ABSTRACT

Title of dissertation: RELATING LEXICAL AND SYNTACTIC PROCESSES IN LANGUAGE: BRIDGING RESEARCH IN HUMANS AND MACHINES

Allyson Kate Ettinger, Doctor of Philosophy, 2018

Dissertation directed by: Professor Colin Phillips
Department of Linguistics

Professor Philip Resnik
Department of Linguistics, UMIACS

Potential to bridge research on language in humans and machines is substantial—as linguists and cognitive scientists apply scientific theory and methods to understand how language is processed and represented by humans, computer scientists apply computational methods to determine how to process and represent language in machines. The present work integrates approaches from each of these domains in order to tackle an issue of relevance for both: the nature of the relationship between low-level lexical processes and syntactically-driven interpretation processes. In the first part of the dissertation, this distinction between lexical and syntactic processes focuses on understanding asyntactic lexical effects in online sentence comprehension in humans, and the relationship of those effects to syntactically-driven interpretation processes. I draw on computational methods

for simulating these lexical effects and their relationship to interpretation processes. In the latter part of the dissertation, the lexical/syntactic distinction is focused on the application of semantic composition to complex lexical content, for derivation of sentence meaning. For this work I draw on methodology from cognitive neuroscience and linguistics to analyze the capacity of natural language processing systems to do vector-based sentence composition, in order to improve the capacities of models to compose and represent sentence meaning.

RELATING LEXICAL AND SYNTACTIC PROCESSES IN LANGUAGE:

BRIDGING RESEARCH IN HUMANS AND MACHINES

by

Allyson Kate Ettinger

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Colin Phillips
Professor Philip Resnik
Professor Naomi H. Feldman
Professor Ellen Lau
Professor Doug Oard

Acknowledgments

I am tremendously thankful for the many, many people who have helped and influenced me throughout the long process leading up to the writing of this dissertation. Here I will attempt to do some justice to this gratitude.

It would be difficult to overstate the amount that I owe to my advisors, Colin Phillips and Philip Resnik, for their guidance through this process. Both took me on as somewhat of an outlier among their advisees, and proceeded to oversee—with admirable patience and flexibility—my dramatic evolution into a computational linguist. They have served as an invaluable source of insight and advice, whether edifying me on their own areas of research expertise, pushing me to consider alternative perspectives in areas outside their expertise, or lending their considerable experience and wisdom in helping me to navigate the myriad other components of my professional development. I am extremely grateful for all that they have contributed to making me a better researcher and academic.

Tal Linzen has served as a mentor and collaborator since our overlapping time at NYU, when he joined my project in the MEG lab and proceeded to provide a constant source of useful information and insight. Subsequently, as a rare near-neighbor of mine at a very particular interface of research domains, Tal has continued to provide key discussion,

guidance, and collaboration as I've worked to negotiate that interface. He has played a significant role in my development as a researcher, and I'm very grateful for the time and support that he has lent over the years.

Naomi Feldman has been a critical component of my graduate career, providing my first education in computational psycholinguistics, advising my first modeling project, supervising my first interdisciplinary teaching experience, and in general serving as a tireless advocate and role model for women in computation, all of which has had substantial impact on my development and outlook. I am very thankful for everything that her influence has brought to my education and development.

I could not have done this research without the people who saw me through my transition to doing work in NLP. I'm grateful to Hal Daumé III, who provided my first education in computational linguistics and got me quickly hooked, and who subsequently served as a collaborator on multiple projects. I am very grateful also to Marine Carpuat, who agreed to advise me on my first independent NLP project, and who, in so adeptly guiding me through that project, helped me to open up what would become a key component of my research program.

Also in this category are my amazing graduate student collaborators in computer science. Sudha Rao has been a wonderful collaborator and friend since the fateful Language Science Day when we struck up that conversation about anaphora in Mandarin. And Ahmed Elgohary has been my central collaborator in one of the most significant projects of my graduate career, persevering with me through the difficulties of making that project work, providing a key complement to my own expertise, and serving as an invaluable source of insight and critical feedback. I am so grateful for all that these two contributed

to my time at Maryland. Thank you also to Yogarshi Vyas, Mohit Iyyer, Anupam Guha, and other members of the CLIP lab, who kindly provided assistance, information, support and camaraderie as I have proceeded through my education in NLP.

Alexander Williams, Valentine Hacquard, and Ellen Lau have also contributed instrumentally to the process that has led to this point, providing me with valuable guidance, feedback, and education from the start of my time at Maryland. I'm very grateful for their help and influence.

Sophia Malamud, my undergraduate advisor, is genuinely one of the primary reasons that I pursued this career path. Sophia provided my first exposure to linguistics, pushed me to produce quality work as an undergraduate, and continued to serve as a mentor and collaborator in the subsequent years. I am extremely grateful for her impact and mentorship.

Alec Marantz took me on as a research assistant at NYU, and in doing so afforded me some of the most important experience that would shape my research path. I'm very grateful for the opportunity and guidance that he provided. And to all of the other folks who impacted my development and experience at NYU, including but not limited to Chris Barker, Doug Bemis, Christian Brodbeck, Teon Brooks, Dylan Bumford, Paul Del Prato, Keith Doelling, Rebecca Egbert, Joe Fruchter, Itamar Kastner, Miriam Lauter, Kim Leiken, Gwyneth Lewis, Jeff Walker, and Masha Westerlund. Special thanks go to Chris Barker for providing unofficial mentorship on my work in semantics—and also to Joe Fruchter, for showing me that for-loop, in the watershed moment that introduced me to programming.

I want also to thank Tess Wood, Caitlin Eaves, and Shevaun Lewis at the Maryland Language Science Center. They have been there to help me with so many things—especially Tess during my two years of organizing Winter Storm—and I'm very grateful for all of their support. Speaking of Winter Storm, thanks go also to Bob Slevc, our fearless conscripted Winter Storm faculty advisor, who made the whole process more enjoyable. Thank you also to Kim Kwok, our amazing Linguistics administrator who has made all kinds of complicated things to go smoothly.

