*:

$$\mathcal{O} = \{+, -, ?\} \tag{7.4}$$

Polarity correspond to an attitude (or commitment) that speakers may express towards a given relation between two anchors. This set of polarities indicate whether a relation holds ($+$) or its converse holds ($-$), or if the matter is in question ($?$).

In the model, $R_i \subseteq \mathcal{R} \times \mathcal{O} \times \mathbf{A}_i \times \mathbf{A}_i$ is a set of semantic relations. We will write $\mathsf{R}^o(a, b)$ for $\langle \mathsf{R}, v, a, b \rangle$. For example, $\mathsf{Def}^+(a, b) \in R_i$ means that speaker s_i has publicly committed to a as a (positive) partial definition of b.

Given WMN, we can compute a speaker's current commitments. For a pair of anchors (a, b) and relation R, we consider the speaker to be committed to the most recent polarity that has been part of their public commitments. Formally, this is defined as follows:

$$\boldsymbol{R}_{s,0} = \begin{cases} R_0 & \text{if } s = s_0 \\ \varnothing & \text{otherwise} \end{cases} \tag{7.5}$$

and

$$\boldsymbol{R}_{s,i+1} = \begin{cases} \boldsymbol{R}'_{s,i} \cup R_{s,i+1} & \text{if } s = s_i \\ \boldsymbol{R}_{s,i} & \text{otherwise} \end{cases} \tag{7.6}$$

where

$$\boldsymbol{R}'_{s,i} = \{\mathsf{R}^o(a, b) \in \boldsymbol{R}_{s,i} \mid \neg\exists o'.\, \mathsf{R}^{o'}(a, b) \in R_s\} \tag{7.7}$$

Finally, we define the common ground relations at turn i as those relations to which all speakers have publicly committed:

$$\boldsymbol{R}_i = \bigcap_{s \in S} \boldsymbol{R}_{s,i} \tag{7.8}$$

speakers, since speakers can of course be wrong about what is common ground.

7.3.3. Interaction rules

Now that we have a structure for representing the state of a WMN at each turn and a way to compute what is common ground based on the history of those states, we characterise the rules of the WMN as an interaction game.

Formally, there are very few conditions on what A_i and R_i can include. Any number of anchors can be introduced in a turn, although practically the number is usually quite small (see Section 7.4.4). The main restriction on R_i is that it must not result in a cycle in s_i's public commitments; that is, $\{(a,b) \mid \mathsf{R}^o(a,b) \in \boldsymbol{R}_{i,s_i}\}$ must not contain a cycle. This means that $\boldsymbol{R}_{i,s}$, considered as a labeled directed graph, is acyclic, a condition that is necssary for the semantic update function (Section 7.3.4) to be well-defined. Intuitively it would be very strange for speakers to ground such a cycle for exactly that reason—indeed we did not see any such cycles in speaker commitments (let alone grounded cycles) in our annotation study, although the annotation protocol would have allowed it. There are three ways of contributing to R_i:

Propose (or raise) a relation For any two anchors in $\boldsymbol{A}_i$, the speaker either proposes a relation between them ($o \in \{+,-\}$) or poses the question of their relation without asserting anything one way or the other ($o \in \{?\}$).

Ground a relation The speaker makes some indication of their stance (or negative grounding) regarding a relation that another speaker has just committed to. For some $\mathsf{R}^o(a,b) \in R_{i-1}$, $\mathsf{R}^{o'}(a,b) \in R_i$, where $o, o' \neq ?$. If $o = o'$, then it is *positive grounding*, otherwise it is *negative grounding*.

Positive grounding can be accomplished more or less implicitly, though what counts as grounding may depend on the WMN type (NON or DIN), as well as other factors such as the medium of the dialogue and social context.

Answer a question Finally, for $\mathsf{R}^?(a,b) \in R_{i-1}$, s_i can add $\mathsf{R}^o(a,b)$ to R_i for any $o \neq ?$ by answering the question posed by s_i. Note that *grounding a relation* and *answering a question* don't formally add to the possible elements of R_i beyond *posing a relation*, but we characterise them separately because they usually take the form of grounding statements or polar answers which don't include explicit co-reference to an anchor. For that reason, we also annotate them differently (Section 7.4.2).

7.3.4. Semantic update

Our goal is to define a semantic update function that takes WMN as input. We define update functions that apply to the meaning of an anchor, based on a relation with another anchor, if that relation is grounded. Then, we recursively define the update for a whole WMN based on those functions in a straightforward way:

For $a \in \boldsymbol{A}_N$, let

$$\{\mathsf{R}_1^{o_1}(b_1, a), ..., \mathsf{R}_n^{o_n}(b_n, a)\} \subseteq \boldsymbol{R}_N$$

be the common round relations anchoring a at turn N. Then the semantic update given by WMN for a is defined as:

$$\begin{aligned}\Delta(a) = [&I(\mathsf{R}_1, o_1, \Delta(b_1)) \circ ... \\ &\circ I(\mathsf{R}_n, o_n, \Delta(b_n))](\llbracket a \rrbracket)\end{aligned} \tag{7.9}$$

Here, I is the interpretation of R (we assume that for a semantic relation to be common ground implies the existence of an update function):[3]

$$I = \begin{cases} \lambda x.\epsilon^o(b, x) \text{ if } \mathsf{R} = \mathsf{Exa} \\ \lambda x.\delta^o(b, x) \text{ if } \mathsf{R} = \mathsf{Def} \end{cases} \tag{7.10}$$

In essence, Δ, as defined in (7.9) applies the update implied by the semantic relations recursively on $\mathbf{R}_N$ in a straightforward way: the updated meaning of an anchor is computed by sequentially applying each its grounded relations to other anchors, with the caveat that each of those anchors should first have *their* meaning updated, if they were also negotiated as part of the WMN.

7.4. Annotation study

7.4.1. Data

We collected exchanges on Twitter that, based on search heuristics, were likely to involve WMN. In particular, we used the Twitter filtered stream API to find tweets that were in reply to another tweet and that used the indicator phrase *what do you mean by*.[4] This heuristic method is based on that of Myrendal (2015), who used similar phrases in Swedish to build a corpus of WMNs from online discussion forums. The search resulted in a total of 1783 candidate indicator tweets, collected over a 24-hour period (May 5–6, 2021).

After 48 hours (to wait for replies), we used the Twitter search API to collect the rest of the thread, retrieving tweets both upwards and downwards in the reply chain. Since the reply structure on Twitter is a tree (each tweet can be *in reply to* at most one other tweet, but can *have* multiple replies), retrieving the upwards context is easy—we just followed the replies up to the root of the thread. For the downward search (replies to the indicator), we initially look for a reply from the author that the indicator was a reply to, alternating back and forth between these two users for further replies and

[3] We let ϵ^+, ϵ^-, δ^+, and δ^- be as defined in Larsson and Myrendal (2017).

[4] We used a regular expression to allow for some variation in the exact wording (see supplementary materials for details).

taking the first reply in case there were multiple.[5] This resulted in 671 threads with at least one reply after the candidate indicator (38% of threads), of which we randomly sampled 150 for annotation.

7.4.2. Annotation protocol

The annotation protocol, which was developed over a series of pilot studies, aims to be comprehensible for annotators with no linguistic background (see the annotation guide in the supplementary materials). In the pilot studies, small sets of data collected from Twitter were manually annotated using initial drafts of the annotation schema by two annotators (both with a linguistic background). Error analysis sessions were conducted in order to discuss and clarify unclear definitions and inconsistent judgments between annotators. The schema was then refined based on these discussions.

Two additional annotators were added to annotate more data, which we report on in Section 7.4.4. All four annotators are linguists familiar with WMNs. As in the pilot studies, an error analysis was conducted, which we discuss in Section 7.4.5).

Annotators were shown text of the tweets, one thread at a time, in the BRAT annotation tool (Stenetorp et al., 2012). Tweets were separated by a header that included the time of the tweet and the username of the tweet author. We displayed a maximum of 10 context tweets on either side of the candidate indicator.

Annotators were instructed to read the Twitter threads and select and classify text spans as different components of a WMN——as well as to determine whether or not an exchange as a whole was in fact a WMN. The four main points of interest, meant to be evaluated in order, during annotation were the WMN Type, Trigger spans, Anchors (Examples and Definitions), and instances of Grounding. While it was recommended that annotators examine these four points in order, we noted that it is completely acceptable and sometimes necessary to go back and forth to gain a better understanding of the thread.

WMN Type The search phrase (e.g., *what do you mean by*) was automatically prelabeled as an Indicator to help the annotator find the intended focus of the example. Annotators were instructed to tag the Indicator span with the WMN Type of the dialogue as a whole. WMN Type consists of two decision points: First, the annotator must decide whether the thread is a WMN or not. If it *is* a WMN, it must then be classified as a non-understanding (NON) or disagreement (DIS).

Trigger The second task is to identify the word or phrase in question as the Trigger. Annotators must also label every other instance of the Trigger in the discussion, includ-

[5] This is a somewhat brittle heuristic that could improved upon. For example, it breaks if a user makes a "double reply" or if the conversation is between more than two users.

ing anaphoric references. It is not necessary to link Triggers together with co-reference relations since it is implied.

Anchors The next step is to find the Trigger's Anchors and to distinguish between an Anchor's two types, Examples and Definitions. Relations are annotated with a link between the anchor and the Trigger o another Anchor, and the are marked with the polarity of the relation. An Anchor can also appear multiple times within a WMN, including anaphoric reference. In this case, these anchors are linked together using the co-reference relation. Annotators are instructed to try and leave negations out of the anchor and instead annotate the relationship as having negative polarity. When linking anchors, it is important which instance of the Anchor the link originates from, since this indicates which speaker is making the commitment and when. It is recommended that annotators use their best guess when identifying whether or not a URL (which could be an image or external link) is an Anchor and if so, its type based on the textual context.

Grounding Spans of text that explicitly state the speaker does (or does not) understand or agree with the previously offered example or definition must be annotated as Grounding. This span must be linked to the Anchor it refers to. The polarity link of a Grounding statement can be either positive or negative and between an Anchor and a Trigger or between two Anchors. In a non-understanding WMN, a grounding statement with a positive link indicates that the speaker understands the proposed relationship between the Anchor and the Trigger (or another Anchor). A negative link indicates that the speaker does not (or may not) understand the proposed relationship between the Anchor and the Trigger. In a disagreement WMN, a positive link indicates that the speaker agrees with or has adopted the proposed relationship between the Anchor and the Trigger. With a negative link, the grounding statement indicates that the speaker does not agree with or has not adopted the proposed relationship between the Anchor and Trigger.

7.4.3. Post-processing annotations

There are some discrepancies between the annotation schema and the WMN formalisation described in Section 7.3, mainly due to the fact the formalisation is comprised of abstract semantic units, while the annotation is performed directly on the surface form of the WMN.

Text spans annotated as an *Anchor* (Example or Definition) were divided into equivalence classes, based on the co-reference annotations, which constitute the set of anchors in the formalisation. Spans annotated as Trigger were assumed to co-refer and the set of Triggers also constitutes an Anchor.

Relation type (Exa or Def in the formalisation) is coded as property of anchors in the annotation schema. In the pilot studies, we found that it was easier to decide the relational role of the anchor span before determining the polarity and target anchor. It is also more visually legible to separate the relation type (indicated by the color of the anchor span) and polarity (indicated by the color of the relation arrow). In theory, it would be possible for an anchor have multiple relational roles (imagine, for example, a WMN in which *insect* is used as both a partial definition of a *locust* and as an example of an *invertebrate*), but in practice this seems to be vanishingly rare (we have never observed it).

7.4.4. Results

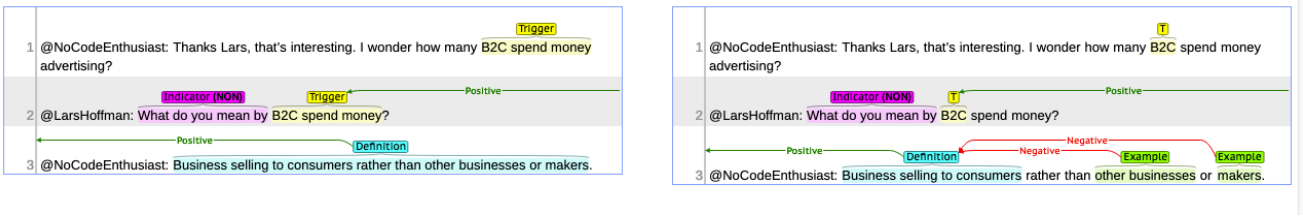


Figure 7.1.: Annotations from two annotators, showing disagreement in the extent of the trigger phrase and anchor structure.

In this section we report the results of the annotation study, particularly inter-annotator agreement.

We measured annotator agreement at two levels of description: the surface-form annotation, and then on the formal WMN representation extracted from the annotations. For agreement statistics, we report the proportion of agreed-upon items (A_0), as well as Cohen's kappa (κ) and Scott's pi (π).[6] Cohen's kappa computes expected agreement (the denominator) using annotator-level priors for the label distribution, whereas Scott's pi assumes a uniform distribution across annotators. Significantly higher κ compared to π would suggest that annotators have different priors for the category labels (Artstein & Poesio, 2008), but we don't observe that to be the case in any of the agreement statistics we measured.

First, we measured agreement on the dialogue level, namely, the classification of whether or not the dialogue was a WMN and if so, what type. Agreement was above chance, but (Table 7.1) with a substantial amount of disagreement. We discuss potential sources of disagreement in Section 7.5.

We measured agreement on span type at the token level. Tokenisation was performed post-hoc—annotators selected spans from the raw character-level text—but we consider a token to be part of a span if a majority of characters in the token overlap with it. This eliminates any artificial disagreements caused by, for example, missing

[6] A_0 is the numerator for both κ and π.

	A_0	π	κ
WMN/Not	0.71	0.40	0.40
NON/DIN	0.79	0.47	0.48

Table 7.1.: WMN type agreement. *WMN/Not* measures agreement on whether or not the dialogue was a WMN, while *NON/DIN* (restricted dialogues both annotators agreed were WMNs) measures agreement on whether the WMN resulted from *non-understanding* or *disagreement*.

the final letter in a word when selecting a span. We also consider it to be more representative than character-level agreement, which would be biased by longer words.[7]

We found a moderate level of agreement on all span types except *grounding* (Table 7.2). Error analysis suggests that this may be primarily due to how much of a tweet the annotator considered to be a part of the grounding span. Additional guidance on this point in the annotation guide may help to raise the level of agreement.

	A_0	π	κ
Anchor	0.93	0.59	0.60
Trigger	0.98	0.63	0.63
Grounding	0.98	0.22	0.22
Overall	0.87	0.64	0.64

Table 7.2.: Token-level span type agreement. *Anchor* (both Definition and Example are considered *Anchor* here), *Trigger*, and *Grounding* only consider the binary choice of whether or not a token is of that type. *Overall* considers all three possibilities together.

At the level of the formal WMN representation, we are interested in whether annotators agree on whether and what kind of relations between anchors participants commit to at each turn, and when they explicitly indicate grounding of those relations. Computing agreement for relations and grounding requires that we align the anchors identified by the two annotators. For this, we take the bijection that maximizes token-level overlap of the spans associated with the anchors. This anchor mapping aligned an average of 89.1% (σ=19.2%) of anchors per dialogue (that is, on average 10.9% of anchors had no counterpart in the other annotation).

For relations, we considered each *potential* relation at each turn; that is, for turn i, we consider each pair of anchors (including the trigger), $\{(a, b) \in \boldsymbol{A}_i \times \boldsymbol{A}_i \mid a \neq b\}$ (with the caveat that $\boldsymbol{A}_i$ only includes *aligned* anchors, since there is no possibility for agreement on unaligned anchors). Annotators agree if they both created a relation

[7] We used the NLTK (v.3.6.2) regex-based TweetTokenizer.

(with the same relation type and polarity) from an a-span originating in turn i to an b-span (regardless of where)—or if they both created no relation at all for that pair. As with the token-level statistics, A_0 is quite high, since relations are sparse, relative to all the opportunities for a relation to be created, but the chance-adjusted scores are also reasonably high (Table 7.3).

For grounding, for each turn i (starting with $i = 1$), we considered each aligned anchor that both annotators agreed was mentioned in turn $i - 1$. Annotators agree if they both thought that the current speaker grounded (with the same polarity) a relation originating in that anchor—or if they both thought no such grounding occurred. Agreement is lower than for anchor relations, but still well above chance (Table 7.3).

	A_0	π	κ
Relation	0.93	0.69	0.69
Grounding	0.88	0.58	0.59

Table 7.3.: Turn-level agreement on relation type and grounding polarity for possible relations and grounding.

7.4.5. Error analysis

After annotating the examples, we conducted some post-hoc discussions in which the annotators attempt to ascertain the reason for certain discrepancies. Based on these discussions, we make suggestions for improvements to the annotation protocol, which should aid in future efforts to annotate WMN. Further observations about the phenomenon of WMN, which came to light in these conversations, can be found in Section 7.5.

WMN Type The phrase *what do you mean by* is often used in a rhetorical way (i.e., not as a genuine question or clarification request), but it can be difficult to determine whether the speaker's objection to using a word to describe some situation under discussion is a disagreement about the meaning of the word (DIN) or a disagreement about the nature of the situation under discussion (not a WMN). The decision could be clarified by emphasizing the *results* of the indicator phrase: Does the meaning of the word subsequently become at-issue? When non-understanding or disagreement is indicated but no meaning negotiation results, this is typcially not considered a WMN (Myrendal, 2015; Varonis & Gass, 1985), but giving such "declined WMNs" their own category could result in better agreement.

Anchor spans Analysis revealed two kinds of discrepancy in anchor spans: (1) where the annotators disagreed on whether something was an anchor, or how much

of the text referred to the anchor (reflected in token-level agreement, Table 7.2), and (2) where the annotators disagreed on whether something was one anchors or two (reflected mainly in the failure to find a bijection between the two annotated sets of anchors).

A particularly notable discrepancy of the first kind involves the extent of the trigger phrase, since the speaker will sometimes repeat some context around the trigger to help locate it in the previous utterance. This can raise the question of how much of what they repeated is context and how much is the trigger. One strategy for annotators could be to observe what is *actually negotiated* subsequent to the indicator, although this too can be ambiguous.

Another common discrepancy was that one annotator would annotate multiple anchors, where another would find only one (see Figure:7.1).

Relation types While agreement on relation type (annotated as anchor span type) was fairly good, there were a few cases where adding more relation types could improve clarity. *Contrasting* is a common pattern in WMNs where the trigger word is compared to an alternative that the speaker thinks better describes the situation under discussion (Myrendal, 2019): *x is really more of a Y than a Z*. In the annotation guide, we suggested such examples be annotated with two relations: $\mathsf{Exa}^{+}(Y, x), \mathsf{Exa}^{-}(Z, x)$, but it could also be its own ternary relation that is interpreted using δ and ϵ, as in Larsson and Myrendal (2017).

7.5. Discussion and conclusion

We conclude by offering some observations on the WMNs in our Twitter corpus, and discussion on the implications these observations may have for negotiated meaning more broadly.

Speaker meaning/token meaning As mentioned in Section 7.4.5, it was often unclear whether *what do you mean by X* was asking what the speaker understands *X* to mean in general, or what they were *using X* to mean in a particular context.[8] This is perhaps related to the phenomenon where the indicator repeats a whole sentence, but the negotiation focuses on one word or short phrase: Since questions about sentence meaning are necessarily about speaker meaning, including the sentence in the indicator may clarify that the question is about speaker meaning. H. H. Clark (1996)'s hierarchical grounding schema, makes the distinction between grounding on the level of *signal meaning* and grounding on the level *uptake* (speaker meaning or illocutionary act). When a WMN is focused on resolving a non-understanding (NON), the issue can be either with the signal meaning or with uptake, however a disagreement (DIN) about

[8] See also: Myrendal (2019) *general* versus *specific explicifications*.

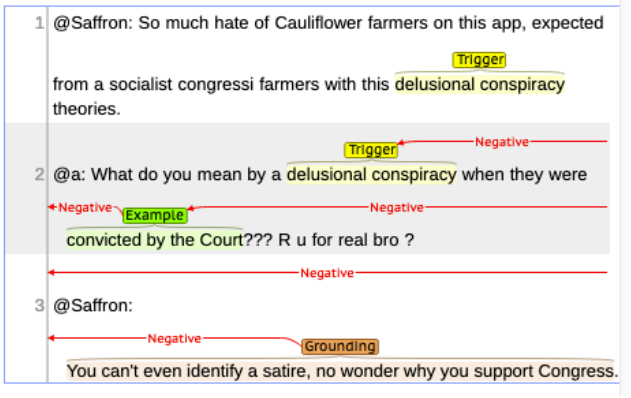


Figure 7.2.: Post-hoc annotation provided by an annotator familiar with Indian social media political discourse. The original annotators of this example, lacking the background knowledge, had different interpretations.

how a word is used is necessarily a disagreement about its *meaning potential* (Linell, 2009)—it doesn't make sense to disagree *that* someone meant something, only *how* they went about meaning it.

Social and cultural context Many of the WMNs in our corpus involved politically or socially controversial topics and the moves made by the participants often required some understanding of the social context in which the conversation was taking place. Consider the example in figure 7.2: Interpreting *convicted by the court* as providing a negative example of a *delusional conspiracy*, requires understanding the role of *conspiracy* in Indian political discourse, what *Congress* refers to (a political party) and even the political alignment implied by *Saffron* in one of the usernames.

Agreement and reliability As the cultural context example demonstrates, annotator disagreement doesn't *necessarily* imply that the annotation schema is incorrect or doesn't reflect the underlying phenomenon. In that case, one of the annotators lacked the context to interpret the WMN correctly, but it is possible for WMNs to be ambiguous (open to multiple possible interpretations), even when both have sufficient background knowledge. Reflecting these different interpretations can make this formalisation a useful tool for analysis, just as first-order logic is a useful tool for analysing certain classes of ambiguous sentences.

Taking that for granted, and considering our somewhat mediocre annotator agreement scores, what can we conclude about this formalisation and annotation schema? Is it in some sense *correct*? The only way to know is probably to continue using it (and where possible, improve upon it)—to carry out further annotation studies on conversational data from different sources, formulate and test hypotheses, and eventually attempt to train artificial agents capable of WMN.

As explicit meta-linguistic discussions, WMNs have potential as window into the processes of semantic alignment, acquisition, and change more generally. By modeling WMNs, we hope to develop conceptual frameworks that apply to the dynamics of lexical semantic resources more broadly.

References

Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, *34*(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Asher, N., & Lascarides, A. (2008). Commitments, Beliefs and Intentions in Dialogue. *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.

Clark, E. V. (2007). Young Children's Uptake of New Words in Conversation. *Language in Society*, *36*(2), 157–182.

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Krause, E., Zillich, M., Williams, T., & Scheutz, M. (2014). Learning to Recognize Novel Objects in One Shot through Human-Robot Interactions in Natural Language Dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, *28*(1).

Larsson, S. (2010). Accommodating innovative meaning in dialogue. *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.

Larsson, S., & Myrendal, J. (2017). Dialogue Acts and Updates for Semantic Coordination. *SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*, 52–59. https://doi.org/10.21437/SemDial.2017-6

Linell, P. (2009). *Rethinking Language, Mind and World Dialogically : Interactional and contextual theories of human sense-making*. Information Age Publishing.

Long, M. H. (1996). The Role of the Linguistic Environment in Second Language Acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press.

Ludlow, P. (2014). *Living words: Meaning underdetermination and the dynamic lexicon* (First edition). Oxford University Press
OCLC: ocn881131932.

Mohan, S., Mininger, A., Kirk, J., & Laird, J. E. (2012). Learning Grounded Language through Situated Interactive Instruction. *2012 AAAI Fall Symposium Series*.

Myrendal, J. (2015). *Word Meaning Negotiation in Online Discussion Forum Communication* (PhD Thesis). University of Gothenburg. University of Gothenburg.

Myrendal, J. (2019). Negotiating meanings online: Disagreements about word meaning in discussion forum communication - Jenny Myrendal, 2019. *Discourse Studies*, *21*(3), 317–339. https://doi.org/10.1177/1461445619829234

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). Brat: A Web-based Tool for NLP-Assisted Text Annotation. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102–107.

Varonis, E. M., & Gass, S. (1985). Non-native/Non-native Conversations: A Model for Negotiation of Meaning. *Applied Linguistics*, *6*(1), 71–90. https://doi.org/10.1093/applin/6.1.71

8. Classification systems: Combining taxonomical and perceptual lexical meaning

Bill Noble, Staffan Larsson, and Robin Cooper

Abstract Lexical meaning includes both perceptual and logical aspects. We present a method for combining a taxonomy with perceptual classifiers, and show that in the few-shot setting, it out-performs other methods of injecting taxonomical information in image classification. We use this method to define witness conditions for types in a rich type system with probabilistic type judgments and suggest how such a type system can be used as the basis for a new type of hybrid NLU architecture.

8.1. Introduction

For words like *red*, *apple*, and *hug*, part of what it means for a person—or indeed an artificial NLU system—to understand the word's meaning is the ability to recognize that some object is red, or an apple, or that some event is one in which hugging is taking place. Marconi (1997) calls this **referential competence**. Another mode of understanding is supported by **inferential competence**, which has to do with the relationship that certain lexical items have with one another—a system that infers that John is not married from the sentence *John is a bachelor* demonstrates inferential competence with the words *bachelor* and *married*. Marconi (1997) argues that neither of these competencies are reducible to the other, meaning that a comprehensive theory of lexical meaning must explain both referential and inferential ability.

In this paper, we propose a framework for combining taxonomical information, which supports an inferential competence, with perceptual classifiers, which implement referential competence. This *classification system* is formalized in a rich type theory with probabilistic type judgments, meaning it can be integrated in a formal semantics based on Type Theory with Records (Cooper et al., 2015).[1]

[1] A PyTTR implementation of a classification systems based on convolutional visual classifiers is available online here: `https://github.com/GU-CLASP/classification-systems`. We also make available the code for the experiments conducted in Section 8.5.

8.2. Classifier-based perceptual meaning

While distributional methods of representing meaning have achieved a lot of success, many have argued that that relying on exclusively *ungrounded* meaning representations has fundamental limitations (Bender & Koller, 2020; Bisk et al., 2020; Harnad, 1990).

Classifier semantics offers a way to ground lexical meaning, operating on the intuition that part of what it means to understand the meaning of a word is to be able to identify instances of it based on perceptual input.

In one approach to classifier semantics (e.g., Schlangen et al., 2016; Silberer et al., 2017), the parameters of a learned classifier (for example, the relevant row of a liner classifier's weight matrix) are regarded as a distributed representation of the meaning of the word. Alternatively, it is possible to regard the classifier itself, as a function of type $f : \mathit{PerceptualData} \rightarrow [0, 1]$, that provides the semantics of the relevant word (e.g., Larsson, 2020a). Here, both the parameters of the classifier and the classification algorithm are considered to be part of the perceptual meaning, whereas in the distributed approach, the classification algorithm is simply a means by which a distributed representation is learned.

In this work, we take a functional approach to classifier semantics. Because they can (at least for one-place predicates) be considered analogous to Montague's $e \rightarrow t$ type, it is natural to integrate classifiers-as-functions in a type-theoretic approach to compositional meaning. Furthermore, classifiers have the nice theoretical property that they can distinguish between intentional identity and extensional equivalence (Lappin, 2012; Larsson, 2020b; Muskens, 2005).

A *multi-class classifier*, C for a set of labels, L is a function that takes an input and produces a prediction among the labels in the form of a probability distribution. We will consider multi-class classifiers that take perceptual data as input:[2]

$$C : \mathit{PerceptualData} \rightarrow (L \rightarrow [0, 1]),$$

subject to the restriction that for any input a, $\sum_{l \in L} C(a)(l) = 1$.

8.3. Folk taxonomies

A *folk taxonomy* is a hierarchically structured collection of conceptual categories that is *common ground*, in the sense of Clark (1996), in a certain speech community. We wish to invoke a more general notion than that of scientific or technical taxonomies that rely on an authoritative reference for their common ground status. Folk taxonomies by contrast can be informal, emerging from the communicative needs of a particular

[2] In the remainder of the paper, we restrict our attention to classifiers and taxonomies of individuals, so we assume that *PerceptualData* is of a kind that corresponds to entities of type *Ind*. In general, however, we can also classify other kinds of entities (events, relations between individuals, etc.).

community and changing in response to changes in the environment. Such a taxonomy can also be established in an *ad hoc* way between a group of speakers, grounded in a particular interaction.

For now, we define a taxonomy in set theoretic terms. A taxonomy takes the form

$$\text{Tax} := \langle \text{Taxon}, \text{Set}(\text{Set}(\text{Tax})) \rangle,$$

where Taxon is the label for a taxonomical category. A taxonomy bottoms out in pairs of the form $\langle \text{Taxon}, \varnothing \rangle$, which we refer to as *leaf taxons*.

Notice that the second element of Tax is a set of *sets* of taxonomies. To see why this is, we will first introduce the notion of a *distinction*, which is a pair that takes the following form:

$$\text{Dist} : \langle \text{Taxon}, \text{Set}(\text{Taxon}) \rangle.$$

Consider this taxonomy:

$$\begin{aligned}
&\langle \textit{object}, \{\{\langle \textit{animal}, \{ \\
&\{\langle \textit{mammal}, \{...\}\rangle, ..., \langle \textit{bird}, \{...\}\rangle\}, \\
&\{\langle \textit{herbivore}, \varnothing\rangle, \langle \textit{omnivore}, \varnothing\rangle, \langle \textit{carnivore}, \varnothing\rangle\}\}, \\
&\langle \textit{vegetable}, \{...\}\rangle, \langle \textit{mineral}, \{...\}\rangle\}\}\rangle
\end{aligned}$$

Here, *animal* is subject to two distinctions: the distinction based on diet, and the one that categorizes animals as *mammal*s, *bird*s, and so on.

In the following, we let $\textit{dist} : \text{Tax} \rightarrow \text{Set}(\text{Dist})$ be the function from a taxonomy to its distinctions.

A *genus-species* relation holds between a taxon and the (first component of) an element of one of its distinctions. In the above example, both *mammal* and *herbivore* are species of *animal*.[3] Conceptually, the key feature of a distinction is that it implies an exhaustive partition of the genus into a set of mutually exclusive species. Note however that we need not assume every species is associated with a lexical item—there can, for example, be a catch-all species in cases where the named alternatives don't cover the entire genus.[4]

This leaves us with two main desiderata for when we start giving content to our taxonomy in the next section.

1. An instance of a species is an instance of its corresponding genus.

2. An instance of a genus is an instance of exactly one species in each of its distinctions.

[3] For word senses, this is referred to as a *hypernym-hyponym* relation.

[4] Generally we would expect a conventionalized taxonomy to make distinctions in a systematic way; that is, where the species within a distinction are differentiated along some common dimension or set of dimensions. This intuition can be traced back at least to Aristotle's *Categories*. However, his is not a formal requirement of a taxonomy at this stage and nor could it be since, taxons are not yet associated with any kind of content that could be considered as features or establish differentia. Such content will come by way of classifiers in Section 8.4.

8.4. Classification systems

By associating a word with a prediction class of a classifier, a system can be endowed with at least some referential competence. Similarly, associating a word with taxon gives a system some inferential competence in relation to other words embedded in the taxonomy. In this section, we describe a *classification system*, which combines classifiers and a taxonomy to integrate these two kinds of competence.

With this in mind, we will formalize a classification system as a rich Martin-Löf (1984)-style type system that allows for probabilistic type judgments (as in Cooper et al., 2015). Furthermore, we will assume that we can provide basic types with *witness conditions* that ground type judgments. From the perspective of an agent, a type's witness conditions are the methods by which an agent may judge something to be of that type (Cooper, forthc).

Suppose we have a taxonomy $\mathbf{T}$, and a classifier, C_d, for each distinction $d \in \mathit{dist}(\mathbf{T})$. For each taxon, t, in the taxonomy, we want to define a type, T_t, with the appropriate witness conditions such that $p(a : T_t)$ estimates the probability that a belongs to the taxon, according to the classifiers.

Intuitively, the classifiers give content to the distinctions of the taxonomy by *distinguishing* between species. The classifier is thus premised on the assumption that the object of classification certainly belongs to *one* the species, s_i, among which it distinguishes, meaning that it must in turn belong to the associated genus, g. In practice, this means that the classifier for a given distinction is trained on the subset of labeled data from the associated genus. The classifier's prediction, $C_d(a)(s_i)$, can thus be interpreted as the conditional probability that a has belongs to s_i, given that it belongs to g.

There is one taxon in the taxonmy—the root taxon—that is not a species in any distinction. Let T_{t^*}, which we will refer to as the *domain* classification system, be the type associated with the root taxon. We will assume that T_{t^*} is *universal* in the sense that it is witnessed by any object:[5]

$$p(a : T_{t_*}) = 1 \tag{8.1}$$

Every other taxon is a species in some distinction, meaning that we have a classifier associated with it. Let $d = \langle g, \{s_1, ..., s_n\}\rangle \in \mathit{dists}(\mathbf{T})$ be a distinction. We define auxiliary types, $T'_{s_1}...T'_{s_n}$ with witness conditions as follows:

$$p(a : T'_{s_i}) = C_d(a)(s_i). \tag{8.2}$$

That is, an object a is judged to be of type T'_{s_i} with probability equal to the probability assigned by the classifier for the corresponding distinction.

[5]This assumption is convenient for simplicity, but it also works if T_{t_*} is given some constant prior or well-defined witness conditions as part of some larger type system in which the classification system is embedded.