Thank you to all of the graduate students, faculty, and staff in Linguistics and the Language Science community as a whole, for creating a wonderful environment within which to work and develop. Thank you to Jeff Green, for being a great officemate and friend, and to the rest of my awesome cohort. And I want to thank my Hyattsville gang, who have provided very much support and made this process infinitely more fun: Christian Brodbeck, Suddhasattwa Das, Lara Ehrenhofer, Phoebe Gaston, Kasia Hitczenko, Ilia Kurenkov, Natalia Lapinskaya, Paulina Lyskawa, Anton Malko, Amit Nag, and Miloš Nikolić.

As for Carl, and my family—Mom, Dad, Kyle, Chelsea, Kerry, Grandmom, Aunt Valerie—I cannot hope to do justice to all that they have done for me, throughout all of the processes that have led to this point. I will simply say that this work is above all a testament to the incredible support that they have provided, and to the tremendous impact that they have had on the person that I am today.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As scientists work to understand how it is that humans process and represent language, engineers work to figure out how to develop computational systems that process language effectively. These are importantly different enterprises, but they have key potential for overlap. In this dissertation I explore aspects of an issue that underlies both efforts: understanding (in the case of humans) and optimizing (in the case of machines) the relationship between lexical-level processes, and syntactically-guided meaning interpretation processes.

In this dissertation I will tackle the problem of lexical versus syntactic processes in a way that bridges approaches, perspectives, and needs from research on language in both humans and machines. Specifically, I will be using computational tools to ask questions about language processing in humans, and I will be using insights from cognitive neuroscience and linguistics to aid in the enterprise of improving language processing in artificial intelligence.

1.1 "Lexical" versus "syntactic" processes in humans

Throughout this dissertation I will be drawing and exploring a distinction between "lexical" processes and a contrasting class of processes that I will refer to as "syntactic", or "message-based". As I will discuss in this section, the purpose here is to draw a general distinction between 1) processes and observed effects that involve (or appear to involve) lexical content but not syntactically-driven constraints on interpretation, and 2) processes and observed effects that *do* appear to involve syntactically-driven constraints on interpretation.

The most obvious existing distinction with which this could align—and indeed the distinction that is sure to be evoked by my use of the terms "lexical" and "syntactic", is the basic distinction in linguistic and psycholinguistic theory between the lexicon—which contains information about words and their meanings—and the syntactic processes that guide the composition of that lexical content. Theories of human language processing typically assume the existence of such a lexicon, which needs to be consulted in a process of "lexical access/retrieval" in order to retrieve word meanings, and these processes are naturally considered to be separate from the syntactically-driven processes that are executed to form composed phrase and sentence meaning representations based on those accessed meanings. This construal of our "lexical/syntactic" distinction is absolutely a key component of what we aim to discuss here.

However, an additional and core driving motivation for our drawing of this distinction, and indeed one of the main motivators of the questions that we will be exploring in Chapters 2-5 is the following fact: when we measure human language processing in real time,

in spite of the fact that we have ample reason to believe that humans have rapid access to a rich set of information (including syntactic information) for generating interpretations and expectations—under certain circumstances, we see patterns of response sensitivity that are consistent with mechanisms that have access to individual words of the context, but not to the syntactic structure that determines how these words should be combined. These phenomena, which will be reviewed in greater detail in this chapter, involve apparent facilitation of words that are semantically incongruent continuations of their preceding context, but that have associative or feature-based relations with individual words of the preceding context, suggesting a mechanism of lexically-based facilitation.

These effects, which have driven much theorizing and debate in psycholinguistics, suggest a potentially different version of the above distinction: a distinction between certain existing processes that are somehow lexically-driven but asyntactic (or less syntactic), and other processes that have access to the full set of available syntactic information. Importantly, when we draw this version of the distinction, based on evidence of this kind, it is not syntax alone that comprises the latter type of process—but rather a combination of syntax, world knowledge, and lexical content, to form a fully compositional and knowledge-informed message interpretation. For this reason, I will at times refer to this category of process or mechanism as "message-based", to avoid confusion with processes that are purely syntactic. However, it is important to keep in mind that even when we draw the distinction in this way, syntactic information remains the key component that distinguishes the two types of processes—as we will discuss below, world knowledge could well be involved in the "lexical" processes, so the critical characteristic that anchors the distinction is the apparent lack (or reduction) of syntactic influence in the observed lex-

ical processes, and the apparent presence of syntactic influence in the "message-based" processes.

A key question that we will need to consider is the extent to which these two versions of the distinction are aligned. In particular, to what extent are the two construals of "lexical" processes—processes involved in access of word meanings for composition, and processes that give rise to lexical facilitation—related or identical? In this section I will lay out evidence motivating the second version of the distinction, and further consider the implications. In Chapter 2, I will delve deeper into relevant issues of lexical representation, combinatorial mechanisms, and facilitation.

1.1.1 Evidence of asyntactic lexical effects in behavioral and ERP measures

A predominant paradigm in psycholinguistic research is to study the processing of words in context, and based on the processing of a given target word, to draw inferences about the preceding mechanisms and representations that generated the state of readiness or expectation that influenced the processing of that word. A variety of measures are used within this paradigm, and although we must necessarily have different linking hypotheses between the mechanisms of interest and the generation of these measures, the basic principle of measures reflecting some kind of facilitation underlies many (if not most) of them.

In the following, we will primarily focus on measures taken from event-related potentials (ERPs)—electrical signals produced by brain activity in response to a stimulus—and

in particular on the N400. The N400 is an ERP component which since its discovery in 1980 has been a frequently-used source of evidence in language processing (Kutas & Federmeier, 2011). The N400 is a negative deflection in the ERP signal, occurring in response to all content words and peaking around 400 milliseconds after word onset. The mechanisms underlying the N400 are by no means fully understood, but this measure is similar to other psycholinguistic measures in seemingly reflecting facilitation in word processing—the greater the facilitation, the smaller the N400 amplitude. This facilitation has been associated with greater congruence or predictability, and along this line, a widespread generalization is that the amplitude of the N400 correlates with cloze probability (Kutas & Hillyard, 1984; Kutas, Lindamood, & Hillyard, 1984), a measure based on the proportion of people who choose a word in a given context during an untimed fill-in-the-blank task. More probable words, as evidenced by cloze, tend to correspond to lower N400 amplitudes. This correlation suggests that facilitation in the N400 reflects something like the extent to which an incoming word matches expectations set up by the context.

At times, however, N400 amplitude deviates from the patterning of cloze probability, suggesting that this neural response does not always reflect facilitation based on all of the information sources that are available for untimed word selections.

A classic example of such a pattern of results is that of Fischler, Childers, Achariyapaopan, and Perry (1985) who find the N400 to be insensitive to the effects of negation. Specifically, in a comparison of true sentences like "A robin is not a tree" with false sentences like "A robin is not a bird", they find that the N400 amplitude fails to reflect the low probability of the false completions (e.g., *bird*), instead showing facilitation for

those completions relative to their true counterparts (e.g., *tree*). It is important to note that by contrast, Nieuwland and Kuperberg (2008) find that with additional contextual support, the N400 reflects facilitation of the true negated sentences rather than the false sentences, suggesting that the N400 *can* reflect these more sophisticated sources of information. However, the insensitivity observed by Fischler et al. suggests the presence of less sophisticated mechanisms that stand to be accounted for.

Another classic example of this apparent insensitivity—one which will be our focus in Chapter 4—is that observed in the case of semantic role reversals. In a literature that will be reviewed in greater detail in Chapter 4, it has been found repeatedly that the N400 fails to reflect the anomaly created by reversing the typical roles of the arguments for a verb. For instance, Chow, Smith, Lau, and Phillips (2015) find no difference in N400 amplitude to target verbs in normal sentences such as "The restaurant owner forgot which customer the waitress had served", as compared to target verbs in anomalous reversed sentences such as "The restaurant owner forgot which waitress the customer had served". In these cases, the N400 thus shows insensitivity to syntactic information indicating the identities of the agent and patient (the performer and the recipient) of the target verb. Again, importantly, recent results have indicated that under some circumstances the N400 *does* show sensitivity to this syntactically-driven role information—these results will be discussed in detail in Chapter 5—suggesting a complex interplay of factors giving rise to this response. But once again, the existence of the observed insensitivities suggests a stage of processing, or a separate mechanism, with a more limited set of information.

In both of the above cases, a ready explanation is one in terms of the relation of the target word to individual words in the context. In the case of Fischler et al. (1985), we

can explain the facilitation in "A robin is not a bird" based on the close relation between *robin* and *bird* (quite plausibly stronger than the relation between *robin* and *tree*). In the case of Chow et al. (2015) and other role reversal results, we can explain the facilitation of *served* in terms of the relation between that word and preceding context words such as *waitress* and/or *customer*.

The above explanation accords well with our understanding of the behavior of the N400, which has long been acknowledged to be sensitive to facilitation by semantic relatedness between words, both within sentence contexts (Kutas & Hillyard, 1980, 1984) and within single-word contexts (Bentin, McCarthy, & Wood, 1985; C. Brown & Hagoort, 1993; Holcomb, 1988; Kutas & Hillyard, 1989). Facilitation effects on the N400 seem to be elicited both via associative relations, such as *car-wheel* (C. Brown & Hagoort, 1993; Kutas & Hillyard, 1989) and by similarity relations, such as *car-truck* (Deacon, Hewitt, Yang, & Nagata, 2000; Federmeier & Kutas, 1999). The relevance of this range of effects, and how it influences our assessment of these lexical processes, will be discussed in greater detail below.

The apparent separation of these lexical facilitation effects—by contrast to more sophisticated expectations reflected in cloze probabilities, which seem to track the full message of the sentence—has long motivated investigations into the relative contributions of lexically-driven and syntactically- (or "message"-) driven effects on the N400. Kutas (1993) isolates lexical-level contributions by examining ERPs to target words preceded only by single-word contexts—where the target words vary in association to the preceding word—and comparing these single-word association effects to the effects of cloze probability of target words within sentential contexts. The results show the qual-

itative properties of the N400 contrast to be similar for single-word association effects and sentence-level cloze effects (though the effect between conditions is much larger for the sentential contexts—explicable by the fact that sentence contexts are more constraining than the single word contexts), and Kutas concludes that the processes underlying lexically-based and sentence-based effects are similar.

Van Petten (1993) investigates the interaction of "lexical-associative" and "sentence-level" effects by crossing sentence congruence with an associative relation between a single context word and the subsequent target word. Examples are shown in Table 1.1. Consistent with the prediction that both sentence-level and lexical-associative facilitation will occur, the results show facilitation for the second word of the pair for all conditions except the "anomalous unassociated" condition, which contains neither an association nor a sentence-level expectation for that word.[1] By contrast to the prediction that sentence-level effects might have a later onset if they result from slower and more strategic mechanisms, the results show no sign of latency differences between the lexical-only (anomalous associated) and the sentence-only (congruent unassociated) effects.[2]

Results of interest come also from studies using behavioral measures such as reaction time and word naming latency. Duffy, Henderson, and Morris (1989) explore the role of lexical versus syntactic effects in the context of a naming task, finding that a single lexical associate in the context is not sufficient to produce a priming effect on the target ("the woman *trimmed* the *mustache*", "the *barber* saw the *mustache*"), but that two lexical

[1]Note that the effect in this experiment is assessed based on a comparison of the second word of the pair with the first word of the pair, which allows for influences of sentence position effects.

[2]Hoeks, Stowe, and Doedens (2004) question how conclusive this result is, given that there doesn't appear to be an additive effect, an observation that inspires the resulting role reversal experiment in Dutch, which we will review in Chapter 4.

Congruent associated When the *moon* is full it is hard to see many *stars* ...
Congruent unassociated When the *insurance* investigators found out that he'd been drinking they *refused* ...
Anomalous associated When the *moon* is rusted it is available to buy many *stars* ...
Anomalous unassociated When the *insurance* supplies explained that he'd been complaining they *refused* ...

Table 1.1: Van Petten (1993) stimulus conditions.

associates do produce priming ("the *barber trimmed* the *mustache*"). They conclude this to be evidence against simple summation of facilitation from associates, but rather in favor of facilitation arising from some kind of combination of the associates. (They suggest that there could be two explanations for the need for two lexical primes: there could be some kind of threshold above which priming is achieved, or the presence of the noun may have constrained the senses of the verb.)

To test whether this is a more modular lexical effect or an effect produced by a higher-level understanding of the sentence message, these authors further compare sentences such as "While she talked to him the barber trimmed the mustache" as opposed to "While talking to the barber she trimmed the mustache". They find priming of *mustache* of the same size in both cases. (Note that this is an early precursor to the role reversal insensitivity results that we will discuss in Chapter 4, in that it holds constant lexical content and changes the message as driven by the syntactic information.) The authors conclude that this is a modular lexical effect independent of the integrated message interpretation.