The interpretation of the classifier as providing a conditional probability suggests that we should define T_{s_i} such that:[6]

$$p(a : T'_{s_i}) = p(a : T_{s_i} \mid a : T_g) \tag{8.3}$$

We also want T_{s_i} to satisfy the desiderata from the end of Section 8.3, which can be restated as follows:

$$p(a : T_{s_i}) \leq p(a : T_g) \tag{8.4}$$

and

$$p(a : T_{s_i} \mid T_g) = 1 - \sum_{j \neq i} p(a : T_{s_j} \mid a : T_g) \tag{8.5}$$

With this in mind, we let the witness conditions for T_{s_i} be defined as the product of the probability assigned to T'_s and T_g:[7]

$$p(a : T_{s_i}) = p(a : T'_{s_i}) \cdot p(a : T_g) \tag{8.6}$$

By induction on the taxonomy and the base case of T_{t*}, this gives us well-defined witness conditions for for every taxon t.

Briefly, we will show that this definition meets each of our desiderata. In the following, let $\langle g, \{s_1, ..., s_n\}\rangle$ be a distinction. Without loss of generality, we consider the case of T_{s_i}.

We get (8.4) directly from (8.6), since $0 \leq p(a : T'_{s_i}) \leq 1$. As a result of (8.4) we may write $T_{s_i} \sqsubseteq T_g$—i.e., that T_{s_i} is a *subtype* of T_g (Cooper et al., 2015). Furthermore, this has the consequence that

$$p(a : T_g | a : T_{s_i}) = 1 \tag{8.7}$$

From Bayes Theorem and (8.7), we can prove (8.3):

$$\begin{aligned} p(a : T_{s_i} \mid a : T_g) & \\ &= \frac{p(a : T_g \mid a : T_{s_i}) \cdot p(a : T_{s_i})}{p(a : T_g)} \\ &= \frac{p(a : T_{s_i})}{p(a : T_g)} \\ &= \frac{p(a : T'_{s_i} \cdot p(a : T_g)}{p(a : T_g)} \\ &= p(a : T'_{s_i}) \end{aligned}$$

[6] This corresponds to the probability that a is of type T_{s_i} given that it is of type T_g, though other notions of conditional judgments are possible in probabilistic type theory. See Larsson and Cooper (2021).

[7] Note that T_{s_i} has different witness conditions from that of the meet type $T'_{s_i} \wedge T_g$, as defined in Cooper et al. (2015), since the witness condition for the meet type is defined by the classical Kolmogorov (1950) equation for conjunction:

$$p(a : T'_{s_i} \wedge T_g) = p(a : T'_{s_i}) \cdot p(a : T_g \mid a : T'_{s_i}),$$

which is different since we can't assume that $C_d[s_i]$ is probabilistically independent from $C_{d'}[g]$, where d' is the distinction of which g is a species.

	Precision	Recall	F1
per-distribution	**0.93**	**0.90**	**0.90**
marginalization	0.90	0.86	0.82
hierarchy-agnostic	0.80	0.84	0.81

Table 8.1.: Macro-averaged precision, recall, and F1 score for the three methods of incorporating hierarchy in classification.

Finally, (8.5) follows from (8.3) and the fact that $\sum_{i \leq n} C_d(a)(s_i) = 1$.

8.5. Empirical comparison

To investigate how well the classification system performs in practice, we compare it with two other plausible methods of combining classification with taxonomical hierarchy. We put aside type theory for the moment and make a comparison based on metrics that are traditionally used for machine learning classification.

Dhall et al. (2020), proposes several possible methods of incorporating hierarchical information, including the *hierarchy agnostic* and *marginalization* methods that we compare against.[8]

The **hierarchy agnostic** method is the simplest and most common way of dealing with a taxonomically organized label set. Every label is considered by a single *multi-label* classifier, without respect to taxonomical hierarchy. There is thus no guarantee that the predicted probabilities will be consistent—the probability assigned to a genus label could be lower than the probability assigned to one of its species, for example. Hopefully the hierarchical relations inherent in the data encourages the classifier to learn a function that approximates the taxonomy.

In the **marginalization** method, a *bottom-up* classifier, is trained on the leaf nodes in the taxonomy. Labels at higher levels are predicted by marginalizing the leaf node probabilities—the probability of a genus label is computed as the sum of the probability of its species labels. Note that this method assumes that the leaf labels are disjoint, meaning that it only works for taxonomies in which there is on distinction per genus.

The system described in Section 8.4 is will be referred to as the **per-distinction** method. As described there, we train a classifier for each distinction and compute the probability of a given label as the product of the classifier output and the probability assigned to its parent label.

We test each method on a simple synthetic dataset shapes with different colors and sizes. The data was generated with a hierarchical stochastic process reflected in the

[8] Dhall et al. (2020) also tests a *per-level* and *masked per-level* method, which are arguably most similar to what we propose here. We do not reproduce those tests because marginalization tended to out-perform them in Dhall et al. (2020)'s experiments. Like marginalization, the per-level and masked per-level methods assume that there is a single distinction per genus.

taxonomy of the labels given to each item. Images were encoded with a convolutional autoencoder, which was pre-trained on images from a larger unstructured sample space.

Each method used simple single-layer linear classifiers trained by stochastic gradient descend through backpropagation. The marginalization and per-distinction classifiers use softmax activations with categorical cross-entropy as the loss function, and the hierarchy agnostic classifier uses a sigmoid activation and binary cross entropy with the indicator function of the item's actual label set. Table 8.1 gives a summary of the results of the classifiers in a few-shot classification scenario with 5 training instances and 100 testing instances for each leaf label. A separate set of 100 development items were used to choose the best model after 10 epochs of training. For the precision, recall and F1 metrics, the predicted classes were chosen in a greedy fashion from the top of the taxonomy, taking the label with the highest probability consistent with the label chosen at the previous level.

Consistent with Dhall et al. (2020), we find that both methods that explicitly take the label hierarchy into account out-perform the hierarchy agnostic method. In the few-shot experiment reported here, our per-distribution method performed best, though we note that this advantage is less pronounced with more training examples.

8.6. Conclusion

In this paper we have focused on the problem of integrating perceptual and logical meaning on a lexical level. To do this, we have embedded perceptual classifiers as witness conditions for types in a type system that respects a taxonomical structure. Our method for doing this is based on the intuition that such a taxonomy gives rise to a collection of *distinctions*, whose content can be defined by multiclass classifiers. We have compared our method of embedding classifiers at each node in a taxonomy to other strategies for classifying in a taxonomically structured label space suggested by (Dhall et al., 2020). Future work should also consider the possibility of learning the label hierarchy on the fly, as Bengio et al. (2010) does. Embedding such a hierarchy in a type system may present additional challenges, but allowing for changes to the taxonomy would be necessary to full model the plasticity of the lexical semantic structures used by natural language speakers.

We have also left open the looming question of compositional semantics. We presented classification systems as a rich type system in order to suggest a way forward in this regard. Our proposal is compatible with Type Theory with Records (TTR), which can be used to define version of compositional semantics (Cooper et al., 2015). Indeed, TTR has been used for compositional semantics with perceptual classifier-based meaning (Larsson, 2013, 2017). The issue remains, however of how to compose the types we define in Section 8.4.

Composing classifiers-as-functions is no easy task.[9] For a given object a, one can

[9]Importantly, it is a different task from the one of composing distributed representations learned through classification.

compute the probability that a witnesses both T_1 and T_2 simply by taking judgments for T_1 and T_2 separately. The difficulty comes when one needs to reason hypothetically, as is necessary in NLI. What is the likelihood that *some* object of type T_1 is also of type T_2? One way forward is to find a way to compose the classifiers for T_1 and T_2 directly, as Monroe et al. (2017) does for color terms. Another option is to use the classifiers to to sample from conditioned space of objects. Something like this is the basis of the system proposed by Bernardy et al. (2019), though it is not perceptually grounded. In order for that to work, the embedding space of *PerceptualData* would have to be regularized in such a way that admits sampling, which could potentially be achieved by using a variational autoencoder (Kingma & Welling, 2014).

Aside from compositionality, there remain many questions on the side of lexical representation, such as that of polysemy. It would seem that certain words may appear in multiple places in a taxonomy. The meaning of a word may be ambiguous among a *set* of such corresponding types. So far we have only discussed predicative nouns. Adjectives, and verbs, including transitive verbs admit a similar treatment, but that leaves quantifiers and function words, among others.

Finally, we only discuss perceptual and taxonomical aspects of meaning, but there are other aspects of meaning, including other inferential aspects. How would we represent, for example, that *being from the Champagne region* is an aspect of the meaning of *champagne* (the beverage)? In Marconi (1997)'s schema, this fact would be treated as an aspect of inferential competence. Certainly we should not expect the inference to be derivative of a perceptual classifier for champagne, but it does not fit neatly as taxonomical information either. A more sophisticated type system is needed to incorporate lexical information of this kind.

Acknowledgements

This work was supported by grant 2014-39 from the Swedish Research Council (VR) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *ACL 2020*.

Bengio, S., Weston, J., & Grangier, D. (2010). Label Embedding Trees for Large Multi-Class Tasks. *Advances in Neural Information Processing Systems*, *23*.

See Moro et al. (2019) for more on that task.

Bernardy, J.-P., Blanck, R., Chatzikyriakidis, S., Lappin, S., & Maskharashvili, A. (2019). Bayesian Inference Semantics: A Modelling System and A Test Suite. *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, 263–272. https://doi.org/10.18653/v1/S19-1029

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., & Turian, J. (2020). Experience Grounds Language. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8718–8735. https://doi.org/10.18653/v1/2020.emnlp-main.703

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Cooper, R. (forthc). *From Perception to Communication: A Theory of Types for Action and Meaning*. Oxford University Press.

Cooper, R., Dobnik, S., Lappin, S., & Larsson, S. (2015). Probabilistic Type Theory and Natural Language Semantics. *Linguistic Issues in Language Technology, Volume 10, 2015*.

Dhall, A., Makarova, A., Ganea, O., Pavllo, D., Greeff, M., & Krause, A. (2020). Hierarchical image classification using entailment cone embeddings. https://doi.org/10.48550/ARXIV.2004.03459

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *Conference Proceedings: Papers Accepted to the International Conference on Learning Representations (ICLR)*.

Kolmogorov, A. N. (1950). *Foundations of the theory of probability*. New York: Chelsea Pub. Co.

Lappin, S. (2012). An Operational Approach to Fine-Grained Intensionality. *UCLA Working Papers in Linguistics, Theories of Everything*, *17*, 180–186.

Larsson, S. (2013). Formal semantics for perceptual classification. *Journal of Logic and Computation*, *25*(2), 335–369. https://doi.org/10.1093/logcom/ext059

Larsson, S. (2017). Compositionality for perceptual classification. *IWCS 2017 — 12th International Conference on Computational Semantics — Short Papers*.

Larsson, S. (2020a). Discrete and Probabilistic Classifier-based Semantics. *Proceedings of the Probability and Meaning Conference (PaM 2020)*, 62–68.

Larsson, S. (2020b). Extensions are Indeterminate if Intensions are Classifiers. *SemDial 2020 (WatchDial) Workshop on the Semantics and Pragmatics of Dialogue*, 10.

Larsson, S., & Cooper, R. (2021). Bayesian Classification and Inference in a Probabilistic Type Theory with Records. *Proceedings of the 1st and 2nd Workshops on Natural Logic Meets Machine Learning (NALOMA)*, 51–59.

Marconi, D. (1997). *Lexical competence*. MIT Press.

Martin-Löf, P. (1984). *Intuitionistic Type Theory*. Bibliopolis.

Monroe, W., Hawkins, R. X., Goodman, N. D., & Potts, C. (2017). Colors in Context: A Pragmatic Neural Model for Grounded Language Understanding. *Transactions of the Association for Computational Linguistics*, *5*, 325–338. https://doi.org/10.1162/tacl_a_00064

Moro, D., Black, S., & Kennington, C. (2019). Composing and Embedding the Words-as-Classifiers Model of Grounded Semantics. *arXiv:1911.03283 [cs]*.

Muskens, R. (2005). Sense and the Computation of Reference. *Linguistics and Philosophy*, *28*(4), 473–504. https://doi.org/10.1007/s10988-004-7684-1

Schlangen, D., Zarrieß, S., & Kennington, C. (2016). Resolving References to Objects in Photographs using the Words-As-Classifiers Model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1213–1223. https://doi.org/10.18653/v1/P16-1115

Silberer, C., Ferrari, V., & Lapata, M. (2017). Visually Grounded Meaning Representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2284–2297. https://doi.org/10.1109/TPAMI.2016.2635138

9. Coordinating taxonomical and observational meaning: The case of genus-differentia definitions

Bill Noble, Staffan Larsson, and Robin Cooper

Abstract Genus-differentia definitions exhibit the dual nature of lexical semantic meaning—they incorporate both "hard" *X is a Y* relations between words, as well as "soft" aspects of meaning which can be supported or challenged by observation. Modeling such definitions as contributions in dialogue requires that we accommodate the fluidity of linguistic resources, while respecting the dual nature of the relations that hold between lexical items. In this paper, we use a Probabilistic Type Theory with Records (ProbTTR) to characterise genus-differentia definitions by describing the update they license to the common ground of a dialogue.

9.1. Introduction

Metalinguistic dialogue is one way for speakers to align on the meaning of words. This is common, for example, between adults and child language learners (E. V. Clark, 2007):

(17) a. Naomi: *mittens*.

b. Father: gloves.

c. Naomi: *gloves*.

d. Father: when they have fingers in them they are called gloves and when they are all put together they are called mittens.

But such interactions also take place between adults engaged in a joint activity (Brennan & Clark, 1996):

Ex.	definiendum	genus	differentia
17	mittens	mittens ∨ gloves	fingers are all put together
18	docksider	shoe	leather
		pennyloafer	preppy
19	raven	corvid	large, black

Table 9.1.: Three examples metalinguistic coordination analysed as genus-differentia definitions. While 19 fits neatly into the paradigm, the other two deviate somewhat. In 17, the genus is not explicitly stated, but can be taken to be a join type encompassing both *mittens* and *gloves* (see Cooper & Larsson, 2009). In 18, two alternative definitions are given, each with their own genus and differentia.

(18) a. A: A docksider.

b. B: A what?

c. A: Um.

d. B: Is that a kind of dog?

e. A: No, it's a kind of um leather shoe, kinda preppy pennyloafer.

f. B: Okay, got it.

In both of these examples, the participants have a joint perceptual scene to help ground the meaning of the word, but that need not always be the case. Definition is also a common coordination strategy in *word meaning negotiations* that take place on text-based social media (Myrendal, 2019).

In this paper, we consider a particular definition paradigm known as a *genus-differentia* definitions. Consider the following (imagined) exchange between an expert ornithologist and aspiring birder:

(19) a. A: You know what a corvid is, right?

b. B: Yeah, sure. We have jays and crows in the garden sometimes.

c. A: A raven is a large black corvid.

d. B: Oh, okay.

Each of the above examples can be analysed as including a genus-differentia definition (Table 9.1). Furthermore, it seems reasonable to expect that each exchange results in some update to the *common ground* (H. H. Clark, 1996) of the participants.

Discussion of genus-differentia definitions can be traced back at least as far as Aristotle.[1] For Aristotle, each genus must be separated into species by some external *dif-*

[1] See especially Books VI and VII of *Topics*.

ferentia. Some species, acting as genera themselves, may be further differentiated into subspecies. We adopt some of the language of the Aristotelian tradition (genus, species, differentia), but rather than metaphysics, we are interested in genus-differentia definitions as a conventionalised resource for linguistic agents to coordinate on the meaning a word or phrase.

Genus-differentia definitions convey two kinds of information about the definiendum:

1. **taxonomical information** – A *X is a Y* relationship between the genus and the definiendum

2. **observational information** – One or more *features* that help to *differentiate* (by observation) the definiendum from other species of the same genus

Marconi (1997) argues that there are two ways for speakers to be competent with the use of a word. *Referential competence* is the ability to map words to individuals or events in the world. If someone can identify a raven by sight (or by call, or by observing its behavior), they might be considered referentially competent with *raven*. This aspect of competence seems to be what is mainly at issue in the argument that at least some aspect of lexical semantic meaning may be associated with a perceptual classifier—a cognitive resource for identifying instances of a class, given some perceptual input (Larsson, 2013; Schlangen et al., 2016). On the other hand, *inferential competence* supports the ability to draw inferences based on the use of a word in context. In a community of bird watchers, one might be expected to infer from an utterance like *I saw a raven* that I saw a corvid. Someone who doesn't make that inference might be considered incompetent with the word *raven*, since part of the meaning of *raven* that they are corvids. Formal semantics in the Montagovian tradition, if it considers lexical semantics at all, focuses on inferential aspects of meaning, for example with meaning postulates (Carnap, 1952; Zimmermann, 1999).

Genus-differentia definitions are interesting to consider from the perspective of interaction because describing the result of grounding an utterance like 19c requires a framework that accounts for the dual nature of lexical meaning. We have essentially two desiderata for the shared meaning of *raven* that results from grounding 19c:

D1 Raven is a species of the genus corvid.[2] This means two things: First, there is an intensional inferential relation from species to the genus. That is, there is no situation (actual or hypothetical) in which something might not be a corvid given that it is a raven, since the definition stipulates that being a corvid is part of *what it means* to be

[2]Since we are interested in lexical meaning, the taxonomical information relevant to us is information about *folk taxonomies*, which are a resource for a particular *community of practice* (Gumperz, 1972). Among botanists, a banana is a species of berry while a strawberry is not. The opposite may hold among cooks or in ordinary discourse.

a raven. Second, being a raven is mutually exclusive with each of the sibling species of corvid.[3]

D2 Given that something is a corvid, being large and black (relative to corvids) is positive evidence for being a raven. However, this does not mean that ravens are a *type of* black thing. Any inference from *raven* to *large and black* is defeasible (for example, the speakers may entertain the possibility of albino raven, even if it happens to be extensionally true that all ravens are black). Furthermore, our account should accommodate the possibility that some differentia are interpreted in a way that is sensitive to the context given by the genus. For the sake of example, we will assume that this is the case for *large* but not for *black*.

Our analysis of 19 and therefore these desiderata is admittedly *ad hoc*. Indeed, the use of genus-differentia definitions as a metalinguistic resource is probably a source of variation across different communities of practice. The analysis that leads to these desiderata is partly motivated by the very fact that it requires us to distinguish between taxonomical and observational information about the meaning of *raven*.

We will come back to these desiderata in Section 9.5 after developing some formal machinery that we can use to express them more precisely. Section 9.2 introduces Probabilistic Type Theory with Records (ProbTTR). Section 9.3 describes a way of representing multiclass classifiers in ProbTTR, and Section 9.4 describes *classification systems*, a kind of ProbTTR type system that encodes a taxonomy with types that refer to multiclass classifiers for their witness conditions. Finally, in Section 9.5, we will put these tools together to give an analysis of example 19.

9.2. Probabilistic Type Theory with Records

Probabilistic Type Theory with Records (ProbTTR) is a type system that allows for probabilistic type judgments of the form

$$p(a : T) = r, \tag{9.1}$$

where $r \in [0, 1]$ is a real number. In settings where the type system is a resource for (or models cognitive processes of) an agent, (9.1) is taken to mean that the agent judges entity a to be of type T with probability r.[4]

Possibilities and witness conditions In ProbTTR, *witness conditions* are used to compute the probability that a given entity is of a given type. For basic types, $T \in$ **BType**, witness conditions assign probability dependent on a *possibility* external

[3] Exactly what the sibling species are may be underspecified in the common ground. In this case, it includes at at least *jay* and *crow*, given the context of 19b. In other cases, the relevant sibling species may be inferable from the differentia.

[4] See Cooper et al. (2015) for a more complete introduction to ProbTTR.

to the type system. A possibility can be a set theoretic model (in which case the witness conditions for basic types is one of set membership) or it can, as in this paper, be based on a collection of classifiers (see Section 9.4.2). Thus, we write

$$p(a :_M T) = r \tag{9.2}$$

to mean that a is of type T with probability r in possibility M. Statements like (9.1) should only be used for judgments that hold regardless of possibility, or as a shorthand where it is clear that only one possibility is being considered.

We have not explicitly introduced a probability space underlying type judgments. In general, this may not be formally necessary (see Scott & Krauss, 1966). However, if we did, the sample space would be the set of all possible sets of pairs of basic types and entities:

$$\Omega = \mathcal{P}(\textbf{BType} \times \mathit{Ind})$$

where, for $A \in \Omega$, $\langle T, a\rangle \in A$ would mean that a is of type T in outcome A.

As long as both **BType** and *Ind* are countable (for the purposes of this paper, we may assume they are finite), the distribution is discrete and there is no difficulty in talking directly about the probability of events.

A key point that is elucidated by considering the sample space of basic type judgments is that probabilistic dependencies between type judgments on basic types are entirely determined by M.

Conditional probability We may speak of the *conditional probability* that an entity a is of type T_1 given that it is of type T_2, written $p(a : T_1 \mid a : T_2)$. If we wish to express the probability (in general) that something is of type T_1 given that it is of type T_2, this is written $p(T_1 \| T_2)$. The use of the double stroke is to distinguish this expression from the probability that something *exists* of type T_1, given that something *exists* of type T_2, which is written $p(T_1 \mid T_2)$. These conditional probabilities are understood extensionally, specific to a particular *possibility*. If, for example, we know that penguins only live in Antarctica, we would, for the types *Penguin* (the type of situation in which there is a penguin) and *Antarctica* (the type of situation in Antarctica), judge $p(\mathit{Antartica} \| \mathit{Penguin})$ to be 1 (or close to 1) on the basis of this contingent fact.

Structured types The witness conditions of structured types are a function of the structure of the type and its components. For example, given types T_1 and T_2, the meet type $T_1 \wedge T_2$ has, witness conditions based on the Kolmogorov (1950) equation for conjunctive probability (Cooper et al., 2015):

$$\begin{aligned} p(a : T_1 \wedge T_2) &= p(a : T_1) \cdot p(a : T_2 \mid a : T_1) \\ &= p(a : T_2) \cdot p(a : T_1 \mid a : T_2) \\ &= p(a : T_2 \wedge T_1) \end{aligned} \tag{9.3}$$

In addition to types defined with $\wedge$, $\vee$ and $\neg$, ProbTTR defines *record types* as structured types—given a record s and record type R, $p(s : R)$ is a function of type judgments of the fields of s (see Cooper et al. (2015) for details).

9.2.1. Hard and soft relations between types

Subtype relation In TTR, T_1 is said to be a *subtype* of T_2, $T_1 \sqsubseteq T_2$ if and only if anything of type T_1 is also of type T_2 for any possibility M, (Cooper, forthc, p. 285). Extending this to ProbTTR, we can say,

$$T_1 \sqsubseteq T_2 \text{ iff } p(a :_M T_1) \leq p(a :_M T_2), \tag{9.4}$$

for any entity a and possibility M.

Naturally, it is not always necessary to check these conditions explicitly.[5] Subtype relations can be implicit in the structure of the types, as in the case of meet types. If $T_3 = T_1 \wedge T_2$, by the definition of the meet type we have $T_3 \sqsubseteq T_1$ and $T_3 \sqsubseteq T_2$.

In other cases, whether two types stand in a subtype relation may depend on what is meant by *all possibilities*. If we literally mean all possible assignments of probability to basic type-entity pairs, then two basic types will never stand in a subtype relation, since there will always be possibilities where $p(a :_M T_1) > p(a :_M T_2)$ and *vice versa*.

If, on the other hand, we restrict our attention to some class of possibilities $\mathcal{M}$, then subtype relations between basic types are possible. Witness conditions are one way to limit the possibilities under consideration and can therefore introduce probibalistic dependency between types.

Evidential relation We introduce a "soft" relation between types in ProbTTR, which captures the notion that T_2 is *evidence for* T_1 in the context of some type T^*. Two types stand in this relation with respect to T^* if learning that something is of type T_2 increases the probability that it is of type T_1:

$$T_1 \prec_{T^*} T_2 \text{ iff } p(T_1 \| T^*) < p(T_1 \| T_2, T^*) \tag{9.5}$$

This relation is also contingent, relative to a particular possibility.

9.2.2. Representing probability distributions

In the next section, we will define a type for probabilistic multiclass classifiers—that is, classifiers that compute the probability that a given entity belongs to each of several mutually exclusive classes. To that end, we must first encode discrete categorical probability distributions in TTR, since the output of the classifier takes that form.

[5] Indeed, it may not even be possible, depending on the notion of possibility since the "extension" of types with witness conditions based on classifiers is indeterminate (Larsson, 2020b).

Larsson and Cooper (2021) introduce a type theoretic counterpart of a random variable in Bayesian inference. To represent a single (categorical) random variable with a range of possible (mutually exclusive) values, ProbTTR uses a *variable type* $\mathbb{A}$ whose range is a set of *value types* $\mathfrak{R}(\mathbb{A}) = \{A_1, \ldots, A_n\}$. We might have, for example, $\mathfrak{R}(\mathit{Animal}) = \{\mathit{Bird}, \mathit{Reptile}, \ldots\}$.

We will use short-hands *Animal*, *Bird* etc, for the situation where some individual is an animal, bird, etc.:

$$\mathit{Animal} = \begin{bmatrix} \text{x} & : & \mathit{Ind} \\ \text{c} & : & \text{animal(x)} \end{bmatrix}$$

$$\mathit{Bird} = \begin{bmatrix} \text{x} & : & \mathit{Ind} \\ \text{c} & : & \text{bird(x)} \end{bmatrix}$$

For a situation s, a probability distribution over the m value types $A_j \in \mathfrak{R}(A), 1 \leq j \leq m$ belonging to a variable type $\mathbb{A}$ can be written (as above) as a set of Austinian propositions, e.g.,

$$\{ \begin{bmatrix} \text{sit} & = & \text{s} \\ \text{sit-type} & = & \text{A}_\text{j} \\ \text{prob} & = & \text{p(s : A}_\text{j}\text{)} \end{bmatrix} \mid A_j \in \mathfrak{R}(\mathbb{A}) \} \tag{9.7}$$

However, we will also have use for an alternative representation of probability distributions, that indexes the probability assigned to each type with a unique label associated with the type:

$$\text{idx}(\{ \begin{bmatrix} \text{sit} & = \text{s} \\ \text{sit-type} & = \text{A}_\text{j} \\ \text{prob} & = \text{p(s : A}_\text{j}\text{)} \end{bmatrix} \mid A_j \in \mathfrak{R}(\mathbb{A}) \}) = \begin{bmatrix} \text{lbl(A}_1\text{)} & = & \text{p}_1 \\ \vdots & = & \vdots \\ \text{lbl(A}_\text{n}\text{)} & = & \text{p}_\text{n} \end{bmatrix}$$

where $p_j = p(s : A_j)$ and $\text{lbl}(A_j)$ is a unique label for $A_j \in \mathfrak{R}(\mathbb{A})$. This means that for a set of probabilistic Austinian propositions P_s, that concern a situation s, $\text{idx}(P_s).\text{lbl}(A_j) = p_j = p(s : A_j)$.

9.3. Multiclass Classifiers in ProbTTR

In this section we extend the TTR classifier defined by Larsson (2013) to give probabilistic type judgments in multiclass setting.

Larsson (2013) shows how perceptual classification can be modelled in TTR and Larsson (2020a) reformulates and extends this formalisation to probabilistic classification. Adapting the notation of a probabilistic TTR classifier to the current setting, a probabilistic perceptual (here, visual) classifier $\kappa_{\mathbb{A}}$ corresponding to a variable type $\mathbb{A}$ provides a mapping from perceptual input (of type $\mathfrak{V}$ e.g., a digital image) onto a probability distribution over value types in $\mathfrak{R}(\mathbb{A})$, encoded as a set of probabilistic Austinian propositions.

We also want to explicitly parametrise our classifier. A classifier $\kappa_{\mathbb{A}}$, would thus be a function of type:

$$\Pi \rightarrow Sit_{\mathfrak{V}} \rightarrow \{ \begin{bmatrix} \text{sit} & : & Sit_{\mathfrak{V}} \\ \text{sit-type} & : & RecType_{A_i} \\ \text{prob} & : & [0,1] \end{bmatrix} \mid A_i \in \mathfrak{R}(\mathbb{A}) \} \tag{9.8}$$

where Π is the type of the parameters needed by $\kappa_{\mathbb{A}}$, and $Sit_{\mathfrak{V}}$ is the type of situations where perception of some object yields visual information, and where $RecType_R$ is the (singleton) type of records identical to R, so that e.g.,

$$T : RecType_{Bird} \text{ iff } T : RecType \text{ and } T = Bird$$

We take classifiers to be part of word meanings. We associate a word like ”bird” with a type *Bird* which is in turn associated with lexical entry in the form of a TTR record:

$$Lex(Bird) = \begin{bmatrix} \text{bg} & = & \text{Sit}_{\mathfrak{V}} \\ \text{par} & = & \pi \\ \text{intrp} & = & \lambda \text{r} : \text{bg} . Bird \\ \text{clfr} & = & \lambda \text{r} : \text{bg} . \kappa_{Animal}(\text{par}, \text{r}) \end{bmatrix} \tag{9.9}$$

Assuming we have a function *Lex* that looks up the lexical entry related to a type (associated with a word), we also define a lookup function that gives us the classifier corresponding to a type:

$$Clfr(T) = Lex(T).\,\text{clfr}$$
$$Intrp(T) = Lex(T).\,\text{intrp}$$

Let us assume a s_{123} situation where a speaker points to a bird a and says "Bird!" (meaning "that is a bird"). We want to classify a perceived situation as being of the type *Bird* or not, or in the probabilistic case, compute the probability of the judgment.

Now, to judge the probability with which a situation s is of a type *Bird* (to continue with our example), the agent looks up the related classifier and applies it to s, which produces a probability distribution over different subtypes of *Animal*. The agent then looks up the probability associated with *Bird*. The general method for doing this can be written as:

$$p(s : T) = \text{idx}(\textit{Clfr}(T)(s)).\,\text{lbl}(\textit{Intrp}(T)(s))$$

In our case:

$$p(s_{123} : \textit{Bird}) = \text{idx}(\kappa_{\textit{Animal}}(\pi, s_{123})).\,\text{lbl}(\textit{Bird})$$

9.4. Classification systems in ProbTTR

To represent both taxonomical and observational relations between types, we will embed a *classification system* in ProbTTR. A classification system has two components, a *taxonomy* (Section 9.4.1), which is a set theoretic object representing an ontological hierarchy, and a collection of *classifiers* (Section 9.4.2) associated with the taxonomy. Ultimately the classifiers will provide witness conditions for certain basic types and the taxonomy will be fully encoded in the type system, but first we define the structure in set theoretic terms so that we can create a ProbTTR system with the correct subtype relations.

9.4.1. Taxonomy

A taxonomy is a rooted tree structure defined by a tuple,

$$\mathbf{T} = \langle T, D, t^* \rangle, \tag{9.10}$$

where T is a set of *taxons*, $D \subseteq T \times \mathcal{P}(T)$ is a set of *distinctions* on T, and $t^* \in T$ is the root taxon.

To elaborate, T is simply a finite set of labels and D provides the hierarchical structure of the taxonomy. *Distinctions* (elements of D) take the form $\langle g, S \rangle$, where $g \in T$ and $S \subset T$, and $|S| \geq 2$. We say that the taxons g and s stand in a genus-species relationship if there is some $\langle g, S \rangle \in D$ such that $s \in S$. Then s can be said to be a *species of* g. Alternatively, we can say that g is *the genus of* s.

This requires certain restrictions on $\mathbf{T}$. Namely, that it is:

- **Acyclic**: There are no cycles. I.e., no chain of distinctions $\{\langle g_1, S_1\rangle, ..., \langle g_n, S_n\rangle\}$ such that $g_2 \in S_1, ..., g_n \in S_{n-1}$ and $g_1 = g_n$.
- **Rooted**: There is no distinction $\langle g, S\rangle \in D$ with $t^* \in S$.
- **Uniquely connected**: For every $t \neq t^*$ there is exactly one $\langle g, S\rangle \in D$ such that $t \in S$.[6]

Importantly, this still allows for multiple distinctions in which the same taxon acts as a genus. In other words, we can have $\langle g, S\rangle, \langle g, S'\rangle \in D$ where $S' \neq S$. For example, we might imagine a taxonomy in which both $\langle \mathit{Animal}, \{\mathit{Bird}, \mathit{Reptile}, ...\}\rangle$ and $\langle \mathit{Animal}, \{\mathit{Carnivore}, \mathit{Herbivor}, \mathit{Omnivore}\}\rangle$ are distinctions.

The *uniquely connected* constraint allows us to define a function

$$\mathit{Dist} : T \setminus \{t^*\} \to D \tag{9.11}$$

that gives, for each taxon, t (other than t^*), the distinction $\mathit{Dist}(t) = \langle g, S\rangle$ such that $t \in S$. For convenience we also define the functions *Genus*, and *Siblings* such that

$$\langle \mathit{Genus}(t), \mathit{Siblings}(t)\rangle = \mathit{Dist}(t). \tag{9.12}$$

Note that under this definition, leaf taxons are those taxons for which there are no distinctions in D where the taxon appears as a genus.

9.4.2. Species Classifiers

In addition to the taxonomy, we have a collection of classifiers, **K** and parameters **P**, each of which we index with elements of D, such that $\kappa_d \in \mathbf{K}$ is the classifier for distinction d provided with the appropriate parameters. This follows the intuition that a distinction in the taxonomy may be accompanied by an ability to *distinguish* among the relevant species. In general, we need only assume that we have classifiers for those distinctions that include at least one leaf taxon, since genus taxons can be defined as the join of their species in certain cases.[7] For now we will assume we have a classifier for each distinction in D.