Complicating the picture, Morris (1994) does a follow-up using eye-tracking reading times, and finds no facilitation for "The gardener talked to the barber and trimmed the

mustache", by contrast to "The gardener talked as the barber trimmed the mustache", which did show facilitation. This result suggests that the reading times in this case were sensitive to the particular message guided by the syntactic information.

We need of course to be careful in comparing results across different types of measures. However, overall what we find in these results is evidence for a complex interplay of factors, such that in some cases the processing of incoming words appears to reflect integration of rich information (including syntax) to generate message-based expectations, while in other cases word processing appears to reflect less sophisticated mechanisms based on lexical relations.

1.1.2 Considerations

Given the above results, how should we think about the mechanisms that are giving rise to these "lexical" effects, and how precisely should we delineate them from other mechanisms?

The primary distinguishing characteristic isolating these lexical processes—refocusing now on the N400 results such as negation and role-reversal insensitivities described above—is that these processes are characterized by a failure to pattern with expectations produced by the syntactically composed representation of the message. This leaves us with two major classes of options for what this asyntactic lexical mechanism could look like.

On one hand, these lexical effects could reflect a mechanism that makes use of all information *but* syntax (or, more precisely, makes use of a larger set of information that excludes at least some components of syntax). By this way of thinking, the N400 could

still involve generating a structured expectation based on world knowledge, but in some cases syntactic information would be omitted from the representation used to consult world knowledge. This conceives of the limitation observed in the N400 as existing within the context of an active process of message inference and expectation generation, and it is reminiscent for instance of Duffy et al.'s suggestion above that these lexical effects consist not just of simple summation but of actual combination processes. This construal also resembles proposed mechanisms such as plausibility heuristics and semantics-only processing streams that are posited by a number of accounts to be reviewed in Chapter 4.

Alternatively, these effects could be based strictly on low-level lexical relations, not reflecting an active message inference process. In this framework, the lexical effects might be considered to arise as a result of more passive, automatic processes, which we could plausibly place under the umbrella of "priming". This would conceive of the lexical effects as facilitation arising automatically as a result of the activation of the words in the context, along the lines of theories of spreading activation (Collins & Loftus, 1975; Quillian, 1967), or theories of priming from feature overlap, as within distributed network models (Cree, McRae, & McNorgan, 1999; Masson, 1995; Plaut & Booth, 2000). Chapter 2 will elaborate further on the details of word relations and mechanisms of facilitation.

If we want to think of the lexical effects as passive, priming-like effects, how can we reconcile this with the observation that priming is typically short-lived, even disappearing with a single intervening word (Neely, 1991)? Various authors have speculated on this question, in light of the unexpected duration of these lexical effects. Kutas (1993), for instance, suggests that N400 reduction based on simple relations between words within a sentence could involve some sentence-level influence on the relevant lexi-

cal processes. Duffy et al. (1989) suggest that during the processing of normal sentences, content words "are maintained in some form of active memory as a byproduct of the higher level syntactic and integrative processes that are required for sentence", continuing to have priming effects as a result. (Though they argue that this is still consistent with a modular view of syntactic and lexical processing, as these effects take place within the lexicon.) We cannot be certain what mechanism precisely might allow for lexical effects in sentence contexts to be carried forward for longer durations than are seen in standard single-word priming—but we are satisfied that such a mechanism could exist, and that the passive construal of lexical effects can consequently be a viable option.

Notably, this suggests that what we are calling "passive" may not in fact be best conceived of as fully passive—or as fully devoid of syntactic influence. At very least, the presence of a structured sentence context may be exerting an influence on whether, or how long, these lexical relations produce facilitatory effects.

Thinking of these lexical effects as arising from (relatively) passive mechanisms has two *prima facie* advantages: 1) this class of account doesn't require an explanation for why there would be active combinatory operations that have access to a wide variety of information, including world knowledge, but for some reason omit key syntactic information. And 2) the passive account affords a ready explanation for the type of timing effects that we will see in Chapter 5, which indicate that lexical effects may drive N400 amplitude when there is less processing time between arguments and verb, while more sophisticated, syntax-/message-based effects may drive the N400 amplitude when more processing time is available. This picture is consistent with lexical processes being fast

and automatic, by contrast to message-based processes that would execute integration of a broader set of information.

On the other hand, potential evidence for a more active construal of these lexical effects comes from Lau, Holcomb, and Kuperberg (2013) who find an increase in lexical-associative facilitation on the N400 under circumstances when lexical relatedness is more reliably predictive of target words—suggesting an active, strategic mechanism influencing lexical facilitation. Because this study uses only single-word preceding contexts rather than sentence contexts, the relevance of these results for sentence processing should be interpreted with some caution.[3] However, bearing that caveat in mind, these results can be taken as possible evidence in favor of an active account of lexical effects in sentence comprehension.

The possibility that lexically-driven effects could be a result of passive, automatic processes brings up key questions about the nature of semantic memory and the relationship of these effects to that semantic memory. If these observed lexical effects arise automatically based on activation of encountered words, then we might very reasonably think of them as simply a byproduct of accessing semantic memory (or the "lexicon"). However, this raises questions about the specific relationship of these processes—which often hinge on associative relations between words—with the access of actual word meanings that will be used for composition of sentence meaning. Along this line, an important related question that of whether these effects are based on relations between words *per se*, or on relations between the concepts denoted by these words. We will not arrive at con-

[3]It could be the case, for instance, that these strategic influences reflect mechanisms typically used for message-based processing, but because the context consists only of a single word, this message-based processing is difficult to distinguish from lexically-based processing.

clusive answers to these questions in this thesis, but we will address the relevant issues in greater detail in Section 2.2.1.

Finally, there is a second set of considerations to keep in mind: that of the distinction between the syntactic processes that drive composition of sentence meaning *per se*, as opposed to the broader processes of accessing world knowledge, which are necessarily involved in generating facilitation based on expectations or predictions in context. Involvement of world knowledge will be a critical component of the "message-based", syntactically-driven effects to be discussed in Chapters 4-5. In Chapter 6, however, our focus will be on processes of composition *per se*, as differentiated from processes that consult general world knowledge to form broader expectations.

1.1.3 Modeling

In the work presented in the subsequent chapters, I use computational modeling to explore the nature and interaction of these different types of processes. Computational modeling allows us to flesh out and test, in a quantifiable way, the hypotheses that we have with respect to observed effects in human experiments—and as we will see in the experiments below, it critically allows us to draw on the properties of individual experimental stimuli, for a finer-grained assessment of the dynamics at play in these experiments. For most of this modeling work, I will be drawing on tools used in the computational field of natural language processing (NLP), which like linguistics and cognitive science is in the business of tackling problems of human language—such that it too has relevant considerations about the nature of lexical and syntactic processes—but which also differs from these

scientific domains in important ways. The role of lexical and syntactic processes in NLP, and the way that the relevant issues manifest in that domain by comparison to the cognitive domains, will be the focus of the next section.

1.2 Lexical versus syntactic processes in machines

1.2.1 Natural language processing

The computational field of natural language processing (NLP) is a domain of artificial intelligence, focused on developing computational systems that are able to perform tasks that require the processing of language.

The fundamental difference between approaches to language in NLP and in linguistics/cognitive neuroscience is that while the latter aim to understand and recreate, as faithfully as possible, the mechanisms and representations employed by the human brain during language processing, the field of NLP is generally not concerned with faithfulness to human mechanisms—except to the extent that such faithfulness contributes toward advancement of engineering goals. So when we talk about problems in NLP, the question is not how to understand or characterize different types of processes in an existing biological system, as it was above. Rather, the question is how we should take a system that is naturally and uniquely human—the system of human language—and design computational systems that are able to process that language and respond in a satisfactory manner in the performance of tasks.

So it is important to note at the outset that the basic problem being tackled in NLP does not intrinsically involve any lexical /syntactic distinction *per se*: NLP systems are built to perform certain tasks, and whether or not the solution to a task involves any notion of lexical representations, syntactic structure, world knowledge, etc. depends on the chosen solution to that problem.

How could a system attempt to solve language tasks without any notion of lexical or syntactic processes? As a concrete example, consider the task of sentiment analysis. The problem in this task is to take as input a passage of some kind (for instance, a product review) and to classify the sentiment of that passage as positive or negative (or somewhere on a non-binary scale). Typical modern sentiment analysis systems will take a machine learning approach, training some kind of classifier to map from features of the input to a sentiment prediction.

In attempting to solve this task, a system could in principle be trained to predict sentiment based, for instance, only on the identities of the characters contained in the passage (how many instances of the letter "a" are in the passage? The letter "g"? The character "!"?). This may not be an approach that would be very successful, but I use it to highlight the fact that there is no in-principle requirement to involve either lexical or syntactic representations or processes when attempting to solve NLP tasks. (And it should be noted that character-level models certainly exist, though they typically involve encoding characters in order and allowing for higher-level representations (e.g., dos Santos & Gatti, 2014; Y. Kim, Jernite, Sontag, & Rush, 2016). But such models are used for a variety of purposes, including text classification tasks such as sentiment analysis (Zhang, Zhao, & LeCun, 2015).)

More realistically, one can imagine a sentiment system that simply receives as input the identities of the words present in the passage, and is trained to predict the sentiment based on these individual words. Because sentiment often correlates well with superficial lexical cues, one can see how this could allow for systems to achieve a reasonable amount of success on this task: many individual words will be strong cues to sentiment (*loved*, *hated*, *amazing*, *terrible*). (It is also easy to see how lack of understanding of the actual message would backfire in the case of more difficult passages—consider, for instance, if any of those terms was being used sarcastically (cf. Ettinger, Rao, Daumé III, & Bender, 2017).)

The important point to make here is that the solution to this problem can, but will not necessarily, involve any notion of lexical or syntactic processes, or the kinds of distinctions that we have been discussing above for humans.

1.2.2 Task-general lexical and sentence-level processes

In actuality, many NLP systems do use lexical representations, and many additionally have some notion of combining those representations to form sentence representations (though this notion of combination may or may not be "syntactic" in any way, as we will discuss below).

What is the relationship of these representations and processes to those of humans? In human language processing, we understand comprehenders not only to be accessing word meanings and composing them to obtain sentence meanings—we understand the word and sentence representations to be fairly uniform regardless of the task for which

they will be deployed. For instance, if a human were to perform the sentiment analysis task, the general understanding is that they would access word meanings, compose the words to obtain interpretations of sentence meanings, and then use additional knowledge of pragmatics, real-world references, etc. to determine whether the writer of the passage is expressing a positive or a negative opinion. If the human's task were instead to come up with a witty retort to a sentence input, we generally assume that they would complete the same first steps of this process to arrive at an interpretation of the sentence, and then do additional processing based on that sentence meaning, in service of generating a reply.

This brings us to the heart of a question that has been the subject of some debate in NLP: whether NLP systems should pursue *task-general meaning representations*. The hypothetical advantages behind this notion are fairly straightforward from a cognitive perspective: if a system can accurately extract and represent the meaning of a sentence, then like a human it should be able to apply that meaning for solving of a variety of tasks (given additional reasoning). Such representations should be more efficient, in being widely applicable across tasks and thus reducing the burden of task-specific learning. Additionally, solving tasks based on meaning representations should be more robust than use of superficial cues, especially for more meaning-bound tasks such as entailment assessment. This is an old idea (see, for instance, the argument in Quillian et al. (1962) in favor of representing invariant meaning properties for machine translation).

However, there are plenty of arguments against taking this route in NLP. A salient practical argument is that we are a long way from being able to do this well—taking this route amounts to mastering a substantial portion of human cognition, in that it requires not only meaning representation and extraction, but also encoding and use of pragmatics

and world knowledge. Consequently, at this point there is often little or no gain observed in system performance from trying to extract meaning *per se*, by comparison to more direct task-specific approaches.

A more fundamental objection would push back on the assumption that there is in fact such a thing as core meaning representations—either at the word or the sentence level—that should be applied across tasks. Perhaps the core meaning components that might be useful across tasks are too few—or the additional post-meaning reasoning occupies too great a portion of the relevant computations—for extraction of task-general meaning to be worthwhile. These are certainly possibilities that must be considered as we strive to improve NLP systems.

Despite these potential objections, recent years have seen increasing interest in learning and deployment of task-general meaning representations, both at the word and sentence level.