9.4.3. The type system

Suppose we have a taxonomy $\mathbf{T} = \langle T, D, t^*\rangle$ and a collection of classifiers **K** on the distinctions of that taxonomy. Let *Dom* be a special type corresponding to the root of the taxonomy. We then define variable types $\mathbb{A}_d$ for each $d = \langle g, S\rangle \in D$

[6] A weakness of insisting on a tree structure is that we cannot have taxons that appear in multiple places in the taxonomy, whereas in folk taxonomies it would appear this is common. We would either need to say that the apparently duplicated taxon is actually part of a distinction at a higher level that encompasses both, or that it corresponds to two senses of the same word.

[7] See Marconi (1997, ch. 6) on "subordinate concepts".

with $\mathfrak{R}(\mathbb{A}) = \{A_{s_1}, ..., A_{s_n}\}$ corresponding to $s_1, ..., s_n \in S$. Classifiers provide the witness conditions for the value types as described in Section 9.3. For a given entity a,

$$p(s : A_t) = \begin{cases} 1 & \text{if } t = t^* \\ \kappa_{Dist(t)}(a)(t) & \text{otherwise} \end{cases} \tag{9.13}$$

In other words, the probability assigned to A_t is 1 in the case of the root taxon, and otherwise determined by the classifier for the distinction corresponding to the variable in which A_t is a value type. These "auxiliary" value types we can give the witness conditions for the associated with the taxonomical categories as the product of the jugment of the genus and the axiliary type. For any object a,

$$p(a : T_t) = p(a : A_t) \cdot p(a : T'_t) \tag{9.14}$$

where

$$T'_t = \begin{cases} \mathit{Dom} & \text{if } t = t^* \\ T_{Genus(t)} & \text{otherwise} \end{cases}$$

This stipulates that the classifiers give us the probability that an individual is of each of the species types, *given* that it is of the genus type. Thus judgments about T_t correspond to an *absolute* judgment about belonging to the taxon.

Taken together, Equations 9.13 and 9.14 imply that for any a, $p(a : T_{t^*}) = p(a : \mathit{Dom})$. In situations where the root taxon corresponds to all individuals (i.e., where $\mathit{Dom} = \mathit{Ind}$), we have $p(a : T_{t^*}) = 1$ for any a. It is also possible, however, to embed a classification system in an existing type system, as long it provides witness conditions for *Dom*. For example, if the classification system is specific to birds, we might embed it in a larger system that gives witness conditions for *Bird*.

9.4.4. Feature classifiers

In addition to the distinction classifiers, a classification system may include some number of types based on feature classifiers. A feature classifier takes any entity $a : \mathit{Dom}$ as input, and receives its witness conditions from a classifier that results in a probabilistic type judgement. In general, feature and distinction classifiers need not interact explicitly though, considered as random variables, there may be probabilistic dependence between them. Distinction classifiers may be defined in terms of feature classifiers, for example as Bayesian classifiers that take the result of feature classifiers as their input (see, Larsson and Bernardy (2021)).

In general, some of these feature types may be dependent types. Consider a type like *Tall*. Whether or not an individual is tall may depend on a comparison class (for example, a type in the taxonomy). Following Fernandez and Larsson (2014), we define

dependent feature types with classifiers that take a threshold function as a parameter. For example,

$$\theta_{Large} : Type \to \mathbb{R}^+ \tag{9.15}$$

This gives the classifier the following type:

$$\kappa_{Large} : (Type \to \mathbb{R}^+) \to Type \tag{9.16}$$

9.5. Combining the observation and taxonomical aspects of genus-differentia definitions

With this formal machinery in place, we return to the project of characterising the result of grounding 19c. First, let's lay out what is shared among speakers A and B before 19c is grounded.

We will assume that A and B share a classification system with *Bird* at its root as part of their common ground. Utterance 19d establishes that a type for the lexical entry of *corvid*, for which we will use *Corvid*, is a type in this system, and that there is a distinction on *Corvid* such that $\mathfrak{R}(Corvid) \supseteq \{Jay, Crow\}$, where *Jay* and *Crow* are the lexical entries for *jay* and *crow*—that is, for all species types of *Corvid* given by the common ground , S (including at least *Jay* and *Crow*), $S \sqsubseteq Corvid$. The witness conditions for each $S \in \mathfrak{R}(Corvid)$ are given by a multiclass classifier κ_{Corvid}. Since $Corvid \sqsubseteq Bird$, we may also assume that $Dist(Corvid)$ exists and that there is a classifier $\kappa_{Dist(Corvid)}$, though it need not be common ground what the genus of *Corvid* is.

Furthermore, we will assume we have types *Large* and *Black*, whose witness conditions are given by feature classifiers. For the purposes of the example, we will assume that *Black* is basic type that gets its witness conditions from a feature classifier, κ_{Black}, whereas $Large : Type \to Type$ is a dependent type with a classifier that depends on threshold function θ_{Large}. Thus, the witness conditions for $Large(Corvid)$ are given by $\kappa_{Large}(\theta_{Large}(Corvid))$. This leaves open the question of exactly how θ_{Large} is defined, but we may assume that the value of $\theta_{Large}(Corvid)$ depends in some way on the parameters of the classifier that defines the witness conditions for *Corvid*, namely $\kappa_{Dist(Corvid)}$.

Returning to our desiderata, we want to construct a type, *Raven*, such that:

$$\sum_{T \in Species(Corvid) \cup \{Raven\}} p(T \| Corvid) = 1 \tag{9.17a}$$

$$Raven \sqsubseteq Corvid \tag{9.17b}$$

$$Raven \prec_{Corvid} Large(Corvid) \wedge Black \tag{9.17c}$$

Here (9.17a) and (9.17b) formalise **D1** and 9.17c formalises **D2**.

9.5.1. Constructive approach

As discussed previously, one motivation for formalising this example and the interactive semantics of genus-differentia definitions in general is to expose some crucial distinctions in lexical semantics that are often overlooked. In this section, we give what is a rather straight-forward and intuitive solution to the challenge we have given ourselves, but one that fails to adequately make the distinction between taxonomical and observational lexical information.

In this solution, we attempt to directly construct a new type *Raven* out of the common ground types already available. The most straight-forward way to do this is with meet types:

$$\mathit{Raven} = \mathit{Corvid} \wedge (\mathit{Large}(\mathit{Corvid}) \wedge \mathit{Black}) \tag{9.18}$$

This definition is intuitively appealing—19c is saying that ravens are large *and* black *and* corvids. Furthermore, this definition does actually satisfy the desiderata stated so far.

To maintain (9.17a), we can redefine each existing species type S as:

$$S' = S \wedge \neg \mathit{Raven} \tag{9.19}$$

We have $\mathit{Raven} \sqsubseteq \mathit{Corvid}$, satisfying (9.17b), since by the Kolmogorov (1950) definition of the meet type (9.3), for any possibility M and any entity a,

$$\begin{aligned} & p(a :_M \mathit{Raven}) \\ =& p(a :_M \mathit{Corvid}) \cdot\ p(a :_M \mathit{Large}(\mathit{Corvid}) \wedge \mathit{Black} \mid \mathit{Corvid}) \\ \leq& p(a :_M \mathit{Corvid}) \end{aligned}$$

Finally, (9.17c) holds since it follows from the definition of *Raven* that,

$$p(\mathit{Raven} \| \mathit{Large}(\mathit{Corvid}) \wedge \mathit{Black}, \mathit{Corvid}) = 1$$

and, assuming there are at least some non-large, non-black corvids,

$$p(\mathit{Raven} \| \mathit{Corvid}) < 1.$$

[8]

However, the definition of the meet type (9.3) implies we also get

$$\mathit{Raven} \sqsubseteq \mathit{Large}(\mathit{Corvid}) \text{ and } \mathit{Raven} \sqsubseteq \mathit{Black}.$$

[8] This assumption is justified by a pragmatic requirement of genus-differentia definitions that the differentia do at least some work to *differentiate* the definiendum from other species of the genus.

It does not make sense for *Raven* to be a *subtype* of large corvids or of black things (consider again the possibility of an albino raven). Put another way, it should be possible to construct a hypothetical possibility M and entity a such that:

$$p(a :_M \textit{Large}(\textit{Corvid}) \wedge \textit{Black}) = 0 \text{ and}$$
$$p(a :_M \textit{Raven}) > 0 \tag{9.20}$$

In the next section, We will consider this a new desiderata along with the constraints in (9.17). Instead of constructing the type directly from existing types, we posit a basic type without explicit witness conditions, but with some constraints that are derived from by the genus-differentia definition.

9.5.2. Underspecifed approach

Cooper (forthc) treats types as having an existence independent of their witness conditions. Two types can share the same witness conditions, for example, and still play different roles in an agent's type system. Part of the motivation for doing this is that an agent can reason about a type and its relation to other types without specifying witness conditions for that type. This is in contrast to predicates in first-order logic, for example, which don't have any meaning independent of the model theoretic entities they are interpreted as.

We would like to interpret definitions like 19c as giving rise to an underspecified type; that is, a type without explicit witness conditions. Instead, we assert the following relationships between the new underspecified type *Raven* and other existing common ground types:

$$\textit{Raven} \sqsubseteq \textit{Corvid} \tag{9.21a}$$
$$p(\textit{Large}(\textit{Corvid}) \wedge \textit{Black} \| \textit{Raven}) = 1 \tag{9.21b}$$

Notice that neither of these two conditions give us direct witness conditions for *Raven*. The first condition says that anything (in any possibility) that is a raven is also a corvid. The second condition says that anything that is a raven is, with probability 1, is large (for a corvid) and black. Note that (9.21b) is a constraint on the type's witness conditions given the current possibility, meaning that we can not infer $\textit{Raven} \sqsubseteq \textit{Large}(\textit{Corvid}) \wedge \textit{Black}$, since nothing prevents us from constructing a possibility in which (9.20) holds. In other words, albino ravens are still possible.

Clearly condition (9.17b) is satisfied by construction. This may be a bit unsatisfying, but it is worthwhile to consider that asserting $\textit{Raven} \sqsubseteq \textit{Corvid}$ amounts to adding *Raven* as a witness condition to *Corvid*. Put another way, for any entity a and possibility M, $P(a :_M \textit{Corvid}) \geq P(a :_M \textit{Raven})$.

In order to satisfy (9.17a), we need to redefine the witness conditions of the existing species types to "make room" in the probability distribution for *Raven*. How to do

this depends somewhat on how completely the distinction is specified in the common ground. If there is an *other corvid* type, *Other*, we might just redefine the classifier for that type so that for any entity a, $\kappa'_{corvid}(a)(other) = \kappa_{corvid}(a)(other) - f(a)$, where f is such that $0 < f(a) < \kappa_{corvid}(a)(other)$. Alternatively, we might take some probability from each class. Either way, the solution should be a function of a that depends on the differentia, but exactly what that function is is not common ground since (9.21b) gives a unidirectional conditional—all ravens are large and black, but there may still be large, black, non-raven corvids.

It remains to be shown that (9.17c) holds. In the following, let $D = Large(Corvid) \wedge Black$ and S be the set of types representing each of the sibling species of *Corvid*, including *Raven*.

$$
\begin{aligned}
& p(Raven \| D, Corvid) \\
=& \frac{p(Raven \| Corvid) \cdot p(D \| Raven, Corvid)}{\sum_{T \in S} p(T \| Corvid) \cdot p(D \| T, Corvid)} && (9.22) \\
=& \frac{p(Raven \| Corvid) \cdot p(D \| Raven)}{\sum_{T \in S} p(T \| Corvid) \cdot p(D \| T)} && (9.23) \\
>& p(Raven \| Corvid) \cdot p(D \| Raven) && (9.24) \\
=& p(Raven \| Corvid) && (9.25)
\end{aligned}
$$

In the above, (9.22) follows from Bayes rule and the fact that $\sum_{T \in S} p(T \| Corvid) = 1$, and (9.23) follows from $Raven \sqsubseteq Corvid$. For (9.24), we must assume that

$$\sum_{T \in S} p(T \| Corvid) \cdot p(D \| T) \leq 1.$$

This is the same assumption we made in the previous approach, which we argue follows from the pragmatics of genus-differentia definitions—namely that not all non-raven corvids are large and black. Finally, (9.25) follows directly from (9.21b).

In this approach, the type for *raven*, *Raven* is defined only in terms of its relationship to types corresponding to other terms in the utterance. A notable feature of this solution is that everything we learn from the definition can be stated in terms of witness conditions for types that already exist: In the case of *corvid*, we know that anything that witnesses the type *Raven* is a witness for the type *Corvid*. This holds intensionally, meaning that it is true independent of possibility. In the case of *large* and *black*, we know *extensionally* that anything that is a raven will be large and black.

Speaker B learns the type *Raven* and the constraints associated with it (9.21) based on the definition offered by A in 19c. After 19d, this type and the associated constrains are added to the common ground.

9.6. Conclusion

The main goal of this paper was to develop a framework that can deal with the distinction between taxonomical and observational lexical information. We argue that this distinction is one that speakers make in metalinguistic interaction, as in genus-differentia definitions. In order to account for this distinction, we use a type system in which intensional relations between types can be reasoned about independently of their witness conditions, which depend on facts about the world.

Our account has been agnostic to the implementation of the classifiers involved. This is justified, in part, by the fact that we describe updates to the conversational common ground, rather than individual agents' abilities. However, it may also be interesting to consider what effect a dialogue like 19 may have on speaker B's ability to recognise ravens. This is related to the machine learning task of *zero-shot classification*, in which an existing classifier is adapted to recognise instances of previously unknown classes based on external information (such as a natural language descriptions). Future work should consider how zero-shot classification can be analysed from an interactive perspective.

References

Brennan, S. E., & Clark, H. H. (1996). Conceptual Pacts and Lexical Choice in Conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482.

Carnap, R. (1952). Meaning postulates. *Philosophical Studies*, *3*(5), 65–73. https://doi.org/10.1007/BF02350366

Clark, E. V. (2007). Young Children's Uptake of New Words in Conversation. *Language in Society*, *36*(2), 157–182.

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Cooper, R. (forthc). *From Perception to Communication: A Theory of Types for Action and Meaning*. Oxford University Press.

Cooper, R., Dobnik, S., Lappin, S., & Larsson, S. (2015). Probabilistic Type Theory and Natural Language Semantics. *Linguistic Issues in Language Technology, Volume 10, 2015*.

Cooper, R., & Larsson, S. (2009). Compositional and ontological semantics in learning from corrective feedback and explicit definition. *Proceedings of the 13th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.

Fernandez, R., & Larsson, S. (2014). Vagueness and Learning: A Type-Theoretic Approach. *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, 151–159. https://doi.org/10.3115/v1/S14-1019

Gumperz, J. (1972). The Speech Community. In P. P. Giglioli (Ed.), *Language and social context: Selected readings*. Harmondsworth : Penguin.

Kolmogorov, A. N. (1950). *Foundations of the theory of probability*. New York: Chelsea Pub. Co.

Larsson, S. (2013). Formal semantics for perceptual classification. *Journal of Logic and Computation*, *25*(2), 335–369. https://doi.org/10.1093/logcom/ext059

Larsson, S. (2020a). Discrete and Probabilistic Classifier-based Semantics. *Proceedings of the Probability and Meaning Conference (PaM 2020)*, 62–68.

Larsson, S. (2020b). Extensions are Indeterminate if Intensions are Classifiers. *SemDial 2020 (WatchDial) Workshop on the Semantics and Pragmatics of Dialogue*, 10.

Larsson, S., & Bernardy, J.-P. (2021). Semantic Classification and Learning Using a Linear Tranformation Model in a Probabilistic Type Theory with Records. *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, 14–22.

Larsson, S., & Cooper, R. (2021). Bayesian Classification and Inference in a Probabilistic Type Theory with Records. *Proceedings of the 1st and 2nd Workshops on Natural Logic Meets Machine Learning (NALOMA)*, 51–59.

Marconi, D. (1997). *Lexical competence*. MIT Press.

Myrendal, J. (2019). Negotiating meanings online: Disagreements about word meaning in discussion forum communication - Jenny Myrendal, 2019. *Discourse Studies*, *21*(3), 317–339. https://doi.org/10.1177/1461445619829234

Schlangen, D., Zarrieß, S., & Kennington, C. (2016). Resolving References to Objects in Photographs using the Words-As-Classifiers Model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1213–1223. https://doi.org/10.18653/v1/P16-1115

Scott, D., & Krauss, P. (1966). Assigning Probabilities to Logical Formulas. In J. Hintikka & P. Suppes (Eds.), *Studies in Logic and the Foundations of Mathematics* (pp. 219–264). Elsevier. https://doi.org/10.1016/S0049-237X(08)71672-0

Zimmermann, T. E. (1999). Meaning Postulates and the Model-Theoretic Approach to Natural Language Semantics. *Linguistics and Philosophy*, *22*(5), 529–561. https://doi.org/10.1023/A:1005409607329

10. Describe me an Aucklet: Generating grounding perceptual category descriptions

Bill Noble and Nikolai Ilinykh

Abstract Human language users can generate descriptions of perceptual concepts beyond instance-level representations and also use such descriptions to learn provisional class-level representations. However, the ability of computational models to learn and operate with class representations is under-investigated in the language-and-vision field. In this paper, we train separate neural networks to **generate** and **interpret** class-level descriptions. We then use the zero-shot classification performance of the interpretation model as a measure of communicative success and class-level conceptual grounding. We investigate the performance of *prototype*- and *exemplar*-based neural representations grounded category description. Finally, we show that communicative success reveals performance issues in the generation model that are not captured by traditional intrinsic NLG evaluation metrics, and argue that these issues can be traced to a failure to properly ground language in vision at the class level. We observe that the interpretation model performs better with descriptions that are low in diversity on the class level, possibly indicating a strong reliance on frequently occurring features.

10.1. Introduction

Language is one of the means we use to achieve our goals when acting in the real world (Chandu et al., 2021). We link forms (symbols) with meaning rooted in other modalities such as perception, sensormotorics, and sounds among others. Mapping words with experiences is often referred to as *grounding* (Harnad, 1990). Large language models trained with tremendous amounts of texts have been criticised for lacking grounded representations (Bender & Koller, 2020; Bisk et al., 2020), and the fast-growing field of multi-modal NLP has been addressing this problem (Beinborn et al., 2018). However, multi-modal and presumably grounded models have several areas for improvement. Recent work suggests that they are affected by the distribution of items

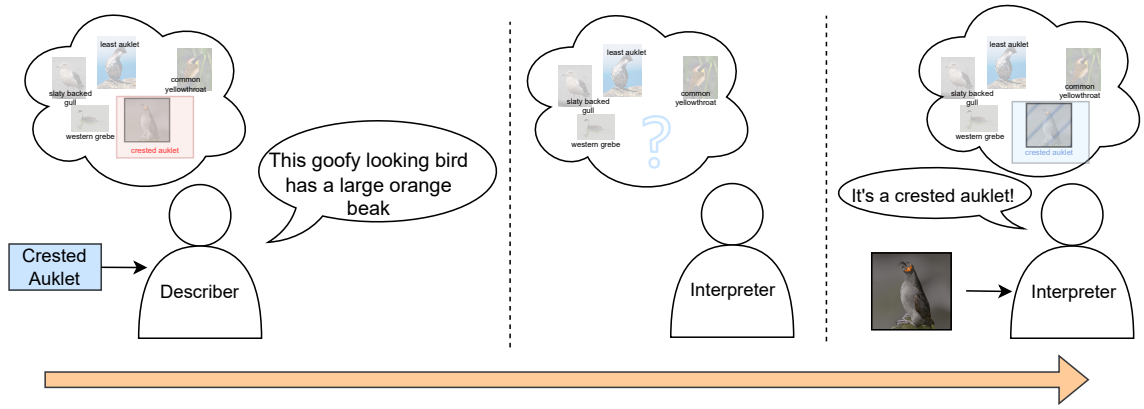


Figure 10.1.: The general pipeline of learning perceptual concepts through simple interactive scenario. The yellow arrow below represents the timeline, while blurred images of class instances represent class representations.

in training data, often over-representing specific scenarios and under-representing others (Agrawal et al., 2018). This ultimately leads to models which rely too much on text and have to be supplied with various mechanisms, enforcing and controlling their attention on other modalities such as vision (Lu et al., 2017; Thomason et al., 2019). This raises the question of what kind of relationship these models learn between linguistic and extra-linguistic information and whether they really can be said to be *grounded* in the same way that humans are.

The problem of imbalance of real-world situations in existing multi-modal datasets is why the models trained on such data tend to *hallucinate*. As shown by Rohrbach et al. (2018), although a neural image captioning system is capable of correctly describing objects in the image (e.g., grounding of words into objects), it is not able to describe the situation, instead falling into the pitfall of using the most common context that the specific configuration of objects would occur. Similar problems are common for other multi-modal tasks such as VQA (Antol et al., 2015) and EQA (Das et al., 2018) models and datasets (Alayrac et al., 2022). These and many more examples indicate that perceptual grounding cannot be achieved in the abstract but must be considered for a particular *communicative context* and *situation*. Even with the current multi-modal datasets which provide perceptual context for learning situation-specific grounding, existing models need help to be more grounded. But perception also plays a role in linguistic communication even when the immediate perceptual context is not an issue. If someone says that *red apples are sweeter than green ones*, this communicates something, even to someone who is not concurrently looking at or tasting apples. We can acquire a (provisional) perceptual concept based on a natural language description. This is sometimes called *fast mapping* (Carey, 1981; Gelman & Brandone, 2010). Human speakers can also *generate* such descriptions with the intent that their interlocutor might later use what they described to classify entities as belonging to the class or not. Can models create and operate with structures similar to provisional

concept representations as humans do?

In this paper, we propose *perceptual category description*, a communicative context that requires grounding in a perceptual modality, emphasising category-level grounding. We view this task as modelling a *very simple* interactive scenario in which (1) a describer, Gen, generates a description of a visual category, then (2) an interpreter, Ipt, learns from the generated description, and (3) classifies among both the `seen` classes they have existing knowledge of and the `unseen` classes that they have only had described by Gen (see Figure 10.1). Ipt performs zero-shot image classification with a natural language description generated by Gen as auxiliary class information. From the perspective of zero-shot learning, using text descriptions as auxiliary data for visual classification is not a novel task, but we focus on description *generation*, using the zero-shot interpreter model as a semi-extrinsic evaluation of the describer. In practice, the task is implemented with disjoint *sets* of `seen` and `unseen` classes (that is, there may be more than one `unseen` class, as is standard in zero-shot learning scenarios). The Gen model is trained on images from both `seen` and `unseen`, but only sees a text from `seen` during training. This ensures that at test time, Gen is using class-level visual information to generate grounded descriptions, rather than merely memorising textual information and using the class representation as a retrieval key. The Ipt model is trained in a standard zero-shot classification setup with no access to images from `unseen` at train time.

In contrast to many visually grounded language generation tasks, the describer is not (necessarily) describing from an image, as in image captioning, but rather from a class-level representation of a visual concept. This brings us to the experimental focus of the current study.

10.2. Background: prototypes and exemplars

Cognitive theories of categorisation are psychologically-motivated accounts of how humans represent perceptual concepts and use them for classification. These theories came about, in part, in reaction to an assumption in certain linguistic traditions that the categories to which predicate-denoting lexical items refer can be defined in terms of a set of necessary and sufficient features. Cognitive theories of categorisation, in contrast, try to account for phenomena like *prototypically effects*, in which certain members of a category are perceived as more representative of the class than others. In *prototype theory*, cognitive categories themselves are defined not by a set of features but rather by a **prototype**, an abstract idealisation of the category. Membership in the class, then, is judged in reference to the prototype (Rosch, 1975). In *exemplar theory*, (e.g., Medin & Schaffer, 1978; Nosofsky, 1984), concepts are still defined in relation to an ideal, but this time the ideal is an **exemplar**, which is a particularly representative *member* of the very category. Put another way, an exemplar is *of the same kind* as the other members of the category, whereas prototypes, in general, are not. There is also

some experimental evidence to suggest that humans use **both** exemplar and prototype-based strategies in classification (Blank & Bayer, 2022; Malt, 1989).

Turning back to the linguistic perspective, it's important to remember that perceptual categories are not only used for classification, but also for linguistic interpretation and (most relevant to the current work) generation. In fact, *classifier-based meaning* has been proposed as a way to ground natural language in perception (Schlangen et al., 2016; Silberer et al., 2017). There are both formal and computational interpretations of this approach that support compositional semantics for lexical items with classifier-based perceptual meanings (Kennington & Schlangen, 2015; Larsson, 2013). In this paper, we are interested in how classifier-based meaning can be used to support generation of class-level descriptions. For this reason, we test three different Gen model architectures. One motivated by prototype theory, one by exemplar theory, and one that uses a hybrid approach.

10.3. Related work

Both our networks operate with textual descriptions: one generates them, and another interprets them. The interpretation model performs zero-shot with the vision as the *primary modality* and text as the auxiliary modality.[1] In zero-shot learning scenarios that use text as auxiliary data, the quality of the text and its relationship to textual information density and richness has been shown to improve model performance. Paz-Argaman et al. (2020), for example, shows that novel concepts are learned better when texts are more perceptual relevant. Similarly, Bujwid and Sullivan (2021) have shown that Wikipedia texts can be used as class descriptions for learning a better encoding of class labels. Similarly, research in computer vision has found that it is possible to learn better visual representations from longer texts rather than labels and use them for a nominally non-linguistic task, e.g. classification (Desai & Johnson, 2021). Forcing a better mapping between visual and linguistic features has also been used for the image classification task (Elhoseiny et al., 2017; Kousha & Brubaker, 2021).

Previously unseen objects can also be learned by deconstructing acquired representations and utilising their parts when seeing novel things. Suglia et al. (2020) and Xu et al. (2021) learn models that recognise out-of-domain objects by learning to compose a new knowledge of representations of previously seen objects and attributes. Also, the descriptiveness and discriminativeness of generated class description can affect the interpreter's performance (Chen et al., 2018; Vedantam et al., 2017; Young et al., 2014). This question is partially investigated in our experiments; see Section 10.5.2.

The closest to ours is the work by Zhang et al. (2018), who train an interpreter

[1] This means that the model has supervised training with visual examples of `seen` classes, and then the model receives text descriptions (one per class) corresponding to the `unseen`. The model is then evaluated in the generalised zero-shot setting. I.e., to classify new visual examples among both `seen` and `unseen` classes. See Xian et al., 2020 for an introduction to different zero-shot learning setups and a recent field survey.

and a speaker to perform continuous learning through direct language interaction. In contrast, our setup is more straightforward in that there is no feedback signal to the describer from the interpreter. We are more interested in the impact that the internal model representations and network types have on the learning process. A different work by Elhoseiny et al. (2013) proposes learning novel concepts without visual representations. For this purpose, they use encyclopaedia entries as information sources since they can fill the information gap when perceptual input is unavailable. Our approach is more challenging as humans often do not have access to extensive textual corpora when interacting in the world.

Our setup with two neural networks is somewhat analogous to the idea of a multi-agent signalling game first introduced by Lewis (1969). The concept of two or more agents developing their language to solve the task has been studied extensively in recent years in NLP (Choi et al., 2018; Lazaridou et al., 2017). The crucial difference with our work is that we do not have a direct learning signal from one model to another, e.g. the agents are not trained simultaneously. Thus, our models do not cooperate in a traditional sense. Instead, we focus on developing a more natural, thus, more complex multi-network *environment* using insights from the research on human cognition and perceptual and communicative grounding. In particular, we (i) explore the ability of the neural model to learn high-level representations of visual concepts, (ii) generate and evaluate concept descriptions from learned representations, and (iii) examine the performance of a different network on the task of interpreting generated descriptions for zero-shot classification.

10.4. Models

Gen and Ipt each have two connected modules: a visual classifier, and a grounded language module. Both networks learn high-quality representations of visual concepts. These representations must be useful for the Gen model as input for a generation. For Ipt, the representations for `seen` classes learned during training help the model extrapolate to interpret descriptions of `unseen` classes.

10.4.1. Label embedding classifier

Both models use a *label embedding classifier*, which represents classes in an embedding. The embedding matrix $\mathbf{V} \in \mathbb{R}^{N \times D}$, contains representations of visual concepts with $N = 200$ being the number of classes and $D = 512$ denoting the size of a single class representation vector.[2] The class embedding parameters, ($\mathbf{V}_G$ for Gen model and $\mathbf{V}_I$ for Ipt) are shared between the classification module and each model's language module, not across models themselves. Both models use the ResNet visual features

[2]We also initialise Gen with $N = 200$ for convenience, but the 20 unseen classes are quickly disregarded during supervised training.

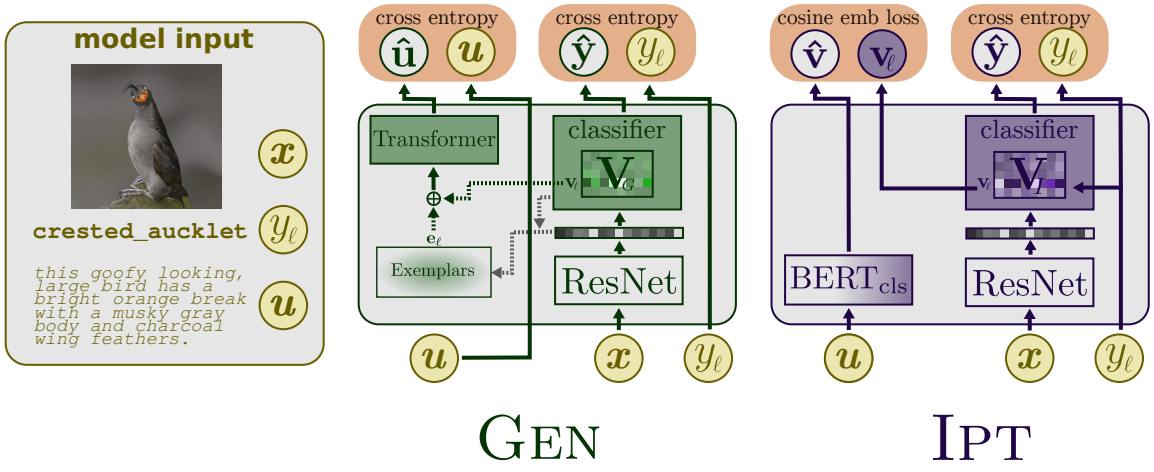


Figure 10.2.: Overview of the training inputs and describer (left) and interpreter (right) models. Learning objectives are pictured above the models. In Gen, dotted lines indicate options for the type of class representation (either $\mathbf{e}_\ell$ $\mathbf{v}_\ell$ or $\mathbf{e}_\ell \oplus \mathbf{v}_\ell$). If the exemplars are used, they are updated based on the classifier at the end of each epoch (gray dotted lines), as described in equation 10.5.

of size 2048 provided by Schönfeld et al. (2019) as inputs to the classifier. These features were extracted from the standard ResNet-101 trained on the ImageNet 1k dataset (Russakovsky et al., 2015). In the following, $\mathbf{x} = \text{ResNet}(\boldsymbol{x})$ is the encoding of input image $\boldsymbol{x}$.

The classifier itself is a simple two-layer feed-forward network which is trained for a multi-class classification task. It takes visual features of the input and concatenates them with each class vector $\mathbf{v}_i$ of $\mathbf{V}$. The network thus outputs N scores, which are converted to class probabilities $\hat{\mathbf{y}}$ with a **softmax** function σ along the label dimension:

$$\hat{\mathbf{y}} = \sigma\big((f_2(f_1(\mathbf{x}) \oplus \mathbf{v}_i))_{i \le N}\big), \tag{10.1}$$

where

$$f_1(\mathbf{x}) = \text{ReLU}(\mathbf{W}_1\,\mathbf{x} + \mathbf{b_1}), \tag{10.2}$$
$$f_2(\mathbf{x}') = \mathbf{W}_3(\text{ReLU}(\mathbf{W}_2\,\mathbf{x}' + \mathbf{b}_2)) \tag{10.3}$$

where $\mathbf{W}_1 \in \mathbb{R}^{2048 \times h_1}$, $\mathbf{W}_2 \in \mathbb{R}^{(h_1+D)\times h_2}$, and $\mathbf{W}_3 \in \mathbb{R}^{h_2 \times 1}$ is the classification output layer.

Both Gen and Ipt use $h_1 = 256$ and $h_2 = 128$.

10.4.2. Generation model

The generation model has two modules: the classifiers described in §10.4.1, and a *decoder* that produces text from a class representation as input. Given a label y_ℓ, the

generation model generates text by passing a class representation, $\mathbf{c}_\ell$, corresponding to the label to the decoder. The class representation is computed differently depending on whether the model uses prototype class representations, exemplars, or both. In all cases, the method of producing a class representation has a relationship to the classifier, but in some cases it is more direct than others.