A major example of this trend is the substantial surge in use of task-general "word embeddings"—vectors of continuous values that are intended to serve as word "meaning" representations. The popularity of these representations is in large part attributable to their capacity for quantifying graded relations between words, which represents a step in the direction of capturing relevant components of meaning. Use of vectors for representing words will be introduced in detail in Chapter 2, and we will make significant use of these types of representations for modeling in subsequent chapters.

With the success of task-general word representations, there has recently been a corresponding upswing of interest in obtaining task-general sentence representations—often framed explicitly as a search for "composition" models able to construct sentence mean-

ing systematically based on word meaning. These models often take the form of neural networks, and are often used to manipulate and output vector representations of "meaning" like the embeddings described above. We will review models of this kind in detail in Chapter 6.

Do these trends represent steps in the right direction? Should NLP models pursue task-general meaning representations, and the humanlike lexical and syntactic processes that this would presumably involve? To the extent that we want to accomplish "natural language understanding" in NLP (as many purport to)—then based on what we know about language as a system, it is difficult to imagine how such a goal could possibly be accomplished without arriving at proper answers to the questions of how word meanings should be represented, and how those meaning representations should be combined to produce compositional phrase and sentence meaning representations. To the extent that meaning exists as a task-independent concept within the language system, the goal of representing that meaning seems not only sensible, but necessary in the long term. This will be the stance that drives the work presented in Chapter 6.

1.2.3 Should NLP imitate humans?

In pursuing task-general meaning representations and framing issues in terms of lexical/syntactic processes, it is necessary that we ask ourselves—are we trying to force NLP to be too much like humans? A frequent argument against basing NLP systems on human cognition (an argument pertaining to engineering in general) draws on the following analogy: airplanes accomplished flight without flapping their wings like birds, and in fact

improved upon the capacity for flight. Perhaps we can take other approaches with computational systems and, as with planes, ultimately accomplish the tasks more effectively than do the original biological organisms.

This notion bears thinking about for a moment. Can we do language "better" than humans? What would this mean? The answer to this question hinges on the related question of how to evaluate the quality of language processing—a question that we will explore in Chapter 6. But what we will see in that discussion is that for the most part, evaluations of NLP systems share a common principle: they are typically based on comparing system performance to performance by humans. Do the systems give responses that are similar to those given by humans? Do they give responses that humans find acceptable or natural (given typical responses by humans)?

From this perspective, I would argue that by contrast to the flight analogy, there is a basic level at which the notion of doing language "better" than humans is ill-defined: language is a naturally and uniquely human system, and to a substantial extent, the goal of NLP is in fact to "do language" as similarly to humans as possible.

One may quickly object that there are ways in which computational systems can surpass human abilities in language tasks, based on greater memory capacity or speed of accessing stored data. Computational systems can (in principle) translate between more languages than can any individual human, or when asked questions, can consult far more data with far greater speed. This is all true, but it also all falls under the umbrella of performing language tasks more quickly or at a larger scale. When we assess the *quality* of a system's response, it will still be based on how similarly the system behaves to

a human—this assessment will simply make the assumption that the human speaks the relevant languages / has access to the relevant information.

The point here is simply that there is a fundamental level at which guiding NLP systems based on what we know about humans is quite sensible—and I would argue that the goals of representing word meaning, and being able to compose those meanings to achieve sentence meanings, fall solidly under the umbrella of useful insights from the human side.

Having established this, it is important to clarify that this is not to say that NLP systems should be constrained to be identical to human language processing systems. Consider, for instance, the cases discussed above, in which the N400 appears at times to be insensitive to certain information in the sentence, producing a response that apparently fails to distinguish good from bad continuations. From an engineering perspective, there would seem to be no use for a component like the N400 that sometimes reflects insensitivity to relevant information. This type of behavior is likely a function of the way that the brain processes language in real time, and there is no in-principle reason to build such a characteristic into NLP systems—instead, in NLP systems we would ideally be able to circumvent effects that arise as a result of real-time processing constraints, skipping straight to the desired final product. In terms of our discussion above, this would presumably mean skipping mechanisms or processing stages that give rise to asyntactic lexical effects, and instead focusing on arriving at an accurate, syntactically-constrained interpretation of the sentence.

The current reality

The irony of the situation is that despite the lack of in-principle need in NLP for humanlike processes that omit key information about the sentence—such as processes that reflect lexical-level associative relations rather than composed sentence meaning—it is in fact these kinds of asyntactic lexical-associative processes that NLP systems often most resemble and excel at. This is in large part because mastering the structured composition processes (and world knowledge-based reasoning) that would lead to an accurate representation of the sentence message is extremely difficult to master—as we will discuss in Chapter 6.

For this reason, what we will actually do in the subsequent chapters is to take advantage of existing representation and combinatory methods from NLP to model asyntactic real-time processes in humans. Then in Chapter 6 we will present work aimed at improving our ability to evaluate—and, by extension, improve—the capacities of NLP systems to do syntactically-informed composition with complex meaning representations. The latter work will also draw on insights from the human side, in that we will leverage methodology from neuroscience analysis and psycholinguistic experimental design in order to analyze and increase the interpretability of opaque representations produced by NLP systems.

1.3 This dissertation

1.3.1 Contributions

The contributions described in this dissertation are twofold. In Chapters 3-5 I will discuss work that leverages vector space models (VSMs), and their ability to quantify relations between words based on corpus data, to simulate lexical processes and test hypotheses about the influences that those processes could be having on the measured response in sentence processing. I will also use these VSMs in combination with human-derived cloze probabilities, in a novel hybrid modeling approach, to ask targeted questions about how lexically-driven and message-driven effects trade off during sentence comprehension. Using these methods, I identify two promising accounts for explaining the relative contributions of these components in giving rise to the complex set of N400 results represented in studies of semantic role reversal.

In Chapter 6 I introduce an analysis method for assessing abstract compositional information in vector representations of sentences, drawing on methodology from cognitive neuroscience and psycholinguistics. This method allows us to probe for information types relevant to achievement of sentence meaning composition, advancing us toward being able to identify characteristics of NLP models that are more and less effective for capturing different aspects of composition and meaning.