Gen-Prot is the simplest model to describe. It simply takes the corresponding row of the label embedding $\mathbf{V}_D$, which is also used for classification.

$$\mathbf{c}_\ell = \mathbf{v}_\ell \tag{10.4}$$

Gen-Ex keeps an additional cache of exemplar image features, one for each class, which is updated after each training epoch. The exemplar image for class ℓ is computed as the image that is most certainly part of that class, according to the classifier:

$$\mathbf{e}_\ell = \arg\max(\{\hat{\mathbf{y}}[\ell] \mid \boldsymbol{x} \in X\}) \tag{10.5}$$

and

$$\mathbf{c}_\ell = \mathbf{e}_\ell \tag{10.6}$$

Gen-Both uses the concatenation of the prototype and exemplar representations:

$$\mathbf{c}_\ell = \mathbf{e}_\ell \oplus \mathbf{v}_\ell \tag{10.7}$$

Regarding the model, we train a standard transformer decoder to generate class descriptions (Vaswani et al., 2017). The decoders of the three models differ only in the type of input representations that they are provided with. At each timestep, t, the model's input is updated with previously generated tokens $(w_1, \ldots, w_{t-1})$ and the current token w_t is predicted. We use a standard setup for the transformer: six self-attention layers with eight heads each. We set the learning rate to 4×10^{-4}. We train the model for 20 epochs; the best model is chosen based on the CIDEr score (Vedantam et al., 2015) on the validation set with beam width 2.

Both the classifier and the decoder are trained jointly with the standard cross-entropy loss:

$$\text{loss}_D = \text{CrossEntropy}(\hat{\mathbf{y}}, \mathbb{1}(\boldsymbol{y}_i)), \tag{10.8}$$

where l is the loss value, $\hat{\mathbf{y}}$ is either the model's logits or output of the classifier, and $\boldsymbol{y}_i$ is the ground-truth representation for either of the tasks. The generation model is trained with teacher forcing during training. For inference, we test multiple decoding algorithms described below.

10.4.3. Decoding algorithms

The quality of the generated texts, specifically their content, is crucial for the successful interaction between the describer and the interpreter. To a large degree, generation depends on the decoding algorithm used. Some of these algorithms can generate more accurate texts, while others introduce randomness and diversity, making the output more "safe" or generic and possibly hurting discriminability (Zarrieß et al., 2021). Here we examine *two* decoding algorithms, which introduce different conditions for accuracy and diversity. While greedy search can generate accurate descriptions, it is sub-optimal on the sentence level, e.g. longer generations tend to be repetitive and "boring" (Gu et al., 2017), therefore, we are not using this search.

Beam search is often used as a standard decoding algorithm because it suffers much less from the problems occurring during long-text generation. At each generation step i, it keeps track of several candidate sequences $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_k)$ and picks the best one based on the cumulative probability score of generated words per sentence:

$$\mathbf{c}_i = \underset{\substack{\mathbf{c}'_i \subseteq \mathcal{B}_i, \\ |\mathbf{c}'_i| = k}}{\operatorname{argmax}} \log p(\mathbf{c}'_i \mid \mathbf{c}_{i-1}, \mathbf{v}_i; \theta). \tag{10.9}$$

The parameter k is used to control the depth of the search tree, and $\mathcal{B}$ is the set of candidate sequences. While beam search generally outperforms greedy, higher k can lead to texts with low diversity (Li et al., 2016). To evaluate whether "more diverse" means "more descriptive" in the context of our two-agent set-up, we generate texts with **nucleus sampling** method (Holtzman et al., 2020) which samples tokens from the part of the vocabulary defined based on the probability mass:

$$p' = \sum_{w_i \in \mathcal{V}'} \log p(w_i \mid \mathbf{w}_{<i}, \mathbf{v}_i; \theta) \geq p, \tag{10.10}$$

where p determines the probability mass value, while $\mathcal{V}'$ is part of the vocabulary $\mathcal{V}$ which accumulates the mass at the timestamp i. Next, a new distribution P is produced to sample the next token:

$$P = \begin{cases} \log p(w_i \mid \mathbf{w}_{<i}, \mathbf{v}_i; \theta)/p' & \text{if } w_i \in \mathcal{V}' \\ 0 & \text{otherwise.} \end{cases} \tag{10.11}$$

With nucleus sampling, we aim to generate more diverse texts than those generated with greedy or beam search. By evaluating the interpreter with texts generated by different algorithms, we consider the impact of generation on the success of information transfer using text from the describer to the interpreter.

10.4.4. Interpretation model

The Ipt model also has two modules: a label embedding classifier with a weight matrix $\mathbf{V}_G \in \mathbb{R}^{N \times D}$, and an interpretation module that maps text to vectors of size D. Ipt uses `[CLS]` token representations extracted from Bert model as text features. In preliminary experiments on the ground-truth test data, we found that the Ipt benefitted greatly from using features from a Bert model (Devlin et al., 2019) which was fine-tuned on descriptions from the `seen` portion of the training set. We fine-tuned the final layer with a learning rate of 2×10^{-5} and weight decay of 0.01 for 25 epochs using the Adam optimiser (Kingma et al., 2015). Rather than the standard masked language modelling objective, the model was fine-tuned using a text classification task among the `seen` classes. Since Bert is not visually grounded, we speculate that this the pre-training task may help it to attend to the visually relevant information in the descriptions in a way that results in an informative `[CLS]` representation. Given a text description $\boldsymbol{u}$, we use $\mathbf{u}$ for the `[CLS]` features (with size 768) extracted from the fine-tuned Bert model.

The interpretation module, then, is a simple fully-connected layer with Tanh activation:

$$\hat{\mathbf{v}} = \text{Tanh}(\mathbf{W}\,\mathbf{u} + \mathbf{b}) \tag{10.12}$$

Given a training example $(\mathbf{x}, y_\ell, \boldsymbol{u})$, the classifier makes a class prediction $\hat{\mathbf{y}}$ from $\mathbf{x}$ and the interpreter predicts the class representation $\hat{\mathbf{v}}$ from $\boldsymbol{u}$. We draw a vector $\mathbf{v}_k$ from $\mathbf{V}_I$ so that with a frequency of 0.5, it is a negative sample (i.e., $k \neq \ell$) and the other half the time $k = \ell$.

The two modules are trained jointly. The loss term for the classifier is computed with the standard cross entropy loss and the term for the interpreter is computed with cosine embedding loss, a variation of hinge loss defined below. The overall loss is computed as follows:

$$\begin{aligned} \text{loss}_I = &\,\text{CrossEntropy}(\hat{\mathbf{y}}, y_\ell) + \\ &\,\text{CosineEmbLoss}(\hat{\mathbf{v}}, \mathbf{v}_k), \end{aligned} \tag{10.13}$$

where

$$\text{CosineEmbLoss}(\hat{\mathbf{v}}, \mathbf{v}_k) = \begin{cases} 1 - \text{Cos}(\hat{\mathbf{v}}, \mathbf{v}_k) & \text{if } k = \ell \\ \max\left(0, \text{Cos}(\hat{\mathbf{v}}, \mathbf{v}_k) - \delta\right) & \text{if } k \neq \ell \end{cases} \tag{10.14}$$

Like hinge loss, the cosine embedding loss includes a margin δ, which we set to 0.1. Intuitively, δ prevents the loss function from penalising the model for placing its class representation prediction close to the representation of a nearby negative class, as long as it isn't too close. After all, some classes *are* similar. The best Ipt model is chosen based on the mean rank for the validation set.

10.5. Experiments

10.5.1. Data

We use the Caltech-UCSD Birds-200-2011 dataset (Akata et al., 2016, hereafter CUB), a collection of $11\,788$ images of birds from 200 different species. The images were collected on Flickr, by searching for the name of the species, and filtered by a process involving crowd workers. In addition to class labels, the dataset includes bounding boxes and attributes, but we do not use those features in the current study, since our focus is on using natural language descriptions for zero-shot classification, rather than from structured attribute-value features.

We also use a corpus of English-language descriptions of the images in the CUB dataset, collected by (Reed et al., 2016). The corpus contains 10 descriptions of each image. The descriptions were written in such a way that they are both very precise (annotators were given a diagram labelling different parts of a bird's body to aid in writing descriptions) and very general (annotators were asked not to describe the background of the image or actions of the particular bird). This allows us to treat the captions as *class descriptions*, suitable for zero-shot classification. We split the dataset into 180 `seen` and 20 `unseen` classes and train, test, and validation sets of each (Table 10.1).

	`seen`	`unseen`	Total
Train	8482	948	9430
Test	1060	119	1179
Val	1060	119	1179
Total	10 602	1186	11 788

Table 10.1.: Number of CUB corpus images by data split.

A single training example is a triple $(\boldsymbol{x}, y, \boldsymbol{d})$, consisting of an image, class label, and description. Since there are 10 descriptions per image, this gives us 94 300 training examples for the generator, which is trained on both the `seen` and `unseen` classes, and 84 820 training examples for the zero-shot interpreter. To mitigate the possibility that `unseen` represents a particularly hard or easy subset of unseen classes, we test 5 folds, each disjointed in their `unseen` classes. All results are reported as the mean over five folds.

10.5.2. Evaluation metrics

Generation and classification We evaluate the generation model with BLEU (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015), which has been shown to correlate the most with human judgements when evaluating multi-modal models. The interpreter is evaluated with different degrees of accuracy, 1, 5 and 10. To give us more information about the model's learning capability, we also use mean rank to indicate how close each model is to the correct prediction in case of incorrect output.

Discriminativity Our generation model is trained to minimise the cross entropy of the next token, given the class label. This learning objective may encourage the model to generate "safe" descriptions, as opposed to descriptions that mention features that would help to identify birds of the given class. To measure this tendency, we define a notion of the *discriminativity* of a class description, which evaluates how helpful the description is in picking out instances of the class it describes. We measure the ability of the describer to generate both category- and instance-aware descriptions, which need to have a specific level of concreteness and generality to be helpful for the interpreter. To compute the metric, we first extract features from the descriptions, where each feature consists of the noun and the set of adjectives used in a noun phrase.[3] We define the discriminativity of a feature as the exponential of the information gain of the feature for the bird class, as measured on the test set; that is,

$$\text{disc}(x_i) = \exp\big(H(Y) - H(Y|x_i)\big),$$

where x is a feature and Y is the bird class.

The maximum discriminativity of a feature (i.e., a feature that uniquely picks out a particular class) is equal to the number of classes, 200. For example,

$$\text{disc}((\text{'bill'}, \{\text{'long'}, \text{'curved'}\})) = 22.9,$$

whereas

$$\text{disc}((\text{'bill'}, \{\text{'short'}, \text{'pointy'}\})) = 2.9,$$

[3] For details on how the noun phrases are extracted, see [paper GitHub repository].

reflecting the fact that more kinds of birds have short pointy bills than long curved bills. We define two metrics for the discriminativity, disc_{max}, and disc_{avg}, which are the maximum and mean discriminativity of the mentioned features, respectively.

10.6. Results

First, we assess the zero-shot performance of the Ipt models to get an idea of how well the models generalise to classifying `unseen` classes based on class descriptions (Table 10.3). We compare the performance of the model using descriptions generated by Gen the model using ground truth descriptions from the unseen portion of the test set. Since multiple descriptions exist per class, we randomly select one for each unseen class in each zero-shot fold.

teacher		mean rank	acc@1		acc@5		acc@10	
random baseline	`seen`	100	0.005	0.005	0.025	0.025	0.05	0.05
	`unseen`	100	0.005		0.025		0.05	
ground truth	`seen`	5.414	0.364	0.251	0.751	0.556	0.888	0.695
	`unseen`	29.766	0.191		0.441		0.571	
best Gen	`seen`	5.504	0.392	0.116	0.748	0.294	0.875	0.419
	`unseen`	70.497	0.068		0.183		0.276	

Table 10.2.: Zero-shot classification results for the Ipt model. These results are computed as macro-averages, averaged first over class, then over the fold. The harmonic mean of the accuracy metrics is reported to the right of the `seen` and `unseen` scores. The ground truth results report zero-shot performance after learning from one randomly sampled ground truth description for each `unseen` class. The best Gen results report zero-shot performance after learning from the best Gen model (Gen-Ex with beam-2 decoding). See Table 10.3 for a full comparison of the generation models, including resulting Ipt performance.

Our first observation is that the model is moderately successful on the zero-shot classification task. When learning from the ground truth descriptions, the model performs well above the random baseline. While 0.19 is not very high for classification accuracy in general, it is not out of line for `unseen` results in zero-shot learning. We should keep in mind that classifying a bird as belonging to one of 200 species based only on a textual description would be a difficult task even for most humans. That the model can use the ground truth text descriptions to learn class representations that are *somewhat* useful for image classification is encouraging for the prospect of using it to evaluate the Gen models. However, we note that the performance of the model using descriptions generated from the best Gen model is quite a lot worse than the ground

truth. This suggests that while the descriptions generated by the best Gen models are not totally useless, they are nevertheless failing to communicate as well as they could.

		Bleu1	Bleu2	Bleu3	Bleu4	CIDEr	mean rank	acc@1	acc@5	acc@10	Disc.
Gen-Both	beam	0.63	0.61	0.57	0.53	7.88	94.43	0.00	0.02	0.06	1.56
	nucleus	0.69	0.58	0.45	0.32	**8.99**	111.25	0.01	0.04	0.07	4.85
Gen-Ex	beam	0.64	**0.62**	**0.60**	**0.58**	5.44	**46.41**	**0.07**	**0.24**	**0.39**	2.15
	nucleus	0.65	0.57	0.47	0.36	6.55	70.50	0.07	0.18	0.28	4.44
Gen-Prot	beam	0.61	0.60	0.58	0.55	6.67	72.63	0.03	0.14	0.23	2.43
	nucleus	**0.70**	0.61	0.50	0.38	7.10	75.24	0.04	0.15	0.23	**4.93**

Table 10.3.: Each generation model uses the same classifier module and decoder architecture, but conditioned on internal class representations created by three different methods. For each architecture we asses two different decoding algorithms. BLEU and CIDEr scores are reported as micro averages over n-grams produced in all 200 class descriptions.

Next, we turn to a comparison of the different Gen models. We can see that overall, Gen-Ex performed best on most of the intrinsic metrics (with the notable exception of BLEU-1) and also in terms of communicative success. Overall, the beam performed better than the nucleus on the intrinsic metrics, and for Gen-Ex it was quite a bit better for communicative success. Interestingly, nucleus-generated texts nevertheless scored much higher in terms of discriminativity. Gen-Prot and Gen-Both performed similarly on the intrinsic metrics, but Gen-Both performed extremely poorly (worse than random baseline in some cases) in terms of communicative success.

10.7. Discussion and conclusion

One of the motivations for adopting this task was that it relies on specifically *class-level* representations in contrast to many language-and-vision tasks that ground in particular images. We wanted to see how well the models can ground language when there aren't any pixels to attend to. It turns out that our models struggled. Strikingly, the models that performed the best were the Gen-Ex models, which essentially transforms the task into one that does involve particular images by picking out exemplars to generate from. Of course, the models we used in this study were relatively simple, and more sophisticated neural models may produce different results. But this raises an interesting question for future work — what *does* it takes to learn grounded representations that are useful in this particular communicative context?

More generally, why are the Gen model descriptions not as good as the ground truth for the zero-shot performance of the Ipt model? There are two possible explanations. One is that the generated descriptions don't carry the visual information that would be needed to classify among unseen classes successfully. Another possibility is that

the text the Gen model produces is not interpretable by the Ipt model. Recall that the Ipt model was trained on ground truth descriptions from `seen`. These descriptions have a structure to them — certain regularities in conveying visual information. If Gen descriptions don't follow that structure, Ipt won't be able to make good use of them, even if they do in some sense "contain" visual information. Indeed, there is some evidence that this is what is happening. We see that nucleus sampling resulted in higher discriminativity scores, including for the Gen-Prot and Gen-Both models. So, although the generator produces sequences containing adjective-noun phrases that pick out the correct class, Ipt cannot make good use of them, perhaps because they appear in texts that are "ungrammatical" for the distribution Ipt was trained on.

A different question to ask is whether generation metrics reflect the communicative power of generated texts as measured by the interpreter's performance. As Table 10.3 demonstrates, Ipt performs the with beam-generated texts. However, these texts overall score very low on discriminativity. Indeed, we found out that beam search generates sentences which mention features which are very common among classes, e.g. "a bird with wings". At the same time, Ipt benefits more from nucleus-generated texts produced by Gen-Ex and Gen-Both models. These texts are more diverse, possibly describing a larger set of class features and our interpreter is able to learn better from that diversity. However, nucleus-generated texts are generally rated lower by intrinsic generation metrics, suggesting a mismatch between task-based evaluation (e.g., interpreter's performance) and intrinsic evaluation (e.g., generation metrics). These results indicate that the "groundedness" of class descriptions and their use for the task might not be adequately captured by the set of NLG metrics and one might want to use the generated texts "in context" (aka interpretation) to get a clearer picture on how much important information such texts carry.

In future work, we plan to investigate whether the interpretation performance can be improved by making sure the model pays attention to differences between classes on a much more fine-grained level of features. Inspecting how the generation of better, more descriptive and more discriminative class descriptions can be achieved is also important. It is worth examining how a more interactive context affects generated class descriptions and their interpretation. This could be studied in a reinforcement learning setup (Oroojlooy & Hajinezhad, 2021) where models receive a reward for producing class descriptions that lead to better zero-shot classification performance.

10.8. Limitations

Here we enumerate some limitations that we think our work has. First, as we were focused on the task and representation learning, we did not expanded our experiments to more complex models. Our analysis can also be performed in the context of different encoder-decoder combinations. The dataset, on the one hand, has fine-grained descriptions of categories. On the other hand, they can be so specific that they lack

a generality, which depends on the domain and even personal preferences and background of those who interact. While this does correspond to the situation in certain real-life situations (bird classification being one of them), the results may look different in a more open-domain setup, such as in the dataset from Bujwid and Sullivan (2021). Experimentation with more datasets of different levels of specificity and with different kinds of ground truth classes descriptions would be necessary to draw more general conclusions.

References

Agrawal, A., Batra, D., Parikh, D., & Kembhavi, A. (2018). Don't just assume; look and answer: Overcoming priors for visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4971–4980. https://openaccess.thecvf.com/content_cvpr_2018/papers/Agrawal_Dont_Just_Assume_CVPR_2018_paper.pdf

Akata, Z., Perronnin, F., Harchaoui, Z., & Schmid, C. (2016). Label-Embedding for Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(7), 1425–1438. https://doi.org/10.1109/TPAMI.2015.2487986

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., … Simonyan, K. (2022). Flamingo: A Visual Language Model for Few-Shot Learning.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2425–2433. https://openaccess.thecvf.com/content_iccv_2015/papers/Antol_VQA_Visual_Question_ICCV_2015_paper.pdf

Beinborn, L., Botschen, T., & Gurevych, I. (2018). Multimodal grounding for language processing. *Proceedings of the 27th International Conference on Computational Linguistics*, 2325–2339. https://aclanthology.org/C18-1197

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., & Turian, J. (2020). Experience grounds language. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8718–8735. https://doi.org/10.18653/v1/2020.emnlp-main.703

Blank, H., & Bayer, J. (2022). Functional imaging analyses reveal prototype and exemplar representations in a perceptual single-category task. *Communications Biology*, *5*(1), 1–13. https://doi.org/10.1038/s42003-022-03858-z

Bujwid, S., & Sullivan, J. (2021). Large-Scale Zero-Shot Image Classification from Rich and Diverse Textual Descriptions. *Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, 38–52.

Carey, S. (1981). The child as word learner. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic Theory and Psychological Reality* (First Paperback Edition, pp. 264–293). The MIT Press.

Chandu, K. R., Bisk, Y., & Black, A. W. (2021). Grounding 'grounding' in NLP. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4283–4305. https://doi.org/10.18653/v1/2021.findings-acl.375

Chen, F., Ji, R., Sun, X., Wu, Y., & Su, J. (2018). Groupcap: Group-based image captioning with structured relevance and diversity constraints. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1345–1353. https://openaccess.thecvf.com/content_cvpr_2018/papers/Chen_GroupCap_Group-Based_Image_CVPR_2018_paper.pdf

Choi, E., Lazaridou, A., & de Freitas, N. (2018). Compositional obverter communication learning from raw visual input. *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. https://openreview.net/forum?id=rknt2Be0-

Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2018). Embodied question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–10. https://openaccess.thecvf.com/content_cvpr_2018/papers/Das_Embodied_Question_Answering_CVPR_2018_paper.pdf

Desai, K., & Johnson, J. (2021). Virtex: Learning visual representations from textual annotations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11162–11173.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Elhoseiny, M., Saleh, B., & Elgammal, A. (2013). Write a Classifier: Zero-Shot Learning Using Purely Textual Descriptions. *2013 IEEE International Conference on Computer Vision*, 2584–2591. https://doi.org/10.1109/ICCV.2013.321

Elhoseiny, M., Zhu, Y., Zhang, H., & Elgammal, A. M. (2017). Link the head to the "beak": Zero shot learning from noisy text description at part precision. *CoRR*, *abs/1709.01148*. http://arxiv.org/abs/1709.01148

Gelman, S. A., & Brandone, A. C. (2010). Fast-mapping placeholders: Using words to talk about kinds. *Language learning and development : the official journal of the Society for Language Development*, *6*(3), 223–240. https://doi.org/10.1080/15475441.2010.484413

Gu, J., Cho, K., & Li, V. O. (2017). Trainable greedy decoding for neural machine translation. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1968–1978. https://doi.org/10.18653/v1/D17-1210

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346. https://doi.org/https://doi.org/10.1016/0167-2789(90)90087-6

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. https://openreview.net/forum?id=rygGQyrFvH

Kennington, C., & Schlangen, D. (2015). Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 292–301. https://doi.org/10.3115/v1/P15-1029

Kingma, D. P., Ba, J., & Amsterdam Machine Learning lab (IVI, FNWI). (2015). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*.

Kousha, S., & Brubaker, M. A. (2021). Zero-shot learning with class description regularization. https://doi.org/10.48550/ARXIV.2106.16108

Larsson, S. (2013). Formal semantics for perceptual classification. *Journal of Logic and Computation*, *25*(2), 335–369. https://doi.org/10.1093/logcom/ext059

Lazaridou, A., Peysakhovich, A., & Baroni, M. (2017). Multi-agent cooperation and the emergence of (natural) language. *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. https://openreview.net/forum?id=Hk8N3Sclg

Lewis, D. K. (1969). *Convention: A philosophical study*. Cambridge, MA, USA: Wiley-Blackwell.

Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119. https://doi.org/10.18653/v1/N16-1014

Lu, J., Xiong, C., Parikh, D., & Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 375–383. https:

//openaccess.thecvf.com/content_cvpr_2017/papers/Lu_Knowing_When_to_CVPR_2017_paper.pdf

Malt, B. C. (1989). An on-line investigation of prototype and exemplar strategies in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 539–555. https://doi.org/10.1037/0278-7393.15.4.539

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *10*(1), 104–114. https://doi.org/10.1037//0278-7393.10.1.104

Oroojlooy, A., & Hajinezhad, D. (2021). A Review of Cooperative Multi-Agent Deep Reinforcement Learning. *arXiv:1908.03963 [cs, math, stat]*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Paz-Argaman, T., Tsarfaty, R., Chechik, G., & Atzmon, Y. (2020). ZEST: Zero-shot learning from text descriptions using textual similarity and visual summarization. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 569–579. https://doi.org/10.18653/v1/2020.findings-emnlp.50

Reed, S., Akata, Z., Lee, H., & Schiele, B. (2016). Learning Deep Representations of Fine-Grained Visual Descriptions. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 49–58. https://doi.org/10.1109/CVPR.2016.13

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object Hallucination in Image Captioning. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4035–4045. https://doi.org/10.18653/v1/D18-1437

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, *7*(4), 532–547. https://doi.org/10.1016/0010-0285(75)90021-3

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Schlangen, D., Zarrieß, S., & Kennington, C. (2016). Resolving References to Objects in Photographs using the Words-As-Classifiers Model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1213–1223. https://doi.org/10.18653/v1/P16-1115

Schönfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., & Akata, Z. (2019). Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders.

Silberer, C., Ferrari, V., & Lapata, M. (2017). Visually Grounded Meaning Representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2284–2297. https://doi.org/10.1109/TPAMI.2016.2635138

Suglia, A., Vergari, A., Konstas, I., Bisk, Y., Bastianelli, E., Vanzo, A., & Lemon, O. (2020). Imagining grounded conceptual representations from perceptual information in situated guessing games. *Proceedings of the 28th International Conference on Computational Linguistics*, 1090–1102. https://doi.org/10.18653/v1/2020.coling-main.95

Thomason, J., Gordon, D., & Bisk, Y. (2019). Shifting the baseline: Single modality performance on visual navigation & QA. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1977–1983. https://doi.org/10.18653/v1/N19-1197

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vedantam, R., Bengio, S., Murphy, K., Parikh, D., & Chechik, G. (2017). Context-aware captions from context-agnostic supervision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 251–260.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4566–4575. https://openaccess.thecvf.com/content_cvpr_2015/papers/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.pdf

Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2020). Zero-Shot Learning – A Comprehensive Evaluation of the Good, the Bad and the Ugly.

Xu, G., Kordjamshidi, P., & Chai, J. (2021). Zero-shot compositional concept learning. *Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing*, 19–27. https://doi.org/10.18653/v1/2021.metanlp-1.3

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, *2*, 67–78. https://doi.org/10.1162/tacl_a_00166

Zarrieß, S., Voigt, H., & Schüz, S. (2021). Decoding methods in neural language generation: A survey. *Information*, *12*(9). https://doi.org/10.3390/info12090355

Zhang, H., Yu, H., & Xu, W. (2018). Interactive language acquisition with one-shot visual concept learning through a conversational game. *Proceedings of the 56th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2609–2619. https://doi.org/10.18653/v1/P18-1243

11. Personae under uncertainty: The case of topoi

Bill Noble, Ellen Breitholtz, and Robin Cooper

Abstract In this paper, we propose a probabilistic model of social signalling which adopts a persona-based account of social meaning. We use this model to develop a socio-semantic theory of conventionalised reasoning patterns, known as *topoi*. On this account the social meaning of a topos, as conveyed in a argument, is based on the set of ideologically-related topoi it indicates in context. We draw a connection between the role of personae in social meaning and the *category adjustment effect*, a well-known psychological phenomenon in which the representation of a stimulus is biased in the direction of the category in which it falls. Finally, we situate the interpretation of social signals as an update to the information state of an agent in a formal TTR model of dialogue.

11.1. Introduction

Consider the (somewhat dramatic) Example 11.1, from Lavelle et al. (2012), a corpus of dialogues where participants are instructed to resolve a moral dilemma. The subjects are asked to decide, based on limited information, who out of four passengers in a hot air balloon to sacrifice in order to save the other three.
Apart from communicating semantic content, arguments often implicitly evoke a *topos*, a pattern of reasoning the speaker draws on to warrant their argument. For example, the argument against sacrificing the pregnant woman (11.1-51) relies on a topos such as *if you have to choose between killing* n *and* m *people and* $m < n$ *then choose* m.

Upon recognizing the evoked topos, an interlocutor may draw certain conclusions about the speaker, namely that they are the *kind of person* who reasons in this way. Given that information, the interlocutor may, in turn, choose to frame their arguments in a way that appeals to the kind of person they infer the speaker to be.

Such topoi can be seen as signals conveying *social meaning* by association with *personae* or stereotypical categories of people (Eckert, 2012). The use of personae as a *semantic medium* in a theory of social meaning is analogous to how possible worlds (Lewis, 1970), infons, or situation types (Barwise & Perry, 1983) are used in truth-theoretic accounts of propositional meaning. Just as declarative sentences restrict the

Example 11.1.

39 C: Well I'm not throwing a kid out [I just couldn't cope with it].

42 A: And the other thing is I mean what what what she achieves er in her life if she becomes as famous as famous as Mozart erm will go on er [forever]=

45 A: So I mean the person it seems like the person with least value is the .

48 B: [she's] pregnant.

51 B: [So you're] killing two people instead of one.

52 C: Yhh and another thing is would he be able to pilot the balloon if his wife is overboard?

set of possible worlds or situation types, the social meaning of a social signal restricts the personae attributed to the speaker. Recent work by Burnett (2017), for example, uses game theoretic modelling to formalise social meaning in terms of personae. In contrast to Burnett (2017), who considers dialectical variables orthogonal to semantic content, we consider the social meaning of topoi in argumentation, following Breitholtz (2014). We develop a probabilistic model that formalises the relationship between topoi and personae through Bayesian inference and integrate this account into a formal TTR model of dialogue by defining an update to the information state of an agent.

11.2. Personae, topoi, and social meaning

In this section, we give background on the rhetorical and sociolinguistic phenomena we seek to model.

11.2.1. Personae

The variationist branch of sociolinguistics is interested in the construction of linguistic style through the use of *linguistic variables* Hudson, 1996. A variable is any axis along which an individual's language may differ from someone else in the same community. Linguistic variables can be found at all levels of linguistic analysis, including phonetics (e.g., accent), prosody, lexical choice, morphology, and syntax.

Some of the earliest work in variationist sociolinguistics, for example, studies phonetic variations different groups of speakers on the island of Martha's Vineyard (Labov, 1963). This *first wave* of variationist sociolinguistics (Eckert, 2012), is principally

concerned with variation across macrosociological categories such as race, class and gender .

The second wave of variationist study was interested in more fine-grained social categories, sometimes referred to as *personae* Eckert, 2012. A persona is a widely recognised social category which is available as a reference point for the expression of social identity in a given community. For example, Eckert (1989, 2008) identifies the personae of "jock" and "burnout" as central to the social semiotic system of an early-2000s Detroit-area high school. Through their dress, behaviour, and linguistic style, students signal identification with or distance from the established personae.

Third-wave sociolinguistics considers the role of variation in the expression of social meaning, rather than merely reflective of social categories Eckert, 2012. Personae are the semantic common ground that makes communicating social meaning possible. In a given speech community, a linguistic *variant* (the expression of a linguistic variable) constitutes a *social signal* in virtue of its association with one or more personae. Speakers identify themselves as ideologically aligned with a given persona by adopting variants associated with it. This is referred to as *projecting* a persona. Speakers typically do not identify uniquely with one persona, however. Each individual constructs a unique style, mixing and matching variants associated with different personae in a process Eckert (2000) refers to as *bricolage*.

While previous work assumed that linguistic variables were orthogonal to propositional meaning, third-wave sociolinguistics acknowledges that that separation is not always possible. Eckert (2008) writes that her view of linguistic style "precludes the separation of form from content, for the social is eminently about the content of people's lives". In the following section we present *topoi*, a pragmatic phenomenon that play a role in semantic content, but that we argue can also be viewed as a constituent of linguistic style.

11.2.2. Topoi

Argumentation and reasoning in dialogue is predominantly *enthymematic*, that is, it partly relies on what is "in the mind" ἐμθύμημα of the listener Breitholtz, 2014. Aristotle referred to the principles of reasoning which enthymematic arguments are based on as the *topoi* of the arguments. For Aristotle, a topos was a "place" or "field", where a public speaker or a participant in a dialectic debate could find ideas on which to build his argument.

In the 20th century the idea of topos has been taken up in linguistics by Ducrot (1980) and Anscombre et al. (1995) who suggest that every link between a statement and another statement, or between a statement and (for example) an exhortation in discourse is a topos and that topoi are thus essential to any theory of semantics beyond the sentence, as well as important for contextual interpretation of lexical meaning. One of the leading ideas in Ducrot's take on topoi, is that topoi are not part of factual

knowledge about the world, but part of "ideology", that is the agent's conception of acceptable ways to make inferences. This does not mean that topoi are unrelated to facts—for example, a topos of gravity is not likely to be unrelated to the way gravity works. However, it is clear that a large number of topoi are related to ethical considerations such as what is good or beneficial, and these cases are clearly ideological in the sense that they are relative to context.

For example, Ducrot discusses different ways of arguing about giving tips. One individual might encourage another to give a tip to a porter who "carried the bags all the way here", while someone else might advise against it, for the reason that the porter is already paid to carry bags, and why should you pay someone for something they are already paid to do? This is an example of how different topoi may apply in one situation, and lead to inconsistent results or conclusions. Which topoi we appear to draw on while making an argument in a given situation thus gives our interlocutors information of an ideological nature. This is true both in situations where we reason from a context (a set of premises present in a context) to a conclusion, and when we have a particular conclusion in mind that we argue for. In the first case, applying different topoi might lead to different conclusions, but in both cases the implicit ideological information conveyed might differ depending on the topos used.

We argue that topoi, in virtue of their ideological association, constitute social signals that contribute to the persona projected by a dialogue participant, much like use of particular linguistic variants. Topoi are an attractive subject of study as social signals since, unlike social variables like physical appearance or pronunciation, they may be extracted from written text or transcribed dialogues.

11.3. Two probabilistic models of social meaning

In this section, we develop a simple probabilistic model that associates topoi with personae. In particular, we model how the use of a topos by one agent results in an update to another agent's model of their persona. Since we restrict our attention to a single utterance, we refer to the listener, whose internal state is updated, as *Self* and the speaker, who evoked the topos, as *Other*.

We present two versions of the model. In the *first-order model*, Self models Other as a simple categorical probability distribution over personae. In the *second-order model*, Self represents Other as a Dirichlet distribution over *possible* categorical distributions over personae.

In both cases, the event being modelled is the same: Other (O) invokes a topos (τ) in a dialogue with Self (S). Then, Self's updates their model of Other as a result of that social signal.

Unlike the social signaling game from Burnett (2017)'s, which is based on rational speech acts (Frank & Goodman, 2012), we do not assume any level of social recursion in the speaker; that is, the speaker does not consult a model of the listener's model of

themself when producing an utterance.

We assume that each agent has access to a set of personae, $\Pi = \{\pi_1, ..., \pi_K\}$, and topoi, $\Psi = \{\tau_1, ..., \tau_N\}$. A probability distribution φ_π is assigned to each persona π such that:

$$\varphi_\pi(\tau) = P_S(\tau \mid \pi). \tag{11.1}$$

The probability given by $\varphi_\pi(\tau)$ is the likelihood that someone projecting π will evoke τ. This distribution models the idealogical association between topoi and personae—it is what gives the topoi their social meaning. For now, we assume that Π, Ψ and $\boldsymbol{\varphi}$ are shared community resources.

We begin with the first-order model as a demonstration of the setting and, after discussing its weaknesses, move on to the second-order model.

11.3.1. First-order model

In the first-order model, Self models Other as categorical probability distribution over personae. Let $\theta_{S,O}$ be S's model of O; that is, the probability, according to S, that O will project the persona π:

$$\theta_{S,O}(\pi) = P_S(\pi \mid O) \tag{11.2}$$

When O evokes τ, S updates their prior model of O accordingly. Intuitively, S learns that O is more likely to project personae that are likely to evoke τ:

$$\Delta_1(\theta_{S,O}, \varphi, \tau) = \lambda\pi.\frac{\varphi_\pi(\tau) \cdot \theta_{S,O}(\pi)}{\sum_{\pi'} \varphi_{\pi'}(\tau) \cdot \theta_{S,O}(\pi')} \tag{11.3}$$

In Bayesian terms, the update function gives the posterior distribution of $\theta_{S,O}$, given τ:

$$\begin{aligned}\frac{\varphi_\pi(\tau) \cdot \theta_{S,O}(\pi)}{\sum_{\pi'} \varphi_{\pi'}(\tau) \cdot \theta_{S,O}(\pi')} =& \frac{P(\tau \mid \pi) \cdot P_S(\pi \mid O)}{P_S(\tau)} \\ =& P_S(\pi \mid \tau, O)\end{aligned}$$

To make the situation more concrete, consider again utterance 51 from Example 11.1. Among the topoi elicited by this utterance is the assumption that, given the choice, it's always better to kill fewer people. Let's call this utterance τ_3 and let $\Delta_1(\theta_{S,O}, \varphi, \tau_3) = \hat{\theta}_{S,O}$.

We may imagine any number of personae associated with τ_3, but most relevant are those personae based on different kinds of moral reasoning. In this case, S believes that the *humanist* and *cold rationalist* personae give some prior probability to the evoked topos (see figure 11.1). Self updates their model of Other in proportion to the product of the likelihood of the topos given the persona, and the persona's prior probability for Other.

In this first-order model, $\theta_{S,O}$ has two possible interpretations:

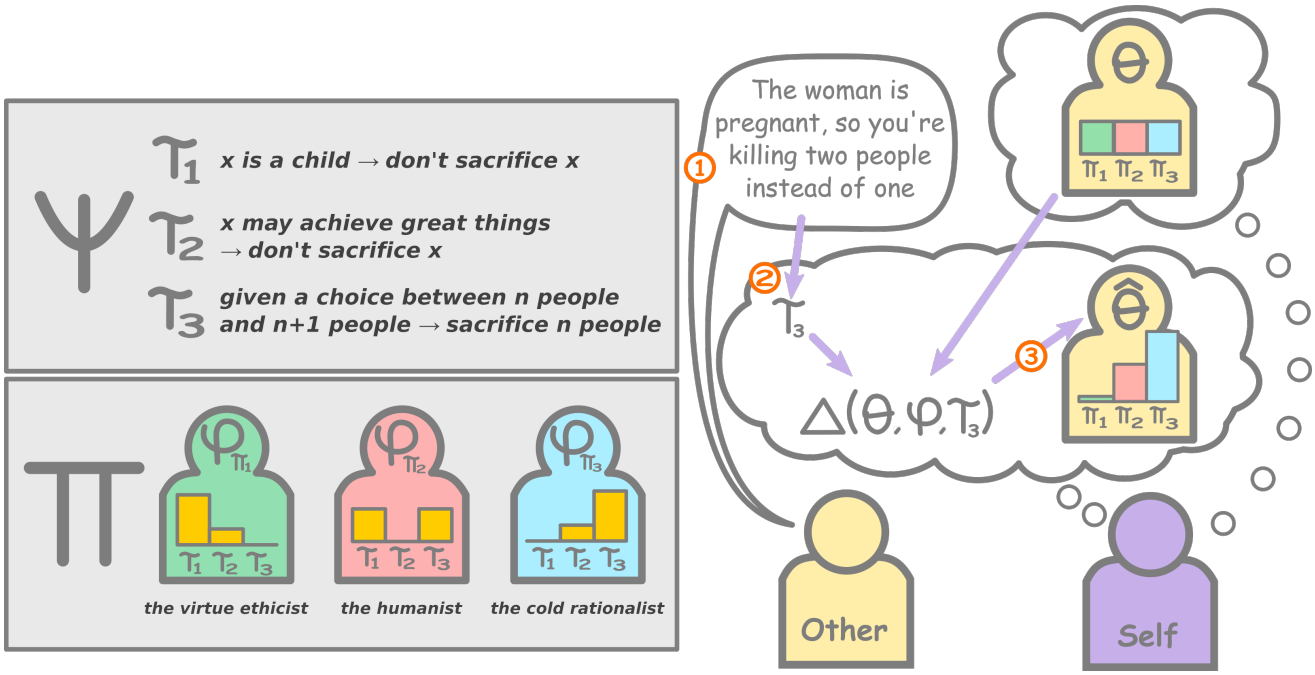


Figure 11.1.: Using the shared topoi, personae, and topos distribution for personae (left), Self updates their representation of Other (right) as follows: (1) Other utters 11.1-51. (2) Self interprets 11.1-51 as evoking τ_3. (3) Other applies the update function from Equation 11.3, incorporating their prior model of Other, the topos distributions for personae, and the evoked topos.

1. It represents Self's uncertainty about which persona Other projects (but Self assumes that Other uniquely projects one persona).

2. It represents Self's belief about Other's persona tendencies—i.e., their *bricolage* (but no uncertainty is modelled).

Both of these interpretations have drawbacks. The (false) assumption that each person projects a unique persona results in inconsistency when an agent observes an interlocutor evoke both τ_1 and τ_2 that don't appear in any of the same personae. However, if $\theta_{S,O}$ instead represents Self's take on Other's bricolage of personae, the lack of uncertainty leaves the Bayesian belief revision given by Equation 11.3 unfounded. To simultaneously account for bricolage and uncertainty, we must add a second layer of analysis to the agent model.

11.3.2. Second-order model

In the second-order model, we assume that Self attributes some particular distribution over personae to Other, but that their representation captures uncertainty about exactly what distribution it is. Thus, instead of a prior over personae, S's model of O is a prior over distributions over personae. For this, we use a Dirichlet distribution parametrized by K-dimensional positive real-valued $\boldsymbol{\alpha}_{S,O}$.

The Dirichlet distribution is a probability density function defined as follows:

$$f(\theta; \boldsymbol{\alpha}_{S,O}) = \frac{1}{B(\boldsymbol{\alpha}_{S,O})} \prod_{i=1}^{K} \theta(\pi_i)^{\alpha_{S,O,i}}$$

where the domain, θ, is defined on the K-simplex—the space of all possible categorical probability distributions in $\mathbb{R}^K$. Unlike the parameter for a categorical distribution, there is no requirement that $\boldsymbol{\alpha}_{S,O}$ sum to 1. In general, higher overall values for $\boldsymbol{\alpha}_{S,O,i}$ tend to produce flatter distributions, whereas lower values favour sparser ones. For this reason, the Dirichlet parameter is sometimes referred to as a *concentration parameter*.

A higher relative value for a given $\alpha_{S,O,i}$ means the Dirichlet is biased in favour of θ's that assign a high probability to π_i. In fact, by integrating over θ, we arrive again at the marginal probability that S assigns a given persona for O:

$$\begin{aligned} P_S(\pi_i \mid O) &= \int \mathcal{D}(\theta; \boldsymbol{\alpha}_{S,O}) \theta(\pi_i) d\theta \\ &= \frac{\alpha_{S,O}, i}{\sum \boldsymbol{\alpha}_{S,O}} \end{aligned} \tag{11.4}$$

As before, Self updates their model of Other based on the topos they evoked. This time, Self interprets τ by way of a particular persona—the persona *projected* by the social signal. We define the persona projected by τ (according to S) as the as the most

likely persona, given the topos and Self's model of Other. This is given by Bayes rule and Equation 11.4:

$$\begin{aligned}\text{Proj}(\alpha_{S,O}, \varphi, \tau) &= \arg\max_{i \leq K} P(\pi_i \mid \tau) \\ &= \arg\max_{i \leq K} P(\tau \mid \pi_i) \cdot P_S(\pi_i \mid O) \\ &= \arg\max_{i \leq K} \varphi_{\pi_i}(\tau) \cdot \frac{\alpha_{S,O,i}}{\sum \boldsymbol{\alpha}_{S,O}}\end{aligned} \tag{11.5}$$

Now let $\text{Proj}_S(\alpha_{S,O}, \varphi, \tau) = \hat{\pi}$. The projected persona is used to update S's model of O as follows:

$$\Delta_2(\alpha_{S,O,i}, \hat{\pi}) = \begin{cases} \alpha_{S,O,i} + 1 & \text{for } \pi_i = \hat{\pi} \\ \alpha_{S,O,i} & \text{otherwise} \end{cases} \tag{11.6}$$

Note that the updated model O is equal to the Bayesian posterior distribution, given that $\hat{\pi}$ was observed. This is a result of the conjugacy of the Dirichlet distribution over the categorical. For proof, let $\Delta_2(\boldsymbol{\alpha}_{S,O}, \tau) = \hat{\boldsymbol{\alpha}}_{S,O}$ in the following:

$$\begin{aligned} D(\theta, \hat{\boldsymbol{\alpha}}_{S,O}) &= \int \prod_{i=1}^{K} \theta(\pi_i)^{\hat{\alpha}_{S,O,i}} d\theta \\ &= \int \theta(\hat{\pi}) \prod_{i=1}^{K} \theta(\pi_i)^{\alpha_{S,O,i}} d\theta \\ &= P(\hat{\pi} \mid \theta) \cdot P(\theta \mid \boldsymbol{\alpha}_{S,O}) \\ &= P(\theta \mid \hat{\pi}, \boldsymbol{\alpha}_{S,O}) \end{aligned}$$

This conjugacy result means that updating the persona model is very simple—we simply add 1 to $\boldsymbol{\alpha}_{S,O}$ in the position corresponding to the projected persona (as in Equation 11.6).

In Δ_1, Self updates their model of Other considering all of the personae that Other *might have* been projecting by evoking τ—propagating uncertainty about the projected persona to the update function. In Δ_2, Self assumes that Other is using projecting the maximum likelihood (given τ) persona, $\hat{\pi}$, and updates the posterior accordingly. It would be interesting to compare Δ_2 to a second-order model that is uncertain about the projected persona. Unfortunately, the Dirichlet distribution is not conjugate over the likelihood, $P(\tau \mid \theta)$, meaning that the traditional Bayesian posterior, $P(\theta \mid \tau, \boldsymbol{\alpha}_{S,O})$, is not itself Dirichlet, but rather a mixture of Dirichlet distributions.

Nevertheless, the second-order model performs better than the first-order model in preliminary signaling games simulations. After ten exchanges, a second-order listener's model of the speaker is closer to the speaker's actual persona distribution than that of a first-order listener. Furthermore, as discussed in the following section, similar probabilistic models of the *category adjustment effect* make a similar assumption.

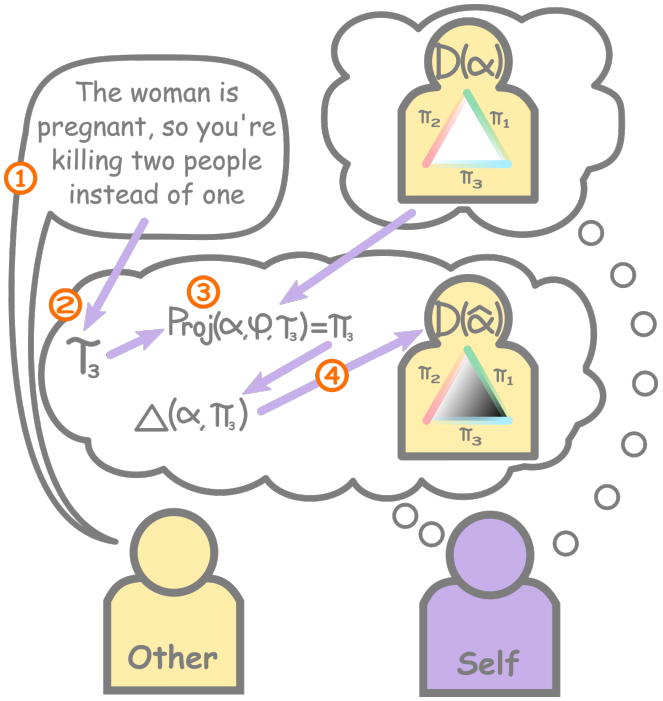


Figure 11.2.: Using the same Ψ, Π, and φ_π's as Figure 11.1, Self updates their second-order representation of Other as follows: (1) Other utters 11.1-51. (2) Self interprets 11.1-51 as evoking τ_3. (3) Self interprets τ_3 as projecting π_3, according to Equation 11.5. (4) Self updates their prior according to Equation 11.6.

11.4. The category adjustment effect

The category adjustment effect is a phenomenon in which the perception of a stimulus is biased in the direction of the centre of the category in which it falls. Category effects are, for example, an explanation for why phonetic differences are easier to detect when they cross phoneme boundaries (Feldman et al., 2009; Liberman et al., 1967).

The category adjustment model (Huttenlocher et al., 2000), describes the category adjustment effect in explicitly Bayesian terms, with the category acting as a prior distribution over stimuli. Cibelli et al. (2016), use this model to test a version of the Sapir-Whorf hypothesis. In a series of experiments they show that the semantics of colour terms have an effect on colour perception. For example, when asked to recall the colour of a displayed colour swatch, speakers were biased towards the mean of the colour category in which the swatch fell.

Eckert (2008) defines the meaning of a linguistic variable, its *indexical field*, as the "constellation of ideologically related meanings" that arises in virtue of the variable's relationship with one or more personae. Viewed through the lens of category adjustment, the social interpretation of a linguistic variable is mediated by the social categories (personae) associated with it. In our model, idealogical relatedness is represented by the conditional distribution of topoi given a persona (φ from §11.3). This

distribution corresponds to the prior from the category adjustment model. This framework suggests two empirical questions for future work.

The first question concerns the propagation of uncertainty about the projected persona. Is it the case that, as in the original category adjustment model, the interpretation of a social signal is mediated by a *single nearest category* (the projected persona from §11.3.2), or does it take into account all of the personae that the speaker *might be* projecting (as in §11.3.1)?

Second, is there a category adjustment effect on the listener's judgment of *which* topos is being evoked? Since we are focused on updates to the listener's model of the speaker's persona, we don't model uncertainty about the evoked topos, but a given argument may have multiple possible warrants. It seems reasonable to assume that the listener would take their persona model of the speaker and associations between personae and topoi into account when judging which topos was evoked.

11.5. Information state update

In order to use the above technique to account for social meaning dynamics in interaction, we integrate our model with an information state update account of dialogue, an approach successfully used to model various dialogue phenomena (Ginzburg, 2012; Larsson & Traum, 2000). We see this as pointing to a general method for incorporating previous work on social meaning, for example, Burnett (2017), into an account of incremental updates of social meaning in a idealogical context. This continues the work of Breitholtz and Cooper (2019).

To represent the evolving information states of agents involved in interaction, we use dialogue gameboards (Ginzburg, 1994, 2012; Larsson, 2002; Lewis, 1979). In order to account for coordination phenomena in dialogue, such as misunderstandings and clarifications, it is important that the information state of the participants are modelled as separate gameboards, representing each agent's view of the conversational game currently being played. The gameboards are split into two fields, one for information that the speaker takes to be private, one field for information that he or she takes to be shared in the dialogue. On our account dialogue participants are represented twice on the DGB. In Figure 11.3 we see that the shared information about the participants is just referential. The information about perceived personae of the dialogue participants can be found in the private-field of the DGB, where the labels 'other' and 'self' are associated with the corresponding individuals in the shared field. The superscripted up arrow indicates that the path points to an object three levels up in the record type.

As an interaction progresses the DGBs of the participants evolve in accordance with update rules. In Figure 11.4 we represent the update rule '$f_{\text{UpdatePersonae}}$' which is a function which takes an information state and an utterance event and returns a type for the updated information state. This function is used in the action rule 'UpdatePersonae' given in Figure 11.5. This action rule has three conditions. The first one requires that

$$\left[\begin{array}{l}\text{private:}\left[\begin{array}{l}\text{topoi:set}(\textit{Topos})\\ \text{participants:}\left[\begin{array}{l}\text{other:}\left[\begin{array}{l}\text{x=}\Uparrow^3\text{shared.participants.O:}\textit{Ind}\\ \text{pcf:}\textit{PersConcFun}(\Uparrow^2\text{topoi})\end{array}\right]\\ \text{self:}\left[\begin{array}{l}\text{x=}\Uparrow^3\text{shared.participants.S:}\textit{Ind}\\ \text{pcf:}\textit{PersConcFun}(\Uparrow^2\text{topoi})\end{array}\right]\end{array}\right]\end{array}\right]\\ \text{shared:}\left[\begin{array}{l}\text{topoi:}\left[\begin{array}{l}\text{prev:}\textit{RecType}\\ \text{curr:}\left[\begin{array}{l}\text{topos:}\textit{Topos}\\ \text{speaker:}\textit{Ind}\end{array}\right]\end{array}\right]\\ \text{participants:}\left[\begin{array}{l}\text{O:}\textit{Ind}\\ \text{S:}\textit{Ind}\end{array}\right]\end{array}\right]\end{array}\right]$$

Figure 11.3.: Representation of participants on the DGB

$$\lambda r\text{:}\left[\text{private:}\left[\begin{array}{l}\text{topoi:set}(\textit{Topos})\\ \text{participants:}\left[\text{other:}\left[\begin{array}{l}\text{x:}\textit{Ind}\\ \text{pcf:}\textit{PersConcFunc}(\Uparrow^2\text{topoi})\end{array}\right]\right]\end{array}\right]\right].$$
$$\lambda u\text{:}\left[\text{s-event:}\left[\begin{array}{l}\text{sp=}r\text{.private.other.x:}\textit{Ind}\\ \text{topos:}\textit{Topos}\\ \text{proj-pers=proj(topos, s-event.sp):}\textit{Topos}\end{array}\right]\right].$$
$$\left[\text{private:}\left[\begin{array}{l}\text{topoi=}r\text{.private.topoi:set}(\textit{Topos})\\ \text{participants:}\left[\text{other:}\left[\begin{array}{l}\text{pcf=}\Delta_2(r\text{.private.participants.other.pcf,}\\ u\text{.proj-pers):}\textit{PersConcFunc}(\Uparrow^2\text{topoi})\end{array}\right]\right]\end{array}\right]\right]$$

Figure 11.4.: $\mathrm{f}_{\text{UpdatePersonae}}$

$$\frac{s_{i,S} :_S T \qquad T \sqsubseteq \left[\text{private:}\left[\begin{array}{l}\text{topoi:set}(\textit{Topos})\\ \text{participants:}\left[\text{other:}\left[\begin{array}{l}\text{x:}\textit{Ind}\\ \text{pcf:}\textit{PersConcFunc}(\Uparrow^2\text{topoi})\end{array}\right]\right]\end{array}\right]\right] \qquad u^* :_S \left[\text{s-event:}\left[\begin{array}{l}\text{sp=}s_{i,S}\text{.private.other.x:}\textit{Ind}\\ \text{topos:}\textit{Topos}\\ \text{proj-pers=proj(topos, s-event.sp):}\textit{Topos}\end{array}\right]\right]}{s_{i+1,S} :_S T \boxed{\wedge} \mathrm{f}_{\text{UpdatePersonae}}(s_{i,S})(u^*)}$$

Figure 11.5.: UpdatePersonae: Updating personae on the DGB according to the second-order model

the agent's, S, current information state, $s_{i,S}$, is judged by S to be of some type, T. The second condition requires that T is a subtype of the type required for r in '$f_{\text{UpdatePersonae}}$'. The third condition requires that the current utterance, u^*, is of the type required for u in '$f_{\text{UpdatePersonae}}$'. If these conditions are fulfilled S is licensed or "afforded" (indicated by the wavy line) to make a judgement about S's updated information state, $s_{i+1,S}$, namely that it of the type which S judged the current information state to be of asymmetrically merged, indicated by $\boxed{\wedge}$ with the result of applying the update function to the current information state and the current utterance. The operation of asymmetric merge on record types in TTR corresponds to priority unification in feature based systems. It will preserve all the information in both types except that if the two types have different information on a given path then the information from the second type will be in the result but not that from the first type. (See Cooper and Ginzburg, 2015 and Cooper, in prep for more details.)

These definitions rely on two types which depend on the set of topoi, Ψ, which are currently under consideration. The first type is *Persona*(Ψ). A witness for this type is a distribution over Ψ. (In a more complete treatment this would just be one of a number of components that make up a persona.) That is,

> f : *Persona*(Ψ) iff f is a function with domain Ψ and range in $[0, 1]$ such that $\sum_{t \in \Psi} f(t) = 1$

The second type we use is *PersConcFunc*(Ψ), the type of Persona Concentration Functions for Ψ. This is defined as

> $(\textit{Persona}(\Psi) \rightarrow \textit{Real}_{(0,\infty+)})$

That is, *PersConcFunc*(Ψ) is the type of functions from distributions over Ψ to positive real numbers greater than 0.

11.6. Conclusion

In this paper we present a probabilistic model that accounts for the social meaning of topoi. We suggest that, as in the case of colour perception, the interpretation of social signals is subject to a category adjustment effect induced by social categories, or personae. Finally, we incorporate this model into an integrated account of linguistic interaction. We do this by defining a TTR update rule which is referenced in an action rule showing how speakers change their model of their interlocutor based on social signalling.

We see three major avenues for future work stemming from the basic model presented here. First, systems of social meaning are not monolithic or static—we should account for variation and change in the available personae, topoi, and the associations between the two. Second, this model could be used in a game-theoretic analysis of

argumentation. Based on the persona that a speaker projects, which topoi should an interlocutor use to warrant their arguments? Finally, as discussed in §11.4, how does the listener's model of the speaker persona affect which topos they interpret as warranting the speaker's argument?

Acknowledgements

This work was partly funded by Riksbankens Jubileumsfond for the Advancement of the Humanities and Social Sciences, Project P16-0805:1 *Dialogical Reasoning in Patients with Schizophrenia*.

References

Anscombre, J.-C., et al. (1995). Théorie des topoï. *Hermès*, *15*, 185–198.

Barwise, J., & Perry, J. (1983). *Situations and Attitudes*. MIT Press.

Breitholtz, E. (2014). Reasoning with topoi - towards a rhetorical approach to non-monotonicity. *Proceedings of the 50:th anniversery convention of the AISB*, 190–198
ISBN 978-1-908 87-42-0.

Breitholtz, E., & Cooper, R. (2019). Integrating personae in a TTR account of interaction.

Burnett, H. (2017). Signalling games, sociolinguistic variation and the construction of style. *Linguistics and Philosophy*, 1–32.

Cibelli, E., Xu, Y., Austerweil, J. L., Griffiths, T. L., & Regier, T. (2016). The Sapir-Whorf Hypothesis and Probabilistic Inference: Evidence from the Domain of Color. *PLOS ONE*, *11*(7), e0158725. https://doi.org/10.1371/journal.pone.0158725

Cooper, R. (in prep). *From perception to communication: An analysis of meaning and action using a theory of types with records (TTR)* [Draft of book chapters available from https://sites.google.com/site/typetheorywithrecords/drafts]. https://sites.google.com/site/typetheorywithrecords/drafts

Cooper, R., & Ginzburg, J. (2015). Type theory with records for natural language semantics. In S. Lappin & C. Fox (Eds.), *The Handbook of Contemporary Semantic Theory* (second, pp. 375–407). Wiley-Blackwell.

Ducrot, O. (1980). *Les échelles argumentatives*. Les Éditions de Minuit.

Eckert, P. (1989). *Jocks and burnouts: Social categories and identity in the high school*. Teachers college press.

Eckert, P. (2000). *Linguistic variation as social practice: The linguistic construction of identity in Belten High*. Blackwell Publishers.

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, *12*(4), 453–476. https://doi.org/10.1111/j.1467-9841.2008.00374.x

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of variation. *Annual Review of Anthropology*, *41*, 87–100.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782. https://doi.org/10.1037/a0017196

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998. https://doi.org/10.1126/science.1218633

Ginzburg, J. (1994). An update semantics for dialogue. *Proceedings of the Tilburg International Workshop on Computational Semantics*.

Ginzburg, J. (2012). *The interactive stance: Meaning for conversation*. Oxford University Press. http://www.dcs.kcl.ac.uk/staff/ginzburg/papers-new.html

Hudson, R. A. (1996). *Sociolinguistics*. Cambridge university press.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology. General*, *129*(2), 220–241.

Labov, W. (1963). The Social Motivation of a Sound Change. *WORD*, *19*(3), 273–309. https://doi.org/10.1080/00437956.1963.11659799

Larsson, S. (2002). *Issue-based Dialogue Management* (PhD Thesis). University of Gothenburg. Gothenburg, Sweden.

Larsson, S., & Traum, D. (2000). Information state and dialogue management in trindi dialogue move engine tool kit. *Natural Language Engineering*, *6*, 323–240.

Lavelle, M., Healey, P. G., & McCabe, R. (2012). Is nonverbal communication disrupted in interactions involving patients with schizophrenia? *Schizophrenia bulletin*, *39*(5), 1150–1158.

Lewis, D. (1970). General Semantics. *Synthese*, *22*(1/2), 18–67.

Lewis, D. (1979). Scorekeeping in a Language Game. In R. Bäuerle, U. Egli, & A. von Stechow (Eds.), *Semantics from Different Points of View* (pp. 172–187). Springer. https://doi.org/10.1007/978-3-642-67458-7_12

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279

12. Conditional language models for community-level linugistic variation

Bill Noble and Jean-Phililppe Bernardy

Abstract Community-level linguistic variation is a core concept in sociolinguistics. In this paper, we use conditioned neural language models to learn vector representations for 510 online communities. We use these representations to measure linguistic variation between communities and investigate the degree to which linguistic variation corresponds with social connections between communities. We find that our sociolinguistic embeddings are highly correlated with a social network-based representation that does not use any linguistic input.

12.1. Introduction

Linguistic communication requires that speakers share certain linguistic conventions, such as syntactic structure, word meanings, and patterns of interaction. Speakers assume that these conventions are *common ground* among their interlocutors, based on joint membership in a community Clark, 1996; Stalnaker, 2002. Such *speech communities* (Gumperz, 1972) range in size from the very small, like members of a friend group, to the very large, like speakers of English. However, as Eckert and McConnell-Ginet (1992) point out, it is *communities of practice* — defined by mutual social engagement in a common activity — that are the primary locus of linguistic variation.

Variation is an important object of study in sociolinguistics, and is naturally amenable to computational analysis (Nguyen et al., 2016). Most previous computational work on linguistic variation has considered variation at the level of macro-social categories, such as gender (Bamman, Eisenstein, et al., 2014; Burger et al., 2011; Ciot et al., 2013), age (Nguyen et al., 2013), and geographic location (Bamman, Dyer, et al., 2014; Eisenstein et al., 2010).In the present work, however, we investigate linguistic variation across online communities in the social media website Reddit.

For this purpose, we introduce (Section 12.2) various Community-Conditioned Language Models (CCLMs for short). These models are conditioned on a vector representation (or embedding), which varies by community. Hence, they learn *community*

embeddings. We report which architectures make best use of the community information (Section 12.3), however our our primary purpose is not to improve language models in terms of perplexity, but rather to extract community embeddings that capture linguistic similarities between communities and test how the resulting embeddings correspond to the social structure of subreddits. To that end, we test the how well the community embeddings correlate with a social network-based representation of communities (Section 12.4).

The contributions of this work are twofold. First, we develop a language model-based community embedding that we show is correlated with (but still different from) an embedding based on community membership alone. Second, the method we describe for testing the correlation between two embeddings from different models is, to our knowledge, novel to computational linguistics.

12.2. Community-conditioned language models (CCLMs)

We experiment with two kinds of model architecture: simple unidirectional LSTM (Hochreiter & Schmidhuber, 1997) and a masked Transformer (Vaswani et al., 2017). Although Transformer-based language models are considered state-of-the-art, they achieve dominance partly thanks to the availability of very large data sets (e.g., Brown et al., 2020; Devlin et al., 2019), which are not available to us.[1] Thus the LSTM is a worthy model to test for us.

In either case, the model is organised as a standard 3-layer neural sequence encoder, where the input for the tth timestep of the $n+1$st layer is the tth hidden state of the nth layer. As usual, the input to the first layer, is a sequence of tokens, encoded with a trainable embedding layer over a pre-determined vocabulary. At the other end, word tokens are predicted using a softmax projection layer. What we have described so far does not take community into account and as such we call them *unconditioned models*, but the same encoder architecture also forms the core of our conditioned models.

In the CCLMs, we add a *community embedding* parameter, which varies depending on the community of origin of the input sample. This parameter is concatenated (at each time step) with the hidden layer of the sequence encoder, at some layer $l_c \leq n$, and passed through a linear layer which projects the resulting vector back to the original hidden layer size. For $l_c = n$, the output of this linear layer is passed directly to the softmax function, just as the final hidden layer of the sequence encoder is in other models. For $l_c = 0$, the community embedding is concatenated with the token embedding. For this reason, we set the hidden size of the sequence encoder and the size of the token embedding to be equal for all models.

[1] Fine-tuning existing models is not compatible with our methodology, because we fundamentally change the structure of the network by concatenating community embeddings with hidden states at various levels.

12.2.1. Data sets

We investigate linguistic variation across various communities from the social media website Reddit.[2] Reddit is divided into forums called *subreddits*, which are typically organised around a topic of interest. Users create *posts*, which consist of a link, image, or text, along with a *comment* section. Comments are threaded: a comment can be made directly on a post, or appear as a reply to another comment. Hereafter we refer to such comments as "messages", matching our convention in mathematical formulas: the letter c stands for a community, and m stands for a message.

Our dataset includes messages from 510 subreddits, the set of all subreddits with at least 5000 messages per month for each month of the year 2015. Ignoring empty and deleted comments, we randomly sampled $42\,000$ messages from 2015 for each community. We reserved 1000 messages from each community for development and testing, leaving a total of 20.4M messages for training.

Using `langid.py` (Lui & Baldwin, 2012), we observe that a majority of the overall messages are classified as English (95% of the test set) and 498 of 510 communities have more than half of their messages classified as English. Given the small amount of non-English data, we decided that the bias introduced by attempting to filter message by language outweighted the potential benefits.[3]

Messages were preprocessed as follows: we excluded the content of block quotes, code blocks, and tables and removed markup (formatting) commands, extracting only rendered text. Messages were tokenized using the default English model for the SpaCy tokenizer Version 2.2.3 (Honnibal & Montani, 2017).

12.2.2. Training scheme

Models used a vocabulary of $40\,000$ tokens (including a special out-of-vocabulary token), consisting of the most frequent tokens across all communities.

We trained the models on a simple auto-regressive language modeling task with cross-entropy loss. Because the Transformer operates on all tokens in the sequence at once, the inputs to the model were masked and incrementally un-masked. We used the AdamW (Loshchilov & Hutter, 2017) optimisation algorithm, with an initial learning rate of 0.001 and no extra control on the decay of learning rate. The batch size was 256 and the maximum sequence length set to 64 tokens, truncating longer messages (16.8% of messages were longer than 64 tokens). During training, a dropout rate of 0.1 was applied between encoder layers and after each linear layer.

All experiments use models with 3 encoder layers, each with hidden (and token embedding) size of 256. The Transformer models had 8 attention heads per layer.[4]

[2] Comments were obtained from the archive at `https://pushshift.io/`. Baumgartner et al., 2020. Code for reproducing our dataset, as well as our pre-trained community embeddings are available at URL.

[3] See Section 12.7 for futher discussion.

[4] This number of attention heads was chosen to give the LSTM and Transformer models a comparable number of

The conditional models were given a community embedding with 16 dimensions. We experimented with every possible value for l_c, the depth of the community embedding, in a three-layer model ($l_c \in \{0, 1, 2, 3\}$).

We trained the models until the validation loss stopped decreasing for two epochs in a row, and used the weights from the epoch with the smallest validation loss for testing. Each training epoch took approximately 1.5 hours of GPU time.

12.3. CCLM Performance

In this section, we report the performance of the conditioned and un-conditioned models on the held out test set. First, we define two performance metrics: perplexity and information gain. In the following, we use M to refer to messages in the combined test set, and M_j for the partition of the test set originating from community c_j.