1.3.2 Overview

In this chapter I have introduced the notion of a distinction between "lexical" and "syntactic" processes in human real-time language processing as well as natural language processing systems. In Chapter 2 I will talk in more detail about dimensions of the hypothesis space that we need to consider when doing computational modeling of human language comprehension, and I will discuss the relations of these dimensions to approaches in NLP. I will then discuss how we will use a particular tool common in NLP—vector space models—for modeling aspects of these dimensions. In Chapter 3 I will demonstrate the use of vector space models for simulating asyntactic lexical processes, both in sentence-level processing and in the context of pairwise semantic priming. In Chapter 4 I will move to discussing the phenomenon of role reversals and their effects on ERP components, which will serve as a key case study for thinking about the interplay of asyntactic lexical processes and syntactically-constrained message-level processes. In that chapter I will introduce the relevant literature as well as two existing computational models, and I will report the results of a replication and exploration of one of these computational models. In Chapter 5 I will review two more recently reported role reversal studies, which complicate the picture of the N400 in response to role reversals, and which present a useful foothold for thinking about the interplay between lexical and syntactic processes in giving rise to observed N400 effects. I will then present the results of a series of modeling simulations based on the nature of the stimuli used in those experiments, helping to shed light on the potential nature of the interplay of these different processes. In Chapter 6 I will then shift to discussing the relation of lexical and syntactic processes within

the context of sentence composition *per se*, and in particular the composition of complex lexical content in NLP systems. In that chapter I will discuss work laying the groundwork to improve the capacity of NLP systems to do sentence composition, drawing on cognitive neuroscience and linguistics to develop a method for better assessing the compositional capacities of existing systems. In Chapter 7 I will conclude the dissertation.

Chapter 2

Bringing things together: the modeling hypothesis space and NLP approaches

2.1 Introduction

In this chapter I will delve more deeply into the cognitive questions that surround the distinctions and phenomena described in Chapter 1, and I will introduce some of the computational approaches of relevance to these questions, which we will apply for various modeling purposes in subsequent chapters.

In the first section I will explore three key dimensions of the hypothesis space that need to be considered in our modeling of human language comprehension, and specifically in the modeling of the lexical/syntactic distinctions and effects described above. For each dimension, I will lay out some of the considerations on the human side, as well as related approaches from the NLP side which will be used and discussed in subsequent chapters.

In the second section, I will focus specifically on vector space models, the computational tool that will see the most substantial use in this dissertation. I will discuss the fundamental properties, advantages, and limitations of this framework, the implications of my using it as I do, and its capacity to model the different aspects of the dimensions from the first section.

2.2 Dimensions of the hypothesis space: human questions and NLP approaches

Chapters 3-5 will tackle complex cognitive issues relating to lexical and syntactic processes, drawing on tools and approaches from NLP to assist in modeling and exploring these questions—while Chapter 6 will pivot to address lexical and sentence-level relationships in the context of composition models in NLP, drawing on methodology from work on language in humans.

Critical to modeling and reasoning about these processes are three key dimensions of the hypothesis space, corresponding to three core issues: 1) the nature of lexical representations, 2) the nature of combinatorial operations on lexical units, and 3) the nature of the facilitation that we quantify with our measures. For each dimension I will discuss some of the relevant questions and possibilities from the human side, and link these to approaches and considerations on the NLP side.

2.2.1 Word representations

A core question underlying our understanding of "lexical" effects is that of how words are, or should be, represented. In humans: what information is encoded in word representations? Is there a single representation that encodes all that we know about a given word, or are there multiple types of representation? What components of word representations participate in composition, and are these the same components involved in other processes (like the asyntactic lexical processes introduced above)?

In NLP: what information should be included in word representations such that they perform as desired—and how should these representations be obtained?

Human side

There are many things that we know, or have access to, when we "know" a word. In this section we will focus primarily on exploring the distinction between the concept / "meaning" that a word denotes, versus the associative relations in which that word (and/or its corresponding concept) participates. This distinction will be our focus because it constitutes one of the primary sticking points in thinking about whether our observed "lexical" processes simply reflect access of word meanings, or alternatively, represent a separate parallel process. More specifically, this question arises because associative relations play a significant role in the asyntactic lexical effects discussed in Chapter 1, but it is not clear that we would consider such associative relations to be part of the "meaning" of a word, or the concept that it denotes. If these associative effects occur as a byproduct of accessing the word meaning/representation, then what precisely is the relationship between the

associative links that these effects suggest, and the word meaning/concept that contributes to the composition of the meaning of the sentence?

Pustejovsky (1991) divides theories of word meaning into primitive-based (Lakoff, 1969; Wilks, 1975) and relation-based (Collins & Quillian, 1969; Quillian, 1967), with feature-based theories falling as a subset of primitive-based theories. Primitive-based theories understand word meaning as defined by decomposition into various primitives, while relation-based theories understand word meaning as defined by relation to other words/concepts. There is a level of correspondence here with theories of "lexical concepts", which Laurence and Margolis (1999) divide into classes of complex and atomic approaches. The first class involves concepts being made up of features: in classical and neo-classical theories, the structure of a concept is definitional, encoding conditions that are necessary and sufficient for something to be an instance of the concept, while in prototype theories, features have a more probabilistic characterization—properties that members of the concept category *tend* to have. In the second class, by contrast to these feature-based hypotheses, there are theories along the lines of conceptual atomism, in which lexical concepts are primitives without internal structure, defined instead by relation to the world (Fodor, 1990).

With respect to our question of concept versus associative information, both primitive-(/ feature-) based approaches and relation-based approaches can in principle accommodate the encoding of both of these types of information.

Consider, for instance, the approach of Collins and Loftus (1975), building on Quillian (1967), in which each concept is represented as a node within a semantic network, with concept properties encoded as links to other concept nodes in the network. These links

are labeled such that they can represent a variety of different types of relations, such as superordinate, subordinate, modifier, etc., which suggests that this structure is able to accommodate a diverse variety of information, including lexical entailment information and virtually any type of associative relationship.

What about feature structures? As in the classical and neo-classical theories of lexical concepts, feature structures lend themselves naturally to definitional or prototypical features that we might consider to constitute core meaning: "is a bird", "has fins", "is spiky", etc. However, one can also imagine feature structures encoding associations, perhaps by labeled relational features such as "is commonly near —", "lives in —", "likes to eat —", or alternatively by a more abstract encoding of associative information, as we will see in the vector space models described below. It should be noted that to the extent that features can be defined relationally with respect to other concepts, as I have just done above, feature structures may in fact be considered to a large extent analogous to the structure defined by a semantic network—features would simply correspond to links (or strengths of links). However, to the extent that features represent non-relational primitives, these frameworks do differ non-trivially.