12.3.1. Perplexity

For a given model, let $H(m)$ be the model's cross-entropy loss, averaged over tokens in m. We define the perplexity on a set of messages, M, to be the exponential of the model's average cross-entropy loss:

$$\mathsf{Ppl}_M = e^{\text{average}_{m \in M}\, H(m)}$$

CCLM Information Gain We also consider the average information gain per token of the CCLM over its baseline un-conditioned counterpart, with the same sequence encoder architecture. For a given message, information gain is defined as the difference between the cross-entropy of the unconditioned model and the conditioned model:

$$H_{\text{LM}}(m) - H_{\text{CCLM}}(m)$$

For a set of messages, M, we consider the average information gain in exponential space (as a ratio of perplexities):

$$\mathsf{IG}_M = \frac{e^{\text{average}_{m \in M}(H_{\text{LM}}(m))}}{e^{\text{average}_{m \in M}(H_{\text{CCLM}}(m))}}$$

$$\mathsf{IG}_M = e^{\text{average}_{m \in M}(H_{\text{LM}}(m) - H_{\text{CCLM}}(m))}$$

Unsurprisingly, the conditioned models mostly have lower perplexity than their respective unconditioned baseline models, (i.e., $\mathsf{IG}_M > 1$, Table 12.1). While the absolute performance (Ppl_M) of the LSTM models is better, the best Transformer models have somewhat higher information gain than their LSTM counterparts.

parameters (22 171 203 and 21 779 523, respectively).

	l_c	test epoch	Ppl_M	IG_M
LSTM	-	12	68.74	-
	0	13	66.16	1.039
	1	7	**66.01**	**1.041**
	2	4	66.19	1.039
	3	4	66.35	1.036
Transformer	-	4	79.13	-
	0	4	**75.66**	**1.046**
	1	4	82.12	0.964
	2	7	83.53	0.947
	3	3	75.90	1.043

Table 12.1.: Performance of baseline (first row for each encoder architecture) and CCLM models. The scope of perplexity and information gain (M) is the entire test set, i.e. 5000×510 messages; 5000 for each community.

The effect of l_c, the depth of the community embedding, is also different across architectures. For the LSTM encoder, the best model concatenates the community embedding after the first encoder layer ($l_c = 1$), but all of the conditioned models perform similarly well. For the Transformer, the best model incorporates the community information first, concatenating it directly to the word vectors ($l_c = 0$). It performs similarly to the model that only integrates the community information after all all the Transformer layers ($l_c = 3$), but the two middle-layer models actually perform worse than the unconditioned model (with $\mathsf{IG}_M < 1$).

We also consider performance stratified by community; that is, Ppl_{M_j} and IG_{M_j}, where M_j is the set of messages originating from community c_j (Fig. 12.1). We observe a lot of variation in baseline perplexity across communities, with Ppl_{M_j} ranging from 3.67 to 93.58 for the best conditional LSTM model (Fig. 12.1; also see Section 12.8 for detailed community-level results). The conditioned models also perform differently across different communities — even among the best models, some communities have $\mathsf{IG}_{M_j} < 1$, meaning that the CCLM performs worse than the unconditioned baseline for messages from that community. For other communities IG_{M_j} is much higher, meaning that the CCLM performs better (Fig. 12.1).[5]

We observe that across all the models we tested, communities where conditioning has the least effect tend to be organised around more general interest topics, such as `/r/relationships` and `/r/advice`, where the subject matter is relevant to a broad range of people. Conditioning the model on community appears to have the most ben-

[5] Some of the communities with consistently high IG_{M_j} across all models are primarily non-English, but surprisingly, not the three most extreme outliers. There are `/r/counting`, `/r/friendsafari`, and `/r/Fireteams`, the later two of which are places where people coordinate to play video games together. The messages in these communites adhere to highly regular formats, which are presumably conventional to the community.

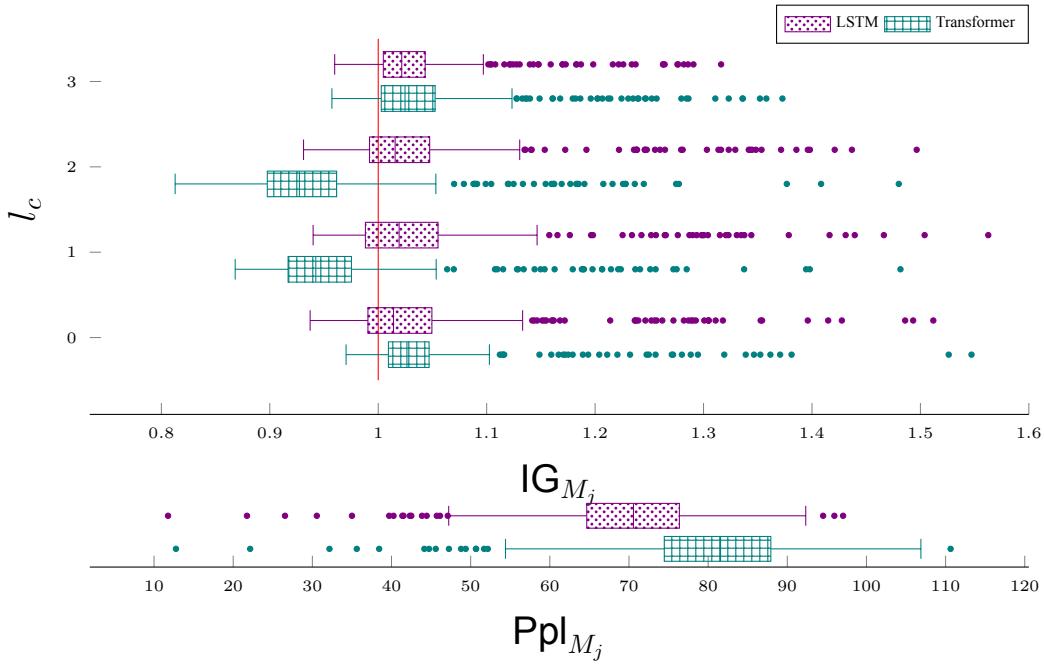


Figure 12.1.: Average model performance by community. The boxes indicate upper and lower quartiles, while the whiskers are placed at the upper and lower maximum, with communities more than $1.5 \times IQR$ (inter-quartile range) above the upper quartile considered outliers (represented as dots). The three most extreme outliers are excluded from this view.

efit for narrower special-interest subreddits, such as those organised around a certain videogame, sports team, or subculture. These empirical observations corroborate the idea that communities of practice are the primary locus of linguistic variation.

12.4. Comparison of CCLM community embeddings with a social network embedding

In this section we investigate the degree to which CCLM community embeddings correlate with the social network structure of Reddit.

To this end, we compare the CCLM-learned community embeddings[6] with the community embedding created by Kumar et al. (2018),[7] which were generated using using a negative-sampling optimization algorithm, with the author-community co-occurrence matrix as ground truth, using data from January 2014 to April 2017. We refer the reader to Kumar et al. (2018) for details, but the important point is that no linguistic infor-

[6] In this section, we only consider the embeddings from the *best* (highest information gain) CCLM from each architecture family; that is, the LSTM with $l_c = 1$ and the Transformer with $l_c = 0$, however we observed similar results for other values of l_c.

[7] Available at `https://snap.stanford.edu/data/web-RedditEmbeddings.html`

mation is used to create these embeddings: they only reflect the social relationship between communities via community membership. In contrast, CCLM community embeddings depend in no way on which user is the author of any given message: we only use the contents of messages, not authorship data.

12.4.1. Comparing embeddings: Cosine similarities

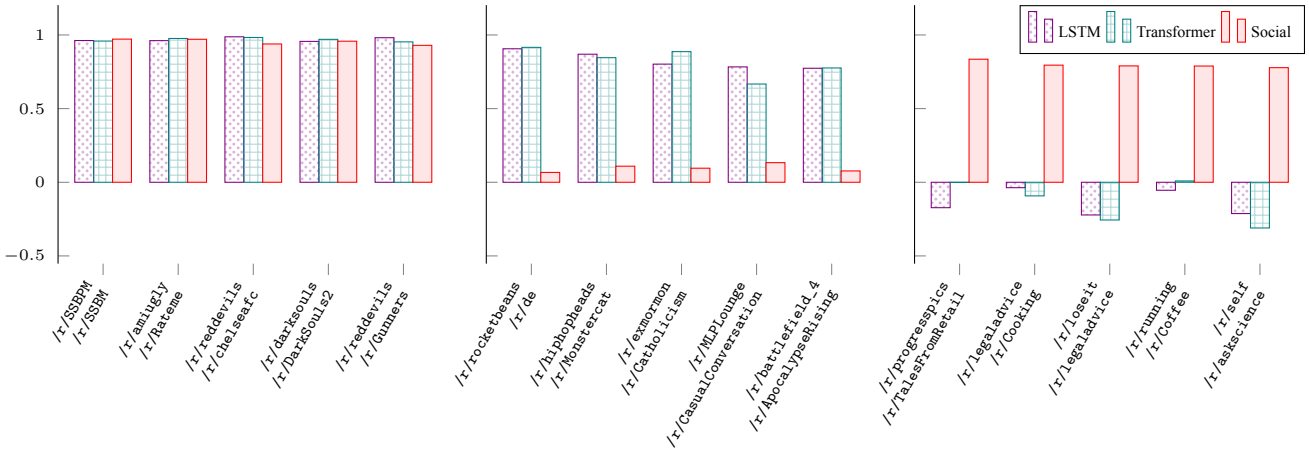


Figure 12.2.: Cosine similarity between pairs of communities, computed for vectors from the best CCLM embeddings (LSTM: $l_c = 1$, Transformer: $l_c = 0$) and the social embedding from Kumar et al. (2018). Communities with high linguistic and social similarity **(left)**, high linguistic but low social similarity **(center)**, and low linguistic but high social similarity **(right)**. See text for details on the selection criteria.

When comparing social embeddings and linguistic embeddings, a difficulty is that they range over completely unrelated spaces. Thus one cannot use the usual cosine similarity metric *between* these spaces. One can, however, use cosine similarity between *pairs* of communities, and verify that the similarities are correlated between linguistic and social embeddings. This gives a way of characterizing the differences between the two kinds of community representation. To get a more concrete sense of what this method yields, we first survey some of the most salient community pairs. We stress that this survey is not meant as a rigorous statistical analysis, as we shall see. Rather it is meant to give a flavor of discrepancies and similarities existing between linguistic and social relations.

We consider communities from three different selection criteria: Those with high linguistic *and* social similarity (where the sum of the two is highest), those with high linguistic and low social similarity (where social similarity is below the median and linguistic similarity is highest), and those with low linguistic and high social similarity (where linguistic similarity is below these median and linguistic similarity is high-

est).[89] We do not consider pairs of communities that are different in both ways, since these don't offer much in the way of understanding the respective embeddings.

Unsuprisingly, the first category (Fig. 12.2, left) yields communities that are qualitatively very similar. The /r/SSBPM and /r/darksouls communities are focused around discussion of a particular videogame, and are paired with communities that discuss a variation of the same game. The /r/amiugly and /r/Rateme communities are both forums where the posts are selfies and the comments are mostly comments on the person's appearance. The two communities paired with /r/reddevils are likewise comprised of fans of a particular English football club.

Communities with similar linguistic embeddings but dissimilar social embeddings (Fig. 12.2, left) tend to share a similar topic, mode of interaction, or language variety, but in all cases we looked at, there is some reason to expect that they might nevertheless attract different members. For example, /r/hiphopheads and /r/Monstercat are both topically related to music, but the music genres are different, and the later has a more geographically local focus (Monstercat is an independent electronic music label based in Vancouver). The interactions in both /r/MLPLounge and /r/CasualConversation could be described as casual conversation, the former is intended specifically for members of a niche internet sub-culture.

The /r/exmormon and /r/Catholicism communities discuss the Mormon and Catholic churches, although their members have different relationships towards those organizations — the former is intended for former members of the church, whereas the later is geared towards practicing Catholics. Finally, both /r/rocketbeans and /r/de are primarily German-language subreddits, but the former is comprised of fans of a computer gaming YouTube channel, while the later is more general-interest.

Differences at the other end of the spectrum (Fig. 12.2, right) are somewhat harder to interpret. It is mostly easy to see why these communities would have different linguistic embeddings — in all cases the topics are quite different. The reason they have similar social embeddings is less obvious, but we can discern some trends in how the communities are premised. The /r/progresspics and /r/TalesFromRetail are premised, in part, on seeking support from other people with similar experiences; /r/legaladvice, /r/Cooking, and /r/loseit all involve sharing knowledge on a particular topic; /r/running and /r/Coffee are hobby-focused; and /r/self (often) and /r/askscience (by premise) are places people ask and answer questions. It may be that there are different patterns in the *social function* that people attribute to this particular social media website — people who use Reddit in one way are more likely to belong to communities that are premised on the same kind of social function, even if the topics (and indeed language) of those communities are quite different. Testing this hypothesis would require a more focused study design and ideally consider communities from multiple social networks (online or otherwise).

[8] We use the LSTM ($l_c = 1$) community vectors for these purposes, but results attain with the best Transformer model.

[9] Median similarity among pairs of communities was 0.177 for the social embedding and 0.010 and 0.012 for the LSTM and Transformer linguistic embeddings, respectively.

In sum, empirical observation simultaneously reveals examples of high and low correlation between social and linguistic embeddings. To quantify correlation and extract the general trends, we must resort to statistical tools, as we do below.

A straightforward (but ultimately flawed) way to measure how similar the two spaces are would be to generalise the above method, by consider each pair of communities (i, j), and compute the correlation between the cosine similarities of both embeddings.

That is, we can compute the Pearson correlation factor of the data set:

$$C = \{(x = L_i \cdot L_j, y = S_i \cdot S_j) \text{ for } i, j \in [1, 510]\}$$

where L_i and S_i are the linguistic and social embeddings for community i. (Thus L is the matrix of (normed) linguistic embeddings and S the matrix of (normed) social embeddings.)

The analysis shows positive correlation for both the LSTM ($r = 0.438$) and Transformer ($r = 0.452$) linguistic embeddings.[10] The correlations are significant with $p < 0.001$ in all cases. However, we note that the number of pairs grows with the square of the number of communities (with 510 communities, we have 129795) pairs), meaning that standard statistical tests on Pearson correlation will assure us of statistical significance in all but the weakest of correlations. A further flaw is that the data points in C are *not* distributed independently — far from it in fact, since each data point is generated from 2 of 510 independent variables. We consider this last flaw fatal, and take a different approach for computing the correlation between community embeddings in the next section.

12.4.2. Comparing embeddings: Procrustes method

In this section, we propose a systematic approach with which we can quantify the correlation between social proximity and linguistic proximity, and measure its statistical significance.

Instead of comparing embedding pairs, as in Section 12.4.1, we will compare embeddings community by community. A naive approach would be to calculate the distance between two embeddings index-wise, which is equal to the Frobenius distance between L and S:

$$||L - S||_F = \sum_i (L_i - S_i)$$

The problem with the above metric is that even if several dimensions of L and S are correlated, they will not coincide in the *representation* of embeddings. That is, re-aligning the embeddings by applying a simple rotation (orthogonal transformation) on either matrix widely changes the $||L - S||_F$ correlation metric.

[10] By comparison, the correlation between the two linguistic embeddings is 0.759.

To make the metric independent of the representation (up to orthogonal transformations, which preserve cosine similarities), we compute the *minimum* distance between L_i and S_i, for any orthogonal matrix Ω applied to L:

$$d(L, S) = \mathsf{argmin}_\Omega ||\Omega L - S||_F$$

Here, the orthogonal matrix Ω gives a map from linguistic embeddings to social embeddings. The problem of computing $d(L, S)$ is known as the orthogonal Procrustes problem (Gower & Dijksterhuis, 2004).[11] The solution is

$$d(L, S) = n - \mathsf{Tr}(\Sigma)$$

where the matrix Σ is obtained by the singular value decomposition (SVD) $U^T \Sigma V = LS^T$. The vectors of U and V give the directions of correlation respectively of L and S. That is, each singular value σ_i (the elements of the diagonal matrix Σ), gives a measure of how much correlation there is between the directions U_i and V_i.

As is common when doing SVD, we arrange U, V and Σ such that $\sigma_i > \sigma_j$ iff $i < j$. Doing so, the largest singular value σ_0 corresponds to the principal directions of correlation (U_0, V_0), σ_1 to the second principal direction, etc.

The $d(L, S)$ metric ranges from 0 (corresponding to perfect correlation, obtained for example if $L = S$) to n (corresponding to perfect orthogonality), where $n = 510$ is the number of communities considered.

Now, to test if $d(L, S)$ corresponds to a significant correlation, it suffices to check if its value is significantly larger than the same value for random linguistic embeddings L'. The distribution of $d(L', S)$ for random embeddings is difficult to compute analytically, but we can instead evaluate it using a Monte Carlo method.

Doing so, we observed that $d(L', S)$ exhibits a mean of $\mu_d = 431.39$ and a (Bessel's-corrected) standard deviation $s_d = 2.90$ in their distance from the social embedding, S.

Thus if the real $d(L, S)$ is below the mean by several standard deviations, we can safely assume that there is statistically significant correlation between L and S. A 4-sigma difference has less than one percent chance of occurring randomly. In our case, we observe a difference of between 61 and 68 standard deviations (Table 12.2). This definitely indicates a significant correlation. Furthermore, by coming back to the definition of $d(L, S)$, we know that, on average, the cosine similarity between ΩL and S is $0.45 = (510 - 232.18/510)$. It further means that if we obtain a linguistic embedding L_k for a new community k, we can estimate its social embedding by ΩL_k, and the cosine similarity with its true social embedding S_k is expected to be $0.39 = (431.39 - 232.18)/510$ — accounting for over-fitting effects by taking the average distance rather than the maximum. In sum, it is clear that the CCLM embeddings predict some aspect the social-network embeddings — but far from all of it.

[11] This approach has also been used to compare word embeddings across representations (e.g., Hamilton et al., 2016).

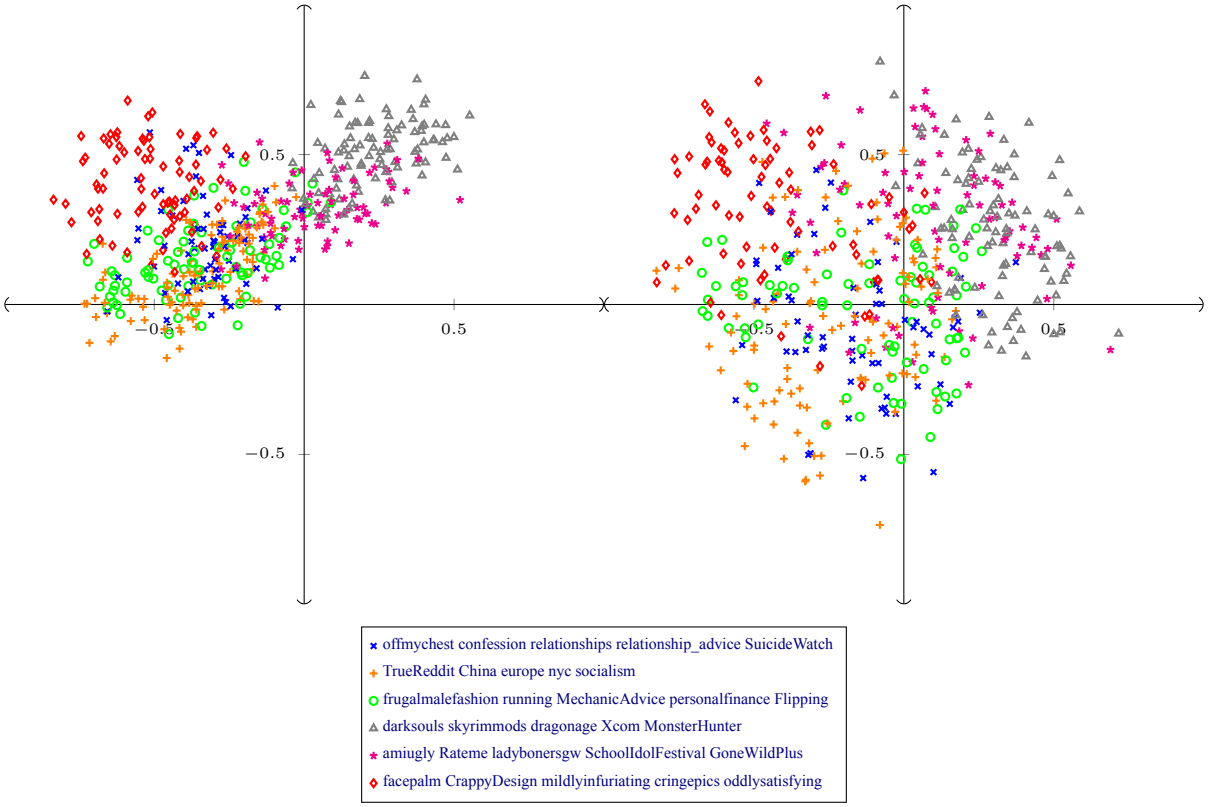


Figure 12.3.: First two components of the aligned social (left) and linguistic (right) embeddings, where the lingusitic embedding is taken from the LTSM with $l_c = 1$. Correlation between these directions is given by $\sigma_0 = 53.4$ and $\sigma_1 = 35.6$. Colors are assigned by k-means clustering of the social embedding. The legend shows the closest 5 communites to each cluster centroid. The legend shows the closest 5 communites to each cluster centroid. The cluster of each community is also available in Section 12.8

To finish, we also give a sense of *how* the correlation is manifested overall, by analysis of the two principal components of correlation in the linguistic embeddings, U_0 and U_1. To do so we plot the projection of each embedding along their first two principle components which, together with the corresponding singular values, gives an idea of how much and in what way they differ (see Fig. 12.3).

12.5. Related work

We have presented results using conditional neural language models to model variation between speech communities. The architecture of these models concatenates a vector representation of the conditioned variable to the input of the sequence model. This approach has been applied in various conditioned text generation domains such as image captioning (Vinyals et al., 2015), machine translation (Kalchbrenner & Blunsom, 2013), but it has not, to our knowledge, been used extensively to study linguistic

		LSTM	Transformer
l_c	0	254.06 (61.21)	239.41 (66.79)
	1	245.14 (64.29)	**232.18** (68.54)
	2	249.17 (62.90)	233.47 (68.32)
	3	**241.13** (65.67)	237.74 (66.84)

Table 12.2.: Distance between CCLM embeddings and the social network-based embedding of Kumar et al. (2018), as measured by $d(L, S)$. In parentheses is the number of standard deviations from the mean distance of our random embedding samples.

variation.

There are, however, related applications of conditional neural language models. Lau et al. (2017) presents a neural language model that jointly learns to predict words on the sentence-level and represent topics on the document level. The topic representation is then fed back into the language model, improving its performance on next word prediction. This is similar to how our model experiences improved performance by learning community representations. Unlike our model, topics are inferred in an unsupervised way, raising the question of whether communities could be identified from unlabeled data as well.

A piece of work with similar goals as ours is that of O'Connor et al. (2010), which uses a Bayesian generative model to infer communities from variation in text data. In contrast to our work, this model treats words as independent events, ignoring the structure (and variation) in the construction of sequences. It does further suggest, however, that community-level variation can be modeled in an unsupervised way.

Del Tredici and Fernández (2017), use a modified skip-gram model to community-level linguistic variation. They show that lexical semantic variation occurs even across different communities organised around the same topic. Their approach does not result in community level representations, however.

There are several other recent studies that aim to measure *linguistic distinctiveness* at the level of speech community (Lucy & Bamman, 2021; O'Connor et al., 2010; Zhang et al., 2017). Distinctiveness is one possible interpretation of the community-stratified information gain of the CCLM over its unconditioned counterpart (Section 12.3.1). Whereas the metrics in previous work are based on lexical frequency (and in the case of Lucy and Bamman (2021), word sense distributions), CCLM information gain is capable of capturing distinctiveness at multiple levels of linguistic analysis. However, further work is needed to investigate exactly what kinds of variation are captured.

While the focus of this paper is sociolinguistic aspects, computational models of variation can also support robust, equitable language technology. Previous work has shown that speaker demographics can improve performance on standard NLP tasks (Hovy, 2015; Yang & Eisenstein, 2017).

12.6. Discussion and Conclusion

To sum up our findings, we have defined community-conditioned language models (CCLMs). These models are generally able to attune to community-specific language, as witnessed by the information gain that they exhibit over baseline unconditioned models.

We find that the layer depth of the community embedding (l_c) has a weak effect on the information gain and the perplexity of the CCLMs.

For LSTM models, the perplexity per word, averaged over messages from all communities, was between 66.01 and 66.35 (with 68.74 for the unconditioned model). For Transformer models, it varies a bit more, between 75.66 and 83.53, but this seems to be mainly due to the poor performance of the models where the community embedding is inserted between Transformer layers ($l_c = 2$ and 3 both test above the unconditioned Transformer's average perplexity of 79.13).

The pattern of information gain by community is similar across architectures; communities that benefit most from the conditioned model behave that way for both the LSTM and Transformer. However, there are some differences. For example, many of the communities with the biggest difference in information gain between the $l_c = 0$ and $l_c = 3$ LSTMs are organised around trading collectables or organising virtual meetups (e.g., `/r/Pokemongiveaway`, `/r/ACTrade`, and `/r/SVExchange`). These communities tended to have highly conventionalized ways negotiating trades and coordinating meetups. It would be interesting to investigate these differences further in future work, since it could reveal differences in the kind of linguistic variation the different model architectures capture.

Our main result is that community representations learned by CCLMs are positively correlation with user co-occurrence patterns. Even though such *homophilic* correlation is a core hypothesis of sociolinguistics (see Kovacs and Kleinbaum (2020), for example), we believe that this study is the first to test it at the level of communities of practice using computational methods. Furthermore, it appears that our method (correlating linguistic embeddings and social embeddings) is novel. Indeed, even though the Procrustes method has been used to correlate two sets of linguistic embeddings *for the same model*, we find no evidence of the method being applied to embeddings for widely different models, as we have done.

12.7. Ethical considerations

Data privacy Our work uses publicly available data from Reddit, collected from the API made available by Baumgartner et al. (2020). Additional considerations apply, however (see Gliniecka et al. (2021) for discussion). Reddit users are not, in general, aware of the possibility that their data will be used for research purposes, and deleted posts can persist in archive formats. We do not release any data, since the it is already

publicly available and duplicating the dataset increases the likelihood that deleted posts will persist.

The paper does not include any text that could be linked back to personally identifiable information. We do release our trained community embeddings, but they have low dimensionality and pose a low risk for exposing personally identifiable information.

Language identification As mentioned in section Section 12.2.1, we decided not to filter our data for non-English comments. Although our focus in this paper is intra-language variation, language identification has the potential to introduce bias by reinforcing hegemonic language classes and the boundaries between them. In our case, filtering out messages classified as non-English would introduce bias by disproportionately removing messages in non-standard and code-switched language varieties, which are of interest in the current work.

Nevertheless, the representations learned by our model are (necessarily) relative to the other communities in the dataset. Thus the learned representations for non-English communities tend to be more similar to each other than to other communities that use mostly English, even if their predominant language is not the same. This would probably not be the case if the distribution of messages was more varied across hegemonic language classes; our work cannot be used to conclude, for example, that there is more variation within English than between Dutch and German.

Subjective analysis In the qualitative discussion offered in Section 12.4.1, our comparative characterization of the topic, mode of interaction, and language varieties used in the pairs of communities were formed by reading comments from the data our language models were trained on. This included Googling words and phrases that were unfamiliar. Where we make claims about the how the community is "premised" or what kinds of members it is "geared towards" or "intended for", these are based on the text of the sidebar on the community's Reddit page. While we believe this methodology, aggregated over many pairs of communities, is appropriate for making a qualitative comparison of the community features encoded by different representations, to make conclusions about *particular* communities based on such an analysis would be dubious and potentially harmful.

Acknowledgements

This work was supported by grant 2014-39 from the Swedish Research Council (VR) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

12.8. Appendix: Community-level results

The following table shows results at the community level. The baseline Ppl_{M_j} is computed from the unconditioned LSTM and the CCLM results (Ppl_{M_j} and IG_{M_j}, use the LSTM with $l_c = 1$). "Social cluster" is determined by k-means clustering of the social embedding.

Subredddit	baseline Ppl_{M_j}	CCLM Ppl_{M_j}	IG_{M_j}	Social embed. cluster
counting	11.77	3.67	3.21	5
friendsafari	26.55	9.76	2.72	4
Fireteams	43.85	20.01	2.19	4
randomactsofcsgo	41.37	26.47	1.56	4
RandomActsOfGaming	39.71	26.4	1.5	3
Pokemongiveaway	47.1	32.12	1.47	4
SVExchange	44.44	30.88	1.44	4
ukraina	21.74	15.19	1.43	4
ACTrade	47.82	33.77	1.42	4
gameswap	69.99	50.77	1.38	3
france	68.34	50.85	1.34	1
Rateme	60.74	45.41	1.34	0
italy	59.56	44.62	1.33	1
Romania	58.34	43.84	1.33	1
argentina	70.14	53.01	1.32	4
hardwareswap	57.3	43.39	1.32	3
EDH	76.55	58.21	1.32	3
de	71.06	54.49	1.3	1
GoneWildPlus	61.76	47.5	1.3	4
podemos	55.71	42.89	1.3	4
makeupexchange	53.03	40.85	1.3	4
rocketbeans	95.95	74.17	1.29	4
thenetherlands	69.16	53.61	1.29	1
circlejerk	53.48	41.55	1.29	5
ecigclassifieds	51.95	40.7	1.28	4
supremeclothing	85.59	67.65	1.27	4
ladybonersgw	55.83	44.16	1.26	4
pokemontrades	52.37	41.69	1.26	4
gonewildcurvy	61.23	48.92	1.25	4
gonewild	62.73	50.47	1.24	4
sweden	53.21	43.12	1.23	1
csgobetting	80.89	66.01	1.23	4
millionairemakers	42.31	35.32	1.2	5
streetwear	78.19	65.36	1.2	4
Denmark	64.2	54.56	1.18	1
ultrahardcore	64.64	55.47	1.17	4
Sneakers	72.98	63.04	1.16	4
askscience	40.25	35.11	1.15	5
ApocalypseRising	71.35	62.46	1.14	4
NHLHUT	60.6	53.07	1.14	4
gonewildaudio	59.05	51.82	1.14	4

Subredddit	baseline Ppl_{M_j}	CCLM Ppl_{M_j}	IG_{M_j}	Social embed. cluster
shittyfoodporn	68.5	60.2	1.14	5
photoshopbattles	30.56	26.92	1.14	5
amiugly	49.01	43.24	1.13	0
makinghiphop	70.47	62.22	1.13	2
MaddenUltimateTeam	72.05	63.72	1.13	4
Monstercat	70.18	62.12	1.13	4
osugame	68.02	60.33	1.13	4
weekendgunnit	67.8	60.63	1.12	4
4chan	68.54	61.32	1.12	5
fakeid	63.34	56.86	1.11	4
RealGirls	52.95	47.56	1.11	4
WritingPrompts	46.13	41.44	1.11	0
brasil	64.13	57.68	1.11	1
jailbreak	60.47	54.55	1.11	4
ClashOfClans	78.73	71.12	1.11	2
summonerschool	83.28	75.35	1.11	4
Kappa	74.12	67.21	1.1	4
food	65.9	59.79	1.1	5
summonerswar	90.06	81.71	1.1	4
magicTCG	81.65	74.11	1.1	3
Vaping	73.15	66.63	1.1	2
frugalmalefashion	68.61	62.57	1.1	2
hiphopheads	78.14	71.33	1.1	1
nsfw	42.5	38.8	1.1	4
listentothis	35	32.07	1.09	5
worldpowers	50.61	46.38	1.09	4
fivenightsatfreddys	60.84	55.88	1.09	4
longboarding	75.86	69.68	1.09	2
poker	80.61	74.09	1.09	2
progresspics	51.02	46.98	1.09	0
Cricket	89.17	82.26	1.08	1
Celebs	48.89	45.1	1.08	1
Aquariums	68.74	63.47	1.08	2
yugioh	90.35	83.49	1.08	3
bravefrontier	86.87	80.39	1.08	4
BlackPeopleTwitter	68.08	63.08	1.08	5
SSBM	84.02	77.93	1.08	3
nba	74.9	69.47	1.08	1
chelseafc	69.95	64.93	1.08	1
Shitty_Car_Mods	70.04	65.03	1.08	5
subaru	72.5	67.33	1.08	2
buildapcforme	74.75	69.45	1.08	3
tipofmytongue	56.53	52.54	1.08	5
vaporents	69.66	64.79	1.08	2
MechanicalKeyboards	65.95	61.4	1.07	3
reddevils	72.75	67.73	1.07	1
smashbros	78.81	73.38	1.07	3
MakeupAddiction	69.2	64.46	1.07	0
electronic_cigarette	66.98	62.44	1.07	2

Subredddit	baseline Ppl_{M_j}	CCLM Ppl_{M_j}	IG_{M_j}	Social embed. cluster
PuzzleAndDragons	79.77	74.57	1.07	4
Homebrewing	79.8	74.67	1.07	2
cats	59.55	55.74	1.07	5
aww	63.96	59.87	1.07	5
buildapc	66.77	62.5	1.07	3
LiverpoolFC	72.85	68.22	1.07	1
nfl	75.86	71.07	1.07	1
FIFA	65.16	61.11	1.07	4
Jokes	62.37	58.55	1.07	5
CoDCompetitive	76.44	71.77	1.07	4
Gunners	68.4	64.23	1.06	1
Civcraft	72.82	68.39	1.06	4
Coffee	70.76	66.47	1.06	2
DotA2	84.24	79.15	1.06	4
2007scape	74.44	69.96	1.06	4
Boxing	73.85	69.41	1.06	1
pathofexile	91	85.54	1.06	3
Watches	61.97	58.26	1.06	2
Smite	84.11	79.08	1.06	4
Weakpots	74.45	70.04	1.06	4
SchoolIdolFestival	77.47	72.93	1.06	4
headphones	65.24	61.43	1.06	2
steroids	78.23	73.67	1.06	4
Tinder	66.8	62.91	1.06	5
lewronggeneration	74.67	70.38	1.06	5
Cooking	82.3	77.61	1.06	2
bindingofisaac	77.83	73.4	1.06	3
leagueoflegends	76.76	72.42	1.06	4
MMA	69.95	66.04	1.06	1
Indiemakeupandmore	73.19	69.11	1.06	4
Multicopter	73.95	69.85	1.06	2
feedthebeast	81.55	77.04	1.06	3
MechanicAdvice	82.64	78.14	1.06	2
science	45.72	43.24	1.06	1
amiibo	67.55	63.9	1.06	4
rupaulsdragrace	73.74	69.88	1.06	4
SSBPM	81.88	77.59	1.06	3
eagles	69.93	66.27	1.06	1
3Dprinting	73.61	69.77	1.06	2
buildapcsales	66.81	63.34	1.05	3
hearthstone	78.36	74.29	1.05	3
coys	67.54	64.03	1.05	1
playrust	78.58	74.51	1.05	3
battlestations	63.3	60.03	1.05	3
eu4	81.78	77.56	1.05	3
GlobalOffensive	73.65	69.85	1.05	4
techsupport	69.66	66.14	1.05	3
Eve	84.15	80	1.05	3
rugbyunion	83.39	79.28	1.05	1

Subredddit	baseline Ppl_{M_j}	CCLM Ppl_{M_j}	IG_{M_j}	Social embed. cluster
Philippines	78.73	74.85	1.05	1
h1z1	77.21	73.44	1.05	3
NASCAR	70.72	67.28	1.05	1
Justrolledintotheshop	82.13	78.16	1.05	5
Animesuggest	75.84	72.22	1.05	3
Wishlist	55.21	52.58	1.05	4
zen	71.52	68.15	1.05	0
space	54.14	51.6	1.05	1
ProgrammerHumor	73.31	69.89	1.05	5
DarkSouls2	78.64	75	1.05	3
malefashionadvice	68.62	65.47	1.05	2
manga	72.97	69.65	1.05	4
neopets	67.31	64.27	1.05	4
Minecraft	71.11	67.91	1.05	3
Random_Acts_Of_Amazon	58.85	56.22	1.05	4
opieandanthony	73.76	70.47	1.05	4
knifeclub	60.88	58.21	1.05	4
ImGoingToHellForThis	56.18	53.73	1.05	5
soccer	74.33	71.16	1.04	1
Naruto	66.46	63.63	1.04	3
AsianBeauty	70.35	67.43	1.04	4
edmproduction	70.49	67.59	1.04	0
ar15	73.38	70.38	1.04	5
skyrimmods	75.03	71.98	1.04	3
trees	72.34	69.43	1.04	0
Diablo	85.14	81.71	1.04	3
churning	75.02	72.01	1.04	2
LSD	68.01	65.36	1.04	0
MLPLounge	54.02	51.92	1.04	4
windowsphone	77.28	74.32	1.04	2
bodybuilding	77.16	74.22	1.04	2
kpop	72.65	69.91	1.04	4
beer	71.65	68.97	1.04	2
DarkNetMarkets	72.35	69.67	1.04	0
tf2	76.61	73.79	1.04	3
formula1	75.47	72.7	1.04	1
MLS	75.07	72.32	1.04	1
dndnext	94.53	91.07	1.04	3
runescape	73.83	71.15	1.04	4
Pathfinder_RPG	97.05	93.58	1.04	3
golf	77.53	74.76	1.04	2
goodyearwelt	66.06	63.71	1.04	2
trashy	61.95	59.75	1.04	5
hockey	67.11	64.73	1.04	1
teenagers	67.5	65.1	1.04	4
Music	67.94	65.58	1.04	5
Unexpected	59.3	57.25	1.04	5
rawdenim	72.29	69.83	1.04	2
woodworking	73.29	70.82	1.03	2

Subredddit	baseline Ppl_{M_j}	CCLM Ppl_{M_j}	IG_{M_j}	Social embed. cluster
iphone	69	66.68	1.03	2
HomeImprovement	78.86	76.24	1.03	2
baseball	70.74	68.4	1.03	1
survivor	66.58	64.4	1.03	4
civ	87.7	84.83	1.03	3
fountainpens	66.31	64.16	1.03	2
OnePiece	70.44	68.17	1.03	3
oddlysatisfying	62.89	60.87	1.03	5
airsoft	71.09	68.82	1.03	3
nintendo	73.99	71.66	1.03	3
discgolf	76.52	74.18	1.03	5
SkincareAddiction	58.2	56.43	1.03	0
starcraft	79.73	77.33	1.03	3
gifs	63.96	62.03	1.03	5
CrusaderKings	76.55	74.28	1.03	3
polandball	76.38	74.13	1.03	4
KerbalSpaceProgram	74.81	72.6	1.03	3
justneckbeardthings	68.51	66.51	1.03	5
Xcom	92.33	89.68	1.03	3
whatisthisthing	61.1	59.35	1.03	5
Metal	75.65	73.47	1.03	4
cigars	65.73	63.86	1.03	4
pokemon	63.22	61.43	1.03	3
creepy	61.21	59.49	1.03	5
stopdrinking	55.44	53.91	1.03	0
darksouls	74.81	72.79	1.03	3
Whatcouldgowrong	65.73	63.96	1.03	5
linux	79.71	77.57	1.03	1
fireemblem	67.02	65.25	1.03	3
Guitar	73.45	71.53	1.03	0
weddingplanning	66.62	64.88	1.03	0
EarthPorn	48.93	47.65	1.03	5
Android	76.03	74.05	1.03	2
gamegrumps	64.74	63.05	1.03	3
Showerthoughts	71.29	69.45	1.03	5
woahdude	63.53	61.9	1.03	5
gameofthrones	58.92	57.41	1.03	5
anime	71.45	69.63	1.03	3
mildlyinteresting	69.17	67.41	1.03	5
Bitcoin	75.17	73.26	1.03	1
guns	68.94	67.21	1.03	5
keto	68.32	66.66	1.02	0
vinyl	69.04	67.42	1.02	2
watchpeopledie	61.89	60.44	1.02	5
SquaredCircle	77.05	75.25	1.02	4
AskHistorians	53.96	52.75	1.02	0
wiiu	69.32	67.77	1.02	3
minnesotavikings	71.49	69.89	1.02	1
fireemblemcasual	64.14	62.75	1.02	4

Subredddit	baseline Ppl_{M_j}	CCLM Ppl_{M_j}	IG_{M_j}	Social embed. cluster
dayz	69.58	68.07	1.02	3
Twitch	66.02	64.59	1.02	3
heroesofthestorm	80.23	78.53	1.02	3
math	75.52	73.93	1.02	2
Warthunder	85.85	84.04	1.02	3
whowouldwin	89.85	88	1.02	4
elderscrollsonline	80.75	79.12	1.02	3
pcmasterrace	63.58	62.29	1.02	3
interestingasfuck	63.16	61.9	1.02	5
WorldofTanks	82.8	81.15	1.02	3
wow	86.05	84.34	1.02	3
indieheads	84.46	82.78	1.02	1
opiates	69.82	68.43	1.02	0
WeAreTheMusicMakers	70.03	68.65	1.02	0
networking	81.33	79.74	1.02	2
cars	67.48	66.17	1.02	2
battlefield_4	75.42	73.96	1.02	3
dogs	67.25	65.98	1.02	2
Warframe	87.6	85.95	1.02	3
starbucks	78.55	77.12	1.02	0
skyrim	71.84	70.53	1.02	3
WTF	69.05	67.8	1.02	5
aviation	72.42	71.11	1.02	1
Warhammer40k	86.64	85.07	1.02	3
beyondthebump	69.56	68.34	1.02	4
furry	68.32	67.13	1.02	4
OldSchoolCool	58.75	57.78	1.02	5
learnprogramming	68.05	66.95	1.02	2
Flipping	70.57	69.43	1.02	2
GrandTheftAutoV	66.58	65.53	1.02	3
bodyweightfitness	72.5	71.38	1.02	2
funny	62.25	61.28	1.02	5
femalefashionadvice	66.48	65.45	1.02	0
Marvel	77.97	76.78	1.02	1
webdev	73.46	72.34	1.02	2
MMORPG	79.01	77.84	1.02	3
Surface	70.56	69.52	1.01	2
startrek	71.6	70.55	1.01	5
shittyaskscience	80.42	79.24	1.01	5
cordcutters	74.68	73.6	1.01	2
dbz	70.71	69.7	1.01	3
programming	85.82	84.6	1.01	1
3DS	68	67.06	1.01	3
cringe	57.16	56.38	1.01	5
sports	65.9	65.08	1.01	1
russia	74.41	73.52	1.01	1
Patriots	70.71	69.9	1.01	1
cringepics	53.84	53.22	1.01	5
lego	68.43	67.66	1.01	5

Subredddit	baseline Ppl_{M_j}	CCLM Ppl_{M_j}	IG_{M_j}	Social embed. cluster
photography	69.07	68.28	1.01	2
marvelstudios	77.5	76.66	1.01	1
Planetside	80.3	79.45	1.01	3
Fallout	71.83	71.08	1.01	3
NoFap	62.15	61.51	1.01	0
bicycling	72.55	71.81	1.01	2
MonsterHunter	80.89	80.07	1.01	3
ShitRedditSays	66.5	65.84	1.01	1
xboxone	69.04	68.37	1.01	3
paydaytheheist	77.23	76.49	1.01	3
DestinyTheGame	81.85	81.09	1.01	3
halo	75.8	75.12	1.01	3
Bad_Cop_No_Donut	78.04	77.35	1.01	1
asoiaf	65.9	65.32	1.01	1
apple	74.05	73.43	1.01	2
Fitness	64.81	64.27	1.01	2
comicbooks	78.94	78.29	1.01	1
thatHappened	69.04	68.48	1.01	5
india	82.2	81.55	1.01	1
paradoxplaza	75.21	74.63	1.01	3
HistoryPorn	59.95	59.49	1.01	1
investing	81.32	80.7	1.01	2
EDC	67.23	66.72	1.01	5
Steam	70.07	69.55	1.01	3
GameDeals	63.24	62.81	1.01	3
history	59.83	59.44	1.01	1
reactiongifs	54.47	54.14	1.01	5
mylittlepony	64.29	63.91	1.01	4
Guildwars2	81.19	80.72	1.01	3
loseit	59.5	59.16	1.01	0
DCcomics	73.92	73.51	1.01	1
motorcycles	74.3	73.9	1.01	2
PublicFreakout	64.42	64.08	1.01	5
legaladvice	63.01	62.68	1.01	0
bjj	74.59	74.23	1	4
horror	73.1	72.75	1	1
CrappyDesign	66.41	66.1	1	5
pics	65.78	65.55	1	5
JusticePorn	57.81	57.62	1	5
OutreachHPG	85.19	84.9	1	4
islam	69.7	69.48	1	1
Art	47.23	47.1	1	1
iamverysmart	66.58	66.41	1	5
vita	71.64	71.51	1	3
DIY	63.8	63.71	1	2
mildlyinfuriating	67.46	67.37	1	5
forwardsfromgrandma	71.97	71.87	1	5
Silverbugs	69.6	69.52	1	4
BabyBumps	67.72	67.66	1	4

Subredddit	baseline Ppl_{M_j}	CCLM Ppl_{M_j}	IG_{M_j}	Social embed. cluster
letsplay	68.08	68.02	1	3
CFB	71.18	71.17	1	1
MapPorn	77.93	77.93	1	1
DnD	87.77	87.84	1	3
syriancivilwar	77.33	77.4	1	1
vegan	68.74	68.82	1	0
Drugs	64.08	64.16	1	0
scifi	72.28	72.43	1	1
flying	77.44	77.61	1	2
nosleep	57.84	58	1	0
tifu	64.84	65.06	1	5
gadgets	63.79	64.06	1	1
PS4	67.45	67.75	1	3
army	74.93	75.26	1	2
InternetIsBeautiful	64.82	65.13	1	1
TwoBestFriendsPlay	77.9	78.3	0.99	3
Fantasy	69.35	69.73	0.99	1
ffxiv	79.74	80.22	0.99	3
EliteDangerous	82.32	82.86	0.99	3
xxfitness	70.02	70.48	0.99	0
DebateReligion	76.39	76.91	0.99	0
videos	63.11	63.56	0.99	5
gaming	67.38	67.86	0.99	3
starcitizen	82.51	83.11	0.99	3
talesfromtechsupport	74.42	74.99	0.99	5
sysadmin	80.94	81.57	0.99	2
swtor	80.56	81.2	0.99	3
RWBY	69.44	70.03	0.99	4
exmuslim	74	74.63	0.99	1
travel	61.46	61.99	0.99	2
melbourne	75.59	76.25	0.99	2
seduction	61.67	62.26	0.99	0
Christianity	72.59	73.29	0.99	1
ireland	81.86	82.64	0.99	1
sydney	79.44	80.22	0.99	2
sto	83.26	84.11	0.99	3
creepyPMs	53.32	53.87	0.99	0
harrypotter	65.42	66.1	0.99	5
london	78.96	79.78	0.99	1
CCW	69.49	70.23	0.99	5
Economics	90.45	91.44	0.99	1
cscareerquestions	69.79	70.57	0.99	2
Anarcho_Capitalism	83.27	84.21	0.99	1
metalgearsolid	73.5	74.33	0.99	3
serialpodcast	71.76	72.58	0.99	1
relationship_advice	53.94	54.56	0.99	0
movies	71.46	72.31	0.99	1
ukpolitics	79.49	80.44	0.99	1
roosterteeth	63.23	63.99	0.99	3

Subredddit	baseline Ppl_{M_j}	CCLM Ppl_{M_j}	IG_{M_j}	Social embed. cluster
socialism	76.18	77.14	0.99	1
Futurology	77.87	78.86	0.99	1
Libertarian	77.87	78.87	0.99	1
Filmmakers	66.7	67.55	0.99	2
asktrp	70.4	71.31	0.99	0
Military	73.37	74.32	0.99	1
television	69.36	70.26	0.99	1
SuicideWatch	41.51	42.05	0.99	0
boardgames	73.89	74.86	0.99	3
politics	83.72	84.83	0.99	1
rpg	87.28	88.46	0.99	3
pcgaming	73.15	74.15	0.99	3
Conservative	65.03	65.92	0.99	1
personalfinance	61.2	62.06	0.99	2
houston	75.12	76.18	0.99	2
AirForce	74.36	75.44	0.99	2
todayilearned	77.6	78.75	0.99	5
facepalm	64.59	65.56	0.99	5
GamerGhazi	76.13	77.3	0.98	1
australia	90.21	91.61	0.98	1
britishproblems	76.34	77.55	0.98	5
worldpolitics	82.57	83.91	0.98	1
running	74	75.24	0.98	2
unitedkingdom	80.59	81.94	0.98	1
oculus	81.27	82.68	0.98	3
Scotland	82.95	84.4	0.98	1
ADHD	62.9	64.03	0.98	0
Entrepreneur	65.54	66.72	0.98	2
TumblrInAction	72.39	73.72	0.98	5
GetMotivated	53.69	54.68	0.98	2
China	80.86	82.35	0.98	1
PurplePillDebate	75.96	77.38	0.98	0
rage	58.16	59.24	0.98	5
fatlogic	70.99	72.33	0.98	0
AskReddit	74.3	75.72	0.98	5
AskScienceFiction	88.54	90.24	0.98	0
StarWars	66.21	67.48	0.98	5
breakingmom	68.42	69.76	0.98	4
SubredditDrama	69.59	70.97	0.98	1
books	65.13	66.44	0.98	1
LifeProTips	64.54	65.85	0.98	5
jobs	55.59	56.72	0.98	2
philosophy	71.98	73.55	0.98	1
chicago	70.98	72.54	0.98	2
bestof	55.58	56.8	0.98	5
Anarchism	77.17	78.87	0.98	1
conspiracy	78.47	80.21	0.98	1
nottheonion	67.33	68.82	0.98	5
ottawa	70.3	71.88	0.98	1

Subredddit	baseline Ppl_{M_j}	CCLM Ppl_{M_j}	IG_{M_j}	Social embed. cluster
europe	81.43	83.28	0.98	1
worldbuilding	83.56	85.48	0.98	0
Denver	74.52	76.23	0.98	2
CasualConversation	62.58	64.02	0.98	0
worldnews	77.54	79.36	0.98	1
relationships	53.37	54.64	0.98	0
Games	65.99	67.57	0.98	3
nyc	74.19	75.98	0.98	1
Atlanta	68.99	70.68	0.98	2
news	72.41	74.2	0.98	1
Seattle	81.3	83.34	0.98	2
LosAngeles	72.9	74.73	0.98	2
casualiama	60.5	62.03	0.98	0
OkCupid	65.09	66.74	0.98	0
philadelphia	74.4	76.32	0.97	2
raisedbynarcissists	61.4	63.01	0.97	0
CanadaPolitics	81.79	83.99	0.97	1
DeadBedrooms	62.03	63.7	0.97	0
gamedev	72.03	73.98	0.97	3
KotakuInAction	78.84	80.99	0.97	1
washingtondc	74.04	76.07	0.97	2
Portland	79.22	81.43	0.97	2
Parenting	67.68	69.58	0.97	2
Documentaries	65.16	67.01	0.97	1
depression	47.83	49.21	0.97	0
technology	76.7	78.91	0.97	1
vancouver	77.54	79.82	0.97	1
UpliftingNews	58.52	60.32	0.97	1
Frugal	70.53	72.71	0.97	2
toronto	72.25	74.51	0.97	1
TalesFromRetail	61.27	63.24	0.97	0
MensRights	74.53	76.96	0.97	1
boston	75.07	77.55	0.97	2
dragonage	73.97	76.41	0.97	3
atheism	72.62	75.05	0.97	5
dataisbeautiful	67.53	69.83	0.97	1
AMA	61.41	63.51	0.97	0
CFBOffTopic	70.55	72.98	0.97	1
canada	75.02	77.62	0.97	1
truegaming	73.32	75.89	0.97	3
TrueReddit	80.07	82.89	0.97	1
TheBluePill	72.1	74.67	0.97	1
TheRedPill	74.87	77.57	0.97	0
childfree	65	67.36	0.97	0
Advice	53.12	55.05	0.97	0
actuallesbians	57.76	59.88	0.96	0
singapore	81.94	85.02	0.96	2
newzealand	79.58	82.61	0.96	1
PoliticalDiscussion	79.6	82.62	0.96	1

Subredddit	baseline Ppl_{M_j}	CCLM Ppl_{M_j}	IG_{M_j}	Social embed. cluster
rva	68.82	71.47	0.96	2
asktransgender	60.84	63.19	0.96	0
Calgary	77.45	80.48	0.96	2
Austin	78.35	81.43	0.96	2
NoStupidQuestions	66.89	69.53	0.96	0
sex	56.07	58.34	0.96	0
ProtectAndServe	68.97	71.76	0.96	2
AskMen	65.65	68.32	0.96	0
explainlikeimfive	73.28	76.26	0.96	5
TrollXChromosomes	62.66	65.22	0.96	0
AskWomen	62.65	65.26	0.96	0
ForeverAlone	55.78	58.2	0.96	0
Catholicism	75.28	78.56	0.96	1
AdviceAnimals	66.09	69.02	0.96	5
writing	70.1	73.25	0.96	0
exmormon	75.92	79.35	0.96	4
offbeat	73.4	76.73	0.96	1
masseffect	69.64	72.89	0.96	3
TwoXChromosomes	56.24	58.87	0.96	0
askgaybros	59.77	62.74	0.95	0
offmychest	51.86	54.5	0.95	0
IAmA	65.06	68.55	0.95	5
changemyview	74.27	78.29	0.95	0
self	60.25	63.57	0.95	0
confession	51.66	54.61	0.95	0
OutOfTheLoop	60.83	64.38	0.94	5
exjw	75.12	79.59	0.94	4
gaybros	71.02	75.55	0.94	0

References

Bamman, D., Dyer, C., & Smith, N. A. (2014). Distributed Representations of Geographically Situated Language. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 828–834. https://doi.org/10.3115/v1/P14-2134

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*(2), 135–160. https://doi.org/10.1111/josl.12080

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit Dataset. *arXiv:2001.08435 [cs]*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., …

Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating Gender on Twitter. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1301–1309.

Ciot, M., Sonderegger, M., & Ruths, D. (2013). Gender Inference of Twitter Users in Non-English Contexts. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1136–1145.

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Del Tredici, M., & Fernández, R. (2017). Semantic Variation in Online Communities of Practice. *IWCS 2017 - 12th International Conference on Computational Semantics - Long Papers*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Eckert, P., & McConnell-Ginet, S. (1992). Communities of practice: Where language, gender, and power all live. *Locating Power, Proceedings of the 1992 Berkeley Women and Language Conference*, 89–99.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A Latent Variable Model for Geographic Lexical Variation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1277–1287.

Gliniecka, M., Reagle, J., Proferes, N., Fiesler, C., Gilbert, S., Jones, N., Zimmer, M., Xia, H., Sehat, C. M., Prabhakar, T., & Kaminski, A. (2021). AoIR ethics panel 2: Platform challenges. *AoIR Selected Papers of Internet Research*. https://doi.org/10.5210/spir.v2021i0.12096

Gower, J. C., & Dijksterhuis, G. B. (2004). *Procrustes problems*. Oxford University Press
OCLC: ocm53156636.

Gumperz, J. (1972). The Speech Community. In P. P. Giglioli (Ed.), *Language and social context: Selected readings*. Harmondsworth : Penguin.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1501. https://doi.org/10.18653/v1/P16-1141

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Hovy, D. (2015). Demographic Factors Improve Classification Performance. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguis-*

tics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 752–762. https://doi.org/10.3115/v1/P15-1073

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent Continuous Translation Models. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 10.

Kovacs, B., & Kleinbaum, A. M. (2020). Language-Style Similarity and Social Networks. *Psychological Science*, *31*(2), 202–213. https://doi.org/10.1177/0956797619894557

Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Community Interaction and Conflict on the Web. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 933–943. https://doi.org/10.1145/3178876.3186141

Lau, J. H., Baldwin, T., & Cohn, T. (2017). Topically Driven Neural Language Model. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 355–365. https://doi.org/10.18653/v1/P17-1033

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. https://doi.org/10.48550/ARXIV.1711.05101

Lucy, L., & Bamman, D. (2021). Characterizing English Variation across Social Media Communities with BERT. *Transactions of the Association for Computational Linguistics*, *9*, 538–556. https://doi.org/10.1162/tacl_a_00383

Lui, M., & Baldwin, T. (2012). Langid.py: An Off-the-shelf Language Identification Tool. *Proceedings of the ACL 2012 System Demonstrations*, 25–30.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational Sociolinguistics: A Survey. *Computational Linguistics*, *42*(3), 537–593. https://doi.org/10.1162/COLI_a_00258

Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2013). How Old Do You Think I Am?: A Study of Language and Age in Twitter. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 10.

O'Connor, B., Eisenstein, J., Xing, E. P., & Smith, N. A. (2010). A mixture model of demographic lexical variation. *In Proceedings of NIPS Workshop on Machine Learning for Social Computing*, 6.

Stalnaker, R. (2002). Common Ground. *Linguistics and Philosophy*, *25*(5-6), 701–721.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *arXiv:1706.03762 [cs]*.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator. *arXiv:1411.4555 [cs]*.

Yang, Y., & Eisenstein, J. (2017). Overcoming Language Variation in Sentiment Analysis with Social Attention. *Transactions of the Association for Computational Linguistics*, *5*, 295–307. https://doi.org/10.1162/tacl_a_00062

Zhang, J., Hamilton, W. L., Danescu-Niculescu-Mizil, C., Jurafsky, D., & Leskovec, J. (2017). Community Identity and User Engagement in a Multi-Community

Landscape. *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 377–386.

13. Semantic shift in social networks

Bill Noble, Asad Sayeed, Raquel Fernàndez, and Staffan Larsson

Abstract Just as the meaning of words is tied to the communities in which they are used, so too is semantic change. But how does lexical semantic change manifest differently across different communities? In this work, we investigate the relationship between community structure and semantic change in 45 communities from the social media website Reddit. We use distributional methods to quantify lexical semantic change and induce a social network on communities, based on interactions between members. We explore the relationship between semantic change and the *clustering coefficient* of a community's social network graph, as well as community *size* and *stability*. While none of these factors are found to be significant on their own, we report a significant effect of their three-way interaction. We also report on significant word-level effects of frequency and change in frequency, which replicate previous findings.

13.1. Introduction

The mechanisms and patterns of semantic change have a long history of study in linguistics (e.g., Blank, 1999; Bloomfield, 1933; Paul, 1886). However, historical accounts of semantic change typically consider meaning at the language level and, as Clark (1996) points out, referring to Lewis's (1969) account of convention, the meaning of a word "does not hold for a word *simpliciter*, but for a word *in a particular community*". This gives rise questions of how semantic change manifests differently in different communities. In this work, we explore relationship between semantic change and several community characteristics, including social network structure.

Social network analysis has long been a tool of sociolinguists studying variation and change (e.g., Bloomfield, 1933; Eckert, 1988; J. Milroy & Milroy, 1985), but our work differs somewhat from that tradition in both methodology and focus. Sociolinguists typically work with the social networks *of individuals*—their *ego networks*—how many people each speaker is connected to, what kind of relationships they have and, sometimes, how people in their immediate network are connected to each other. The ego network is convenient for sociolinguists using ethnographic methods; it is usually infeasible to recreate the entire social network of a large community (Sharma &

Dodsworth, 2020). By studying online communities, we are able to define and compute several community-level structural characteristics including *size*, *stability*, and *social network clustering* (Section 13.5).

Another way that our work differs from the variationist approach is that we consider change on the level of *meaning*. With a few exceptions (e.g., Hasan, 2009), sociolinguistic research studies variation in linguistic *form* (phonology, morphology and syntax). Indeed, mainstream sociolinguists have expressed skepticism that semantics can be a proper subject of variational analysis at all (Lavandera, 1978; Weiner & Labov, 1983), since the received definition of linguistic variation concerns multiple forms expressing the same content—i.e., *different ways of saying the same thing*. With semantics at the top of the traditional linguistic hierarchy, there is no higher-order constant to which two *meanings* can refer. In this work, we instead consider *semantic shift*, which refers to changes in the meaning of a given lexical form (Newman, 2015).

For more traditional sociolinguistic variables, *social indexicality*—the association of a variant with social identities and ideology—is the main factor that mediates diffusion (Eckert, 2019). Since semantic variation can itself carry social and idealogical meaning (Hasan, 2009), there is good reason to think that it may be sensitive to some of the same aspects of community structure.

The focus on semantic shift is also made possible by computational methodology—we model word meaning with distributional semantics (Section 13.4), which allows us to quantify short-term lexical semantic shifts at the community level.

In this study, we model the social networks of 45 English-language communities from the social media website Reddit,[1] and use diachronic word vectors to measure semantic change between two time periods one year apart. Then, we use a multi-stage linear mixed effects statistical model to test the effect of various community features on word-level semantic change.

13.2. Related work

In this section, we review work that uses computational methods to study linguistic variation and change in social context.

Distributional semantics Distributional methods, which model the meaning of a word with the contexts in which it appears, are a popular way to detect and quantify semantic change.[2] Several recent studies use distributional semantics to examine short-term semantic shift at the community level. Azarbonyad et al. (2017) use diachronic word vectors to study semantic change in political and media discourse, including in UK parliamentary debates, finding that word meaning changes differently depending

[1] https://www.reddit.com

[2] See Tahmasebi et al. (2018), Tang (2018), and Kutuzov et al. (2018) for recent surveys.

on the political viewpoint of the speaker. Stewart et al. (2017) use diachronic word vectors to measure semantic change in the VKontakte social network during the Russia-Ukriaine crisis and find that changes in word frequency are predictive of semantic shift. Del Tredici et al. (2019) studied short-term semantic shift in the /r/LiverpoolFC community on Reddit, empirically validating the diachronic word vector model proposed by Kim et al. (2014) by correlating cosine distance between vectors from two different time periods with semantic change judgments collected from members of the community. In another study Del Tredici and Fernández (2017) find variations in word meaning across different Reddit communities, including communities organized around the same topic.

Social network analysis In an early example of using social network analysis to study the language online communities, Paolillo (1999) categorizes the relationships of users of an IRC channel as strong or weak ties, based on interaction frequency. They find that tie strength predicts the use of some online and community-specific forms but not others and conjecture that this difference is related the social meaning of those forms. Kooti et al. (2012) examined early Twitter conventions for attributing the source tweet to someone else (i.e., indicating that it is a *retweet*). They examined social network features, such as the size of a user's ego network, but did not find such features to be very predictive of convention adoption compared to global trends.

Communication games in a laboratory setting have also been used to examine the effect of social network structure on linguistic change. Raviv et al. (2019) quantified the communicative success, systematicity and stability of languages developed by "communities" of participants, but did not find a significant effect across the three different network structures that were tested. Lev-Ari (2018) found that individuals with larger real-world ego networks had less malleable semantic representations in the lab, and use computer simulations to argue that individuals with smaller ego networks therefore play an important role in the community-level propagation of linguistic change.

13.3. Data

To investigate semantic change in different communities, we use comments collected from the social media website Reddit.[3] On Reddit, users create *posts*, which consist of a link, image, or user-generated text, along with a *comment* section. Comments are threaded: users can comment on the post or reply to another user's comment.

Reddit is divided into forums called *subreddits*, which are typically organized around a topic of interest. While some forums—especially those organized around relatively niche topics—have a small tightly-knit community of users, others have a much looser community structure, with any given user posting and commenting infrequently.

[3] Obtained from pushshift.io Baumgartner et al., 2020.

Our dataset consists of comments from 45 randomly selected subreddits that were active in the years 2015–2017. In addition to the subreddit corpora, we created a generic Reddit corpus, consisting of comments sampled from every subreddit, including communities not in our sample. For both the generic corpus and the community-specific corpora, we constructed separate datasets for 2015 and 2017, leaving a one-year gap between them. The generic corpus consists of 55M comments for 2015 and 54M for 2017. For each of the selected subreddits, we sampled comments from 2015 and 2017 to construct two datasets of 5.4M tokens each (averaging 158K comments).[4]

13.4. Semantic change model

In this section, we describe how we quantify semantic change. We adopt a modeling procedure similar to that of Del Tredici et al. (2019), which is adapted from Kim et al. (2014)'s diachronic skip-gram with negative sampling (SGNS) model (Section 13.4.1). We define *naïve cosine change* for the community-specific and "generic" lexicons (Section 13.4.2). In Section 13.4.3, we use a control procedure adapted from Dubossarsky and Weinshall (2017) to account for noise in the naïve metric.

13.4.1. Diachronic SGNS

The strategy laid out by Kim et al. (2014) is to train a standard skip-gram language model on a corpus from some time period t_0, and then for each subsequent time period t_{n+1}, initialize a model with the same architecture with word vectors from time period t_n.[5] Del Tredici et al. (2019) adapts this procedure for a low-data setting by first training a *base* model on some large corpus, and initializing the t_0 model with vectors from that model. We follow the same framework. We train a base model, $M_{G,2015}$, on the generic 2015 corpus. Then, for each community, c, $M_{c,2015}$ is initialized with word vectors from $M_{G,2015}$ and trained on the community-specific 2015 corpus. Then, $M_{c,2017}$ is initialized from $M_{c,2015}$ and trained on the community-specific 2017 corpus. Additionally, we train a generic 2017 model, $M_{G,2017}$, which is initialized from $M_{G,2015}$ and trained on the generic 2017 corpus. See the supplementary materials for details on vocabulary and skip-gram model hyperparameters.

In the following, will write $\vec{w}_{c,t}$ for the word vector from $M_{c,t}$, corresponding word w.

[4] See Appendix 13.8.1 and 13.8.2 for details on community selection and data preprocessing. Code for downloading the data and the running experiments can be found at `https://github.com/GU-CLASP/semantic-shift-in-social-networks`.

[5] It is not clear in the original paper if the t_{n+1} model is initialized with only the word vectors from the previous time period, or if internal weights and context vectors are included as well. It seems that most subsequent implementations only carry over the word vector weights, though, which allows for more flexibility with the vocabulary. We follow this approach.

13.4.2. Naïve cosine change

We define *naïve cosine change* as the angular distance between corresponding word vectors from the two different time periods.[6]

For a community c, naïve cosine change is defined for all words in the vocabulary as follows:

$$\Delta_c^{\cos}(w) = \frac{\cos^{-1}(\text{cos sim}(\vec{w}_{c,2015}, \vec{w}_{c,2017}))}{\pi} \tag{13.1}$$

where

$$\text{cos sim}(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\|\|v_2\|} \tag{13.2}$$

Generic naïve cosine change, $\Delta_G^{\cos}$, is defined analogously.

Generally speaking, naïve cosine change has a strong track record as a semantic change metric, performing well in both human-annotated and synthetic evaluations (Hamilton et al., 2016a; Schlechtweg et al., 2020; Shoemark et al., 2019). Especially relevant to this work, Del Tredici et al. (2019) found cosine change to correlate with aggregated semantic change judgments collected from members of the /r/LiverpoolFC community on Reddit.

Model drift can distort cosine change, although this is mainly a problem with many serially-trained time periods (Shoemark et al., 2019). In a pilot study, we experimented with post-hoc aligned vector spaces and a neighborhood-based change metric (Hamilton et al., 2016b), but found minimal differences from the naïve metric.

A more serious concern for our purposes is the fact that naïve cosine change is inherently biased towards words that appear in more variable contexts. In the following section, we examine this issue more closely and define a *rectified change* metric that controls for noise. We discuss other limitations of the model in the final discussion section.

13.4.3. Rectified change score

Consider Figure 13.1 (left). Although naïve cosine change ranges *a priori* from 0 to 1, very few words score below 0.1. Even some of the most common function words have naïve cosine change above 0.2. Dubossarsky and Weinshall (2017) demonstrate that this bias is due to differences in the variance of different words' context distributions—if a word appears in highly variable contexts, the SGNS model is more likely to pick up on differences between time periods, even if those differences are mere happenstance and not reflective of actual change. This is especially a problem in our case where the amount of data is relatively small.

We adapt the *shuffle control condition* described by Dubossarsky and Weinshall (2017) to address this problem. For each subreddit, we shuffle the 2015 and 2017

[6] Some authors use $1 - \text{cos sim}$ as the cosine change metric, but angular distance is easier to interpret since it is a distance metric and ranges from 0 to 1.

corpora together and split them randomly to create *pseudo-diachronic* corpora with two "time periods". Then, we train diachronic SGNS models just as before, including initializing the "first" model with word vectors from $M_{G,2015}$. We do this $n = 10$ times for each community, giving us, for each sample i, and each vocabulary item w, a *pseudo*-naïve cosine change, $\Delta^{\cos}_{c,i}(w)$. Since no *genuine* change can possibly have taken place between the shuffled corpora, $\Delta^{\cos}_{c,i}(w)$ is a sample from the noise distribution that contributes to w's naïve cosine change, based purely on the nosiness of its context distribution in c.

Next, we take the mean, $\bar{x}_{c,w}$ and sample standard deviation (using Bessel's correction of $n-1$ degrees of freedom), $s_{c,w}$, of the samples and compute *rectified change*, which we define as the t-statistic of the genuine naïve cosine change, given the estimated noise distribution:The resulting metric, although it is still more variable for less frequent words, is unbiased by the variance of the underlying context distribution (Figure 13.1, right).

$$\Delta^*_c(w) = \frac{\Delta^{\cos}_c(w) - \bar{x}_{c,w}}{s_{c,w}\sqrt{1 + 1/n}} \tag{13.3}$$

We perform this same procedure with the generic change models (shuffling together the generic 2015 and 2017 corpora) and define *generic rectified change*, Δ^*_G, analogously.

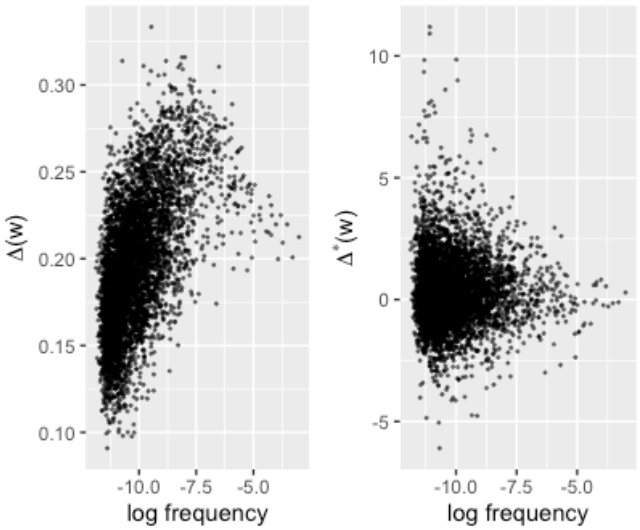


Figure 13.1.: Naïve cosine change versus rectified change for words in the /r/toronto subreddit.

Rectified change is a measure of how much higher (or lower) the measured naïve cosine change is than would be expected if the word's underlying context distribution hadn't changed at all. In other words, it quantifies the strength of the evidence that the word has changed. In our setup with 10 samples from the noise distribution, rectified change scores above 4.781 correspond to a 99.95% confidence that the change detected

	$\Delta^{\cos}$	rank	Δ^{*}	rank	freq.
possibly	0.333	1	4.19	81	7.78
;	0.316	2	0.30	2519	33.89
definitely	0.316	3	2.23	450	29.68
heck	0.314	4	2.58	311	2.19
except	0.314	5	1.60	860	14.78
2016	0.260	303	11.19	1	1.54
rentals	0.245	576	10.91	2	1.53
foreign	0.218	1414	9.84	3	4.60
admission	0.221	1330	9.83	4	1.23
screening	0.245	582	9.34	5	1.21

Table 13.1.: Top five tokens from /r/toronto, according to naïve cosine change and rectified change. Frequency is per 100k tokens.

by the diachronic SGNS model was genuine. In addition to the analytical reasons for preferring rectified change and previous empirical work on historical change, we note that the highest scoring words for each community in our data are intuitively more varied and community-specific for rectified change. The naïve cosine change frequently ranks words with some kind of rhetorical or discourse connective function as the having changed the most (see Table 13.1 for examples).

13.5. Community features

In this section we characterize the structural features of the online communities in our dataset. Many of the features we define use the notion of *active members*. For a community c and time period t, the active members, $U_{c,t}$, is the set of members who made at least 10 posts in that period.

Size The size of a community may have an effect on semantic change. In communication game experiments, Raviv et al. (2019) found that larger communities of participants developed linguistic structure faster and more consistently than when they were grouped in smaller communities.

We define community size, $S_{2015} = |U_{c,2015}|$, as the number of active members in 2015.

Stability Community stability may also have an effect on semantic change. For example, communities with stable membership have a better chance of building up community-specific common ground. On the other hand, stable communities may experience less change if such change tends to come from new community members, as some studies have suggested (Danescu-Niculescu-Mizil et al., 2013).

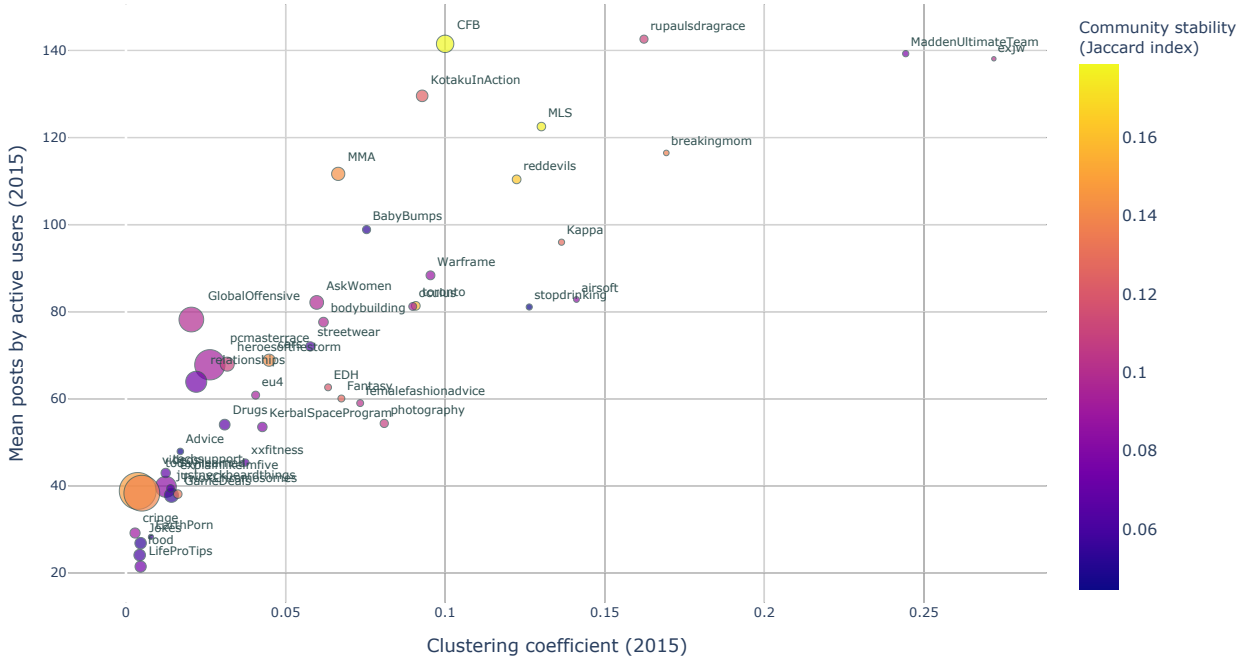


Figure 13.2.: Community-level features for each of the 45 subreddits in our experiments. Dot size represents the community's active membership in 2015 (smallest = 1 679; largest = 118 625).

We define community stability as the Jaccard index between the sets of active members in 2015 and 2017. This metric, ranging from 0 to 1, captures how similar the community membership is between the two time periods.

$$T = \frac{|U_{c,t_0} \cap U_{c,t_1}|}{|U_{c,t_0} \cup U_{c,t_1}|} \tag{13.4}$$

Mean posts P_{2015} is the average number of posts per active members over the course of 2015.

13.5.1. Social network model

In this section, we define our model of social network structure and a measure of network connectivity, which we consider along with the other community features. First, we give some background and motivation for including this feature.

Social network connectivity can have seemingly-contradictory influences on linguistic change. Bloomfield (1933) observed that densely connected networks and strong social ties have a conservative influence on an individual's speech.

It is not clear whether this pattern will hold for semantic change since, as discussed by Sharma and Dodsworth (2020), different variables respond differently to different

social network structures. We must also consider the evidence that an encounter with a novel or subtly unfamiliar word usage gives a speaker about the community's lexical *common ground* (Clark, 1996; Stalnaker, 2002). In more densely connected communities, such an exposure is better evidence that other speakers have been exposed to similar uses of the same word, either by the same speaker or, especially in the case of communities on social media, to the very same occurrence. For this reason, it could be that semantic change occurs faster in communities with dense clusters of strong social ties.

Clustering coefficient For each community, we define a graph model of its social network. For $a, b \in U_{c,2015}$, let $I(a, b)$ be the number of interactions between a and b in that community in 2015. Interactions are considered undirected (regardless of who is replying to whom) and we don't consider self-replies, meaning that $I(a, b) = I(b, a)$ and $I(a, a) = 0$. The two networks are thus defined:

$$G = \{\{a, b\} \mid I(a, b) > 1\} \tag{13.5}$$

Note that we do not consider a top-level comment to be an interaction between the commenter and the creator of the post for two reasons: First, posts frequently do not contain any text written by the author—they are often just a link or photo. Second, the author of the post is not always the addressee of top-level comments, whereas in replies to comments, the author of the parent comment is always salient (though replies may of course be made with a wider audience in mind).

The *clustering coefficient* (Watts & Strogatz, 1998), measures the graph's tendency to form dense, interconnected clusters of nodes. For an individual, i, the clustering coefficient C^i is defined as the proportion of possible connections that exist between individuals connected to i in G:

$$C^i = \frac{|\{\{j, k\} \in G \mid j, k \in N(i)|\}}{|N(i)|(|N(i)| - 1)} \tag{13.6}$$

where $N(i) = \{j \in U \mid \{i, j\} \in G\}$ is the *neighborhood* of i. The clustering coefficient for the community as a whole is the mean clustering coefficient of its members:

$$C_G = \frac{\sum_{i \in U} C^i_G}{|U|} \tag{13.7}$$

Note that C^i is precisely the measure of ego network density used in many sociolinguistic studies (L. Milroy, 1987), meaning that we would expect communities with higher clustering coefficients to exhibit less sociolinguistic change. We don't know whether the same effect holds for semantic change.

13.6. Predictive model

We perform an exploratory analysis of the data using multi-stage regressions and model selection by backwards elimination with semantic change, as measured by Δ^*, as the dependent variable.[7]

Since we fit the mixed effects model at the word level, in addition to the community-level independent variables described in Section 13.5, we consider two word-level features as fixed effects. See Table 13.2 for the full list of fixed effects.

Word frequency Since word frequency known to interact with semantic change (Hamilton et al., 2016a). we include the frequency of the token in the 2015 community corpus (f_{2015}) as a feature.

Change in frequency Additionally, we include the change in frequency between 2015 and 2017 ($f_\Delta = f_{2017} - f_{2015}$) as a feature since previous work suggests that increases in the frequency of a word often accompany semantic change (Del Tredici et al., 2019; Kulkarni et al., 2015; Wijaya & Yeniterzi, 2011).

Effect		Varies by
Mean posts (2015)	P_{2015}	community
Size (2015)	S_{2015}	community
Stability	T	community
Clustering	C	community
Frequency (2015)	f_{2015}	token, community
Change in Frequency	f_Δ	token, community
Generic rectified change	Δ^*_G	token
Rectified change	Δ^*	token, community

Table 13.2.: Fixed effect inputs to the statistical model. **Rectified change** is the dependent variable.

Community intercepts In addition to fixed effects, we use community-level random intercepts under the hypothesis that community topics have idiosyncratic reasons or lexical reasons for differences in semantic change rates to do with the community topics themselves, which we do not model.

[7] The use of stepwise regression has been criticized for being a fallacious method for one-shot hypothesis testing but is a legitimate way to investigate the explanatory capacity of predictors. See https://dynamicecology.wordpress.com/2013/10/16/in-praise-of-exploratory-statistics/ for a discussion of the issue.

13.6.1. Detecting multicollinearity

Before fitting the full model with interactions, we checked for multicollinearity via linear regressions with the standard `lm` function in R as well as the variance inflation factor (VIF) calculation provided by the `car` package in R. All the predictors were scaled and centered ($n = 201\,240$ word-community combinations). We found that the distribution of Δ^* is fat-tailed (it is likely t-distributed). Nevertheless, it is bell-shaped and large enough that this should not be a problem. We ran a regression under the hypothesis $\Delta^* \sim S_{2015} + T + C + P_{2015} + \Delta^*_G + f_{2015} + f_\Delta$ (see Table 13.2) and calculated the VIF on this model. We found that P_{2015} had VIF higher than 2, the cutoff from Zuur et al. (2010). Removing it produced VIFs below the cutoff for the other predictors.[8]

We fit a linear mixed effects model (using the `lmer` command from the `lme4` package in R; Bates et al., 2015) with the remaining predictors in order to take into account the individual semantic change characteristics of community and word. (Model code and output will be placed on the web upon publication.)

We performed a regression on the model equation $\Delta^*_c \sim (1|\text{community}) + S_{2015} * T * C + \Delta^*_G * f_{2015} * \Delta_f$; that is, we included interactions among the community-level and word-level predictors.

13.6.2. Results

For the regression results (table 13.3), we do not report statistical significance directly from `lmer`. Instead, using R's `anova` function, we performed backwards elimination model selection (by stepwise removal of interactions and factors), and we report statistical significance based on p-values derived from the χ^2 log-likelihood ratio between models.

We found that all word-level fixed effects and their three-way interaction were significant at $p < 0.05$ in the model in terms of a χ^2 likelihood ratio test. The three-way word-level interaction $\Delta^*_G \cdot f_{2015} \cdot f_\Delta$ had a p-value too small to represent ($\chi^2(4) = 6380.751$) relative to a model with all predictors without the interaction (so terms $\Delta^*_G + f_{2015} + f_\Delta$) along with all the other predictors and interactions. Relative to the model without the three-way word level interaction, removing each word-level predictor individually yielded $p_{\Delta^*_G} = 7.059 \times 10^{-81}$ ($\chi^2(1) = 362.759$), $p_{f_{2015}} = 1.605 \times 10^{-26}$ ($\chi^2(1) = 113.587$), and p_{f_Δ} was too small to measure ($\chi^2(2) = 2070.095$).

The three-way interaction for the community-level features was significant at $p = 0.014$ ($\chi^2(4) = 12.530$), but none of the two-way interactions or the individual predictors were significant.[9]

[8] Initially we defined a separate clustering metric C_{weak} for the *weak ties* network, analogous to the network defined in Section 13.5.1 but with edges between community members with exactly one interaction. However, this features was highly colinear with C and had a very high VIF when we tested it at this stage, so it was also excluded from further analysis.

[9] This means that all the individual predictors and two-way interactions must be part of the model, but their significant

We plotted the three-way interaction in Figure 13.3. Clustering coefficient and size are held at the mean and plus or minus one standard deviation from the mean. At low levels of clustering, all levels of size have a positive linear relationship on rectified change with respect to increasing stability.

Predictor	Coefficient	SE
(intercept)	0.250	0.069
S_{2015}	-0.076	0.146
T	0.041	0.046
C	-0.022	0.107
$S_{2015} \cdot T$	-0.088	0.076
$S_{2015} \cdot C$	-0.017	0.192
$T \cdot C$	-0.132	0.056
$S_{2015} \cdot T \cdot C$	-0.056	0.112
f_{2015}	-0.014	0.007
f_{Δ}	0.462	0.005
Δ^*_G	0.055	0.003
$f_{2015} \cdot f_{\Delta}$	-0.026	0.001
$f_{2015} \cdot \Delta^*_G$	-0.012	0.006
$f_{\Delta} \cdot \Delta^*_G$	0.251	0.004
$f_{2015} \cdot f_{\Delta} \cdot \Delta^*_G$	-0.014	0.000

Table 13.3.: Fixed effect coefficients of the mixed effects model with standard errors. p-values for some predictors are reported in the text.

At mean levels of clustering, the lower and mean levels of size retain the positive relationship but flatten out, and the high size level becomes negative. At one standard deviation above the mean for clustering, only the lowest size level remains positively sloped relative to stability. Confidence intervals increase dramatically as clustering increases (as there are fewer examples with higher coefficients).

The effect of the random intercept is small ($\sigma^2 = 0.019$, SD = 0.138). This is the extent to which the type of community causes the intercept of rectified change to vary.

13.7. Discussion and conclusions

We conducted an exploratory statistical analysis of the relationship between semantic change and several word- and community-level predictive features. Rectified semantic change, our independent variable, protects the results from certain systematic biases inherent in the traditional cosine change metric. By looking at online communities, we were able to compute a clustering coefficient on the social network graph of each community, as well as several other community-level structural features.

effect is conditioned on one another.

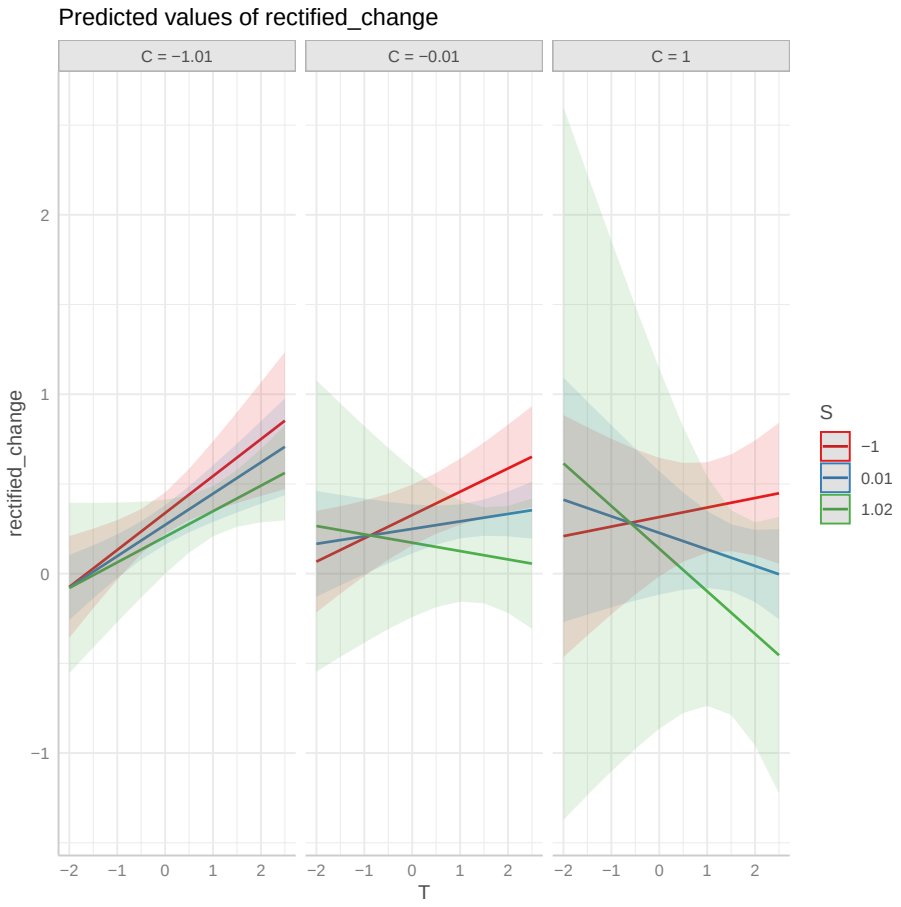


Figure 13.3.: Plot of three-way interaction between community-level predictors vs. the rectified change using the `ggeffects` package. Each panel represents a fixed value for the clustering coefficient, specifically -1 st.dev. from the mean, the mean, and 1 st.dev from the mean. Similarly, each line represents the same three values for the size. The x-axis in each panel represents the group stability.

Community features and semantic change We found all three word-level features to be significant. Together with the intercept, f_Δ dominates the mixed-effects model, with greater changes in frequency associated with higher semantic change. This is in line with previous findings (Del Tredici et al., 2019; Kulkarni et al., 2015; Wijaya & Yeniterzi, 2011), but our study is the first to demonstrate this effect while controlling for noise effects.

Although the effect is much smaller, there is a negative relationship between semantic change and baseline frequency, f. This agrees with previous results about historical change (Dubossarsky & Weinshall, 2017; Hamilton et al., 2016b), but we note that, while we cannot compare the regression coefficients directly, it appears that frequency may have a much smaller effect on semantic change in the short-term setting; however, testing this hypothesis would require further research.

Semantic change in the generic lexicon also predicts community-level change, though it has a smaller effect than f_Δ. The interaction between f_Δ and Δ_G suggests that changes in frequency can predict whether generic lexicon changes in meaning will be

picked up by a particular community.

We found that the three-way interaction between size, stability, and clustering, was significant: For communities with low clustering, there is a positive linear relationship between stability and semantic change (regardless of community size). For communities with average or high clustering, however, the positive relationship between stability and change only appears to hold for smaller communities. Note, however that the confidence intervals increase dramatically as clustering increases, since our sample of communities found fewer examples with high clustering.

We did not find significant correlations for any of the community-level features on their own. It is possible that a larger study with more communities or a more diverse set of communities would reveal some more universal effect, but we cannot make any conclusions from these results. The fact that the three-way interaction has a significant effect while none of the individual features did on their own demonstrates the complexity of relationship between structural community characteristics and semantic change.

Assumptions and limitations of the semantic change model In spite of our efforts to control for biases of cosine change, there are still some caveats when interpreting the results.

Like most distributional models of semantics, the diachronic SGNS model associates each word form with a single vector, meaning it is not sensitive to polysemy or homonymy. If a word with multiple senses undergoes changes in the relative frequency with which those senses are used, this would be reflected in the vector representation of the token that both senses are associated with, even if the meaning of either sense hasn't changed on its own.[10] However, many theories of semantic change emphasize the role of changing sense distributions as a mechanism for lexical semantic change, so it is not necessarily contrary to our aims of quantifying semantic change over the lexicon.

A related weakness of distributional semantics has to do with the distinction between meaning-in-use and lexical meaning. Even if we assume that distributional context is a faithful (if noisy) representation of the situated meaning of a word (cf. Bender & Koller, 2020; Bisk et al., 2020; Lücking et al., 2019), it might not capture the word's full *meaning potential* (Norén & Linell, 2007)—in the extreme, a word may have common ground semantic content that *could* be activated, but that happens not to appear in the corpus.

Moreover, changes in the topics discussed by the community may cause changes in the context distribution of words that don't reflect actual change in meaning. Consider the words at the top of the list for /r/toronto (Table 13.1). It's possible that some of those words appear due to changes in the socio-political topics people were discussing

[10]Contextualized word representations (Devlin et al., 2019; Peters et al., 2018) don't have this shortcoming and have recently been used to investigate semantic change (Giulianelli et al., 2020; Vani et al., 2020), but extracting one vector per occurrence is computationally expensive and has therefore only been applied to small sets of target words.

on the forum between 2015 and 2017. Similarly, the top word, 2016, presumably still refers to the same year, though the year itself went from being in the future to being in the past. Whether or not such a change counts as a *change in meaning* is naturally beyond the scope of this paper.

Future work This work offers some insight into how semantic change and community structure interact, but there are still many open questions, including how these results generalize to communities in different communicative settings and over different time frames. Future work should take a closer look at the *kinds* of change (e.g., Blank, 1999) taking place. For example, are the meanings of words broadening or narrowing? How are existing community-level communicative resources used to create new word uses? Given that we can identify statistically significant changes in meaning over a relatively short period of time, it would also be interesting to investigate the circumstances of individual changes. For example, do community members with more central social network position tend to innovate more? How are early innovative uses received by the community? Is there a correlation between semantic change in a given time period and the frequency of explicit *word meaning negotiation* (Myrendal, 2019) in the same period?

Acknowledgements

This work was supported by grant 2014-39 from the Swedish Research Council (VR) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. This work was also supported by the Marianne and Marcus Wallenberg Foundation grant 2019.0214 for the Gothenburg Research Initiative for Politically Emergent Systems (GRIPES). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455 Awarded to RF).

13.8. Appendix

13.8.1. Subreddit selection

We randomly selected 50 subreddits from the set of all forums with at least 15,000 comments per month for each of the 36 months in the 2015–2017 period. We initially selected 50 subreddits but excluded five from further analysis: two which were primarily non-English, two with particularly short average comment lengths, and one where our procedure for identifying template-generated posts failed (see Section 13.8.2 for details).

13.8.2. Data preprocessing

Below we describe the preprocessing procedure we used to prepare training data for our diachronic SGNS models.

Duplicate comments Before any text normalization steps (described below), we sought to remove duplicate template-generated posts by bots and moderating tools. Since this automated content frequently appears in only one of the two time periods, it can have an outsized effect on the cosine change score of words included in the template.

We identified these posts by comparing the tail (after the first 50 characters) any two posts of more than 50 characters in length. Posts marked as duplicate under this criteria were discarded (keeping one such post in each category). This preserves "natural" human-written duplicates, which tend to be short, while catching most template-generated content, where form-filled deviations tend to be relegated to the beginning of the post. Unfortunately, this criteria missed posts by a bot in the /r/jailbreak subreddit, resulting rectified semantic change score outliers for certain words in the bot's template. As a result, we excluded this community from analysis in the mixed-effects model.

Normalization and tokenization The text of comments was normalized as follows. We removed markdown formatting, extracting only rendered text. We exclude the content of block quotes, code blocks, and tables. We tokenized comments using the SpaCy tokenizer with the default English model (version 2.2.3). We lower-cased all tokens and removed whitespace, including linebreaks. Additionally, we removed tokens containing certain characters present in the 2015 data but absent in 2017, apparently due to text encoding changes made by Reddit. The removed characters were mostly emojis and certain Hangul graphemes and none were particularly common in our data (see [link] for a list of excluded characters).

13.8.3. Vocabulary and SGNS training proceedure

For each community c we maintain a separate vocabulary, V_c. Words with at least 50 occurances in *both* the 2015 and 2017 time periods are included in the vocabulary. Likewise, the generic Reddit models have vocabulary V_G, which includes words with at least 500 occurances in both time periods.

All models were trained with the Gensim (v. 3.8.1) SGNS implementation, with 200 dimensional vectors for 50 epochs (for both the generic and community-spcefic models). For all other hyperparameters, we maintain the default hyperparameters (length 5 context window, 5 negative samples per word, inital learning rate of 0.025, subsampling threshold of 1×10^{-5}, and negative sampling distribution exponent of 0.75).

For $M_{c,2015}$, we randomly initialize vectors for words in $V_c \setminus V_G$. Words in $V_G \setminus V_c$ have no vector representation in $M_{c,2015}$ or $M_{c,2017}$.

References

Azarbonyad, H., Dehghani, M., Beelen, K., Arkut, A., Marx, M., & Kamps, J. (2017). Words are Malleable: Computing Semantic Shifts in Political and Media Discourse. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1509–1518. https://doi.org/10.1145/3132847.3132878

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit Dataset. *arXiv:2001.08435 [cs]*.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *ACL 2020*.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., & Turian, J. (2020). Experience Grounds Language. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8718–8735.

Blank, A. C. (1999). Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. In A. Blank & P. Koch (Eds.), *Historical Semantics and Cognition*. De Gruyter Mouton.

Bloomfield, L. (1933). *Language* (14. impr). Allen & Unwin
OCLC: 256288342.

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. *Proceedings of the 22nd International Conference on World Wide Web*, 307–318. https://doi.org/10.1145/2488388.2488416

Del Tredici, M., & Fernández, R. (2017). Semantic Variation in Online Communities of Practice. *IWCS 2017 - 12th International Conference on Computational Semantics - Long Papers*.

Del Tredici, M., Fernández, R., & Boleda, G. (2019). Short-Term Meaning Shift: A Distributional Exploration. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, *1 (Long and Short Papers)*, 2069–2075.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dubossarsky, H., & Weinshall, D. (2017). Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1136–1145. https://doi.org/10.18653/v1/D17-1118

Eckert, P. (1988). Adolescent social structure and the spread of linguistic change. *Language in Society*, *17*(2), 183–207. https://doi.org/10.1017/S0047404500012756

Eckert, P. (2019). The limits of meaning: Social indexicality, variation, and the cline of interiority. *Language*, *95*(4), 751–776. https://doi.org/10.1353/lan.2019.0072

Giulianelli, M., Del Tredici, M., & Fernández, R. (2020). Analysing Lexical Semantic Change with Contextualised Word Representations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3960–3973. https://doi.org/10.18653/v1/2020.acl-main.365

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016a). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1501. https://doi.org/10.18653/v1/P16-1141

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016b). Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/D16-1229

Hasan, R. (2009). *Collected works of Ruqaiya Hasan. Vol. 2, Semantic variation: Meaning in society and in sociolinguistics*. Equinox.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., & Petrov, S. (2014). Temporal Analysis of Language through Neural Language Models. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 61–65. https://doi.org/10.3115/v1/W14-2517

Kooti, F., Mason, W. A., Gummadi, K. P., & Cha, M. (2012). Predicting emerging social conventions in online social networks. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management - CIKM '12*, 445. https://doi.org/10.1145/2396761.2396820

Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015). Statistically Significant Detection of Linguistic Change. *Proceedings of the 24th International Conference on World Wide Web - WWW '15*, 625–635. https://doi.org/10.1145/2736277.2741627

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. *Proceedings of the 27th International Conference on Computational Linguistics*, 1384–1397.

Lavandera, B. R. (1978). Where Does the Sociolinguistic Variable Stop? *Language in Society*, *7*(2), 171–182.

Lev-Ari, S. (2018). Social network size can influence linguistic malleability and the propagation of linguistic change. *Cognition*, *176*, 31–39. https://doi.org/10.1016/j.cognition.2018.03.003

Lewis, D. K. (1969). *Convention: A Philosophical Study*. Wiley-Blackwell.

Lücking, A., Cooper, R., Larsson, S., & Ginzburg, J. (2019). Distribution is not enough: Going Firther. *Proceedings of the Sixth Workshop on Natural Language and Computer Science*, 1–10. https://doi.org/10.18653/v1/W19-1101

Milroy, J., & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of Linguistics*, *21*(02), 339. https://doi.org/10.1017/S0022226700010306

Milroy, L. (1987). *Language and Social Networks*. Wiley.

Myrendal, J. (2019). Negotiating meanings online: Disagreements about word meaning in discussion forum communication - Jenny Myrendal, 2019. *Discourse Studies*, *21*(3), 317–339.

Newman, J. (2015). Semantic shift. In N. Riemer (Ed.), *The Routledge handbook of semantics*. Routledge
OCLC: 915343861.

Norén, K., & Linell, P. (2007). Meaning potentials and the interaction between lexis and contexts: An empirical substantiation. *Pragmatics*, *17*(3), 387–416. https://doi.org/10.1075/prag.17.3.03nor

Paolillo, J. (1999). The virtual speech community: Social network and language variation on IRC. *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers*, *Track2*, 10 pp.-. https://doi.org/10.1109/HICSS.1999.772680

Paul, H. (1886). *Prinzipien der Sprachgeschichte*. Max Niemeyer.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1907), 20191262. https://doi.org/10.1098/rspb.2019.1262

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1–23.

Sharma, D., & Dodsworth, R. (2020). Language Variation and Social Networks. *Annual Review of Linguistics*, *6*(1), 341–361. https://doi.org/10.1146/annurev-linguistics-011619-030524

Shoemark, P., Liza, F. F., Nguyen, D., Hale, S., & McGillivray, B. (2019). Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 66–76. https://doi.org/10.18653/v1/D19-1007

Stalnaker, R. (2002). Common Ground. *Linguistics and Philosophy*, *25*(5-6), 701–721.

Stewart, I., Arendt, D., Bell, E., & Volkova, S. (2017). Measuring, Predicting and Visualizing Short-Term Change in Word Representation and Usage in VKontakte Social Network. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, 4.

Tahmasebi, N., Borin, L., & Jatowt, A. (2018). Survey of Computational Approaches to Diachronic Conceptual Change. *arXiv:1811.06278 [cs]*, *1*(1).

Tang, X. (2018). A state-of-the-art of semantic change computation. *Natural Language Engineering*, *24*(5), 649–676. https://doi.org/10.1017/S1351324918000220

Vani, K., Mitrovic, S., Antonucci, A., & Rinaldi, F. (2020). SST-BERT at SemEval-2020 Task 1: Semantic Shift Tracing by Clustering in BERT-based Embedding Spaces. *arXiv:2010.00857 [cs]*.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Weiner, E. J., & Labov, W. (1983). Constraints on the agentless passive. *Journal of Linguistics*, *19*(1), 29–58. https://doi.org/10.1017/S0022226700007441

Wijaya, D. T., & Yeniterzi, R. (2011). Understanding semantic change of words over centuries. *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web - DETECT '11*, 35. https://doi.org/10.1145/2064448.2064475

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x