So we see that the available frameworks for word and concept representation are consistent with inclusion of both core meaning and more peripheral associative information. However, word representations don't exist in isolation: they need not only to encode relevant information, but also to participate in various cognitive processes, like the generation of associative facilitation, and the composition of sentence meaning. How does this relate to the nature of these word representation?

Our observed associative effects can be accounted for reasonably straightforwardly by the simple existence of network links or associative features in one or the other of these frameworks, as long as we assume those associations to be activated automatically by the access of representations that encode those associations. This is the idea behind, for instance, spreading-activation theories of semantic priming (Collins & Loftus, 1975).

What about the aspects of word meaning that participate in and inform phrase and sentence composition? Does the composition process filter out associative features/relations and select only core meaning features/relations for participation in the composition process? To what extent do those features inform and influence the composition process?

Pustejovsky (1991) addresses these questions head-on with a theory of word meaning that directly incorporates the manner in which words participate in composition, and the manner in which the composition process affects changes in meaning based on the original word meanings (explaining, for instance, selection of the proper word senses). Pustejovsky's theory divides word meaning into four components: 1) argument structure, 2) event structure, 3) qualia structure, and 4) inheritance structure. Semantic information (primarily that conveyed by nouns and adjectives) is included in the qualia structure, which also incorporates argument and event structure information. These structures then interact with rules of composition that give rise to the meanings of larger expressions. Associative information is also included, within the category of inheritance structure. This theory thus presents one unified example of how word representations might incorporate both conceptual and associative information, and how the composition process could select out and manipulate the relevant components.

NLP side

In thinking about word representations, NLP systems have two primary questions: what should word representations be composed of, and how should they be obtained?

NLP has used a variety of methods for encoding information about words. Some involve hand-construction of resources by humans: examples include semantic networks with hierarchical and other relations, like WordNet (Miller, 1995), as well as databases of feature-based lexical representations more along the lines of Pustejovsky's above, like VerbNet (Schuler, 2005).

A major disadvantage of such approaches from the engineering perspective is that they require painstaking and subjective processes of expert human labor to craft word representations for use by systems. So by contrast, other methods involve automatic learning of representations from text. Examples include topic modeling, which conceptualizes words as being generated probabilistically based on topics (Blei, Ng, & Jordan, 2003; Griffiths, Steyvers, & Tenenbaum, 2007; Hofmann, 1999), and Brown clusters (P. F. Brown, Desouza, Mercer, Pietra, & Lai, 1992), which cluster words into classes so as to maximize mutual information based on the co-occurrence of those classes. There have also been learning-based approaches that arrive at more explicitly structured lexical representations, by defining a number of primitives and rules for forming representations, and then designing algorithms to learn these lexical representations for certain words based on simple synthetic learning environments (e.g., Pustejovsky, 1987; Salveter, 1979; Siskind, 1990, 1996).

An automatic representation learning approach that has seen a particular explosion of interest in the context of task-general word representations is that of vector space models (VSMs). Here I will introduce these in greater detail, as they will figure prominently in subsequent chapters.

VSMs of word representation have existed for decades: they were first popularized within cognitive science by Landauer and Dumais (1997) with the introduction of Latent Semantic Analysis (LSA), which the authors proposed as a model to account for acquisition of knowledge of words. The intuition behind LSA is that necessary knowledge about words is reflected in the types of contexts that they occur in, and consequently, word representations can be derived based on their distributional characteristics. Specifically, LSA takes counts of how many times a word occurs within each of a large number of different paragraphs, transforms these counts by applying a log transformation and dividing by the entropy of that word across contexts, and then performs dimensionality reduction using singular value decomposition.

The result is a multidimensional vector representation for each of the words of the vocabulary, built from the distributional profiles of those words. Because of the dimensionality reduction, the vector dimensions are not easily interpretable, and the precise relationship between the final representation and the original distributional counts of the word is not straightforward. However, words with similar distributional profiles are likely to have similar representations.

Speaking in other terms, the result is the capacity to situate all word representations within a multidimensional vector space, such that words with more similar behaviors will fall more closely together in that vector space. Inspired by the capacity to "embed"

word representations within the vector space, these vector representations have come to be referred to as "word embeddings".

Subsequent years have brought different methods of deriving these representations (e.g., Bengio, Ducharme, Vincent, & Jauvin, 2003; Mikolov, Chen, Corrado, & Dean, 2013; Turney & Pantel, 2010), but they share the same basic advantage: the ability to represent words in a distributed fashion, as sets of dimensions, such that graded similarity relations between individual words can be quantified. For NLP purposes, what this is intended to allow is a rough correlate of meaning, which allows for greater generalizability. As an example of this advantage: in our above sentiment example, the classifier would need to learn sentiment mappings for each individual word, but with word vectors, the classifier can make use of similarities between words—if *loathed* and *hated* have similar representations, and the classifier has learned that *hated* tends to map to negative sentiment predictions, then it can use the similarity of *loathed* to map to a similar prediction.[1]

What do the resulting representations tend to capture in reality? Studies have shown that vectors derived in this fashion often capture not only (and sometimes not necessarily at all) the type of information that we would consider to fall under the "meaning" heading above—rather, they often show behaviors that seem to capture more associative relations (Agirre et al., 2009; Hill, Reichart, & Korhonen, 2015). This is a characteristic that we will exploit for modeling purposes, in trying to capture the types of lexical relations that characterize our asyntactic lexical processes.

[1] See related discussion on the value of modeling synonymy in, e.g., K. S. Jones (1965).

2.2.2 Combinatorial

Our next question involves how word representations are combined to form more complex representations, from the perspective of structured sentence meaning composition, as well as the incremental processes of real-time language comprehension.

Human side

As we comprehend a sentence, we encounter words incrementally, presumably developing a form of complex representation reflecting the string of words that we have seen. What is the nature of the combinatorial processes that give rise to these new representations? I will describe two classes of possibilities, inspired by the observations and distinctions made in Chapter 1.

The first class of combinatorial hypotheses is that involving structured, syntactically-driven composition of lexical items. This general notion of composition is dominant in the domain of linguistics, and is typically understood to involve a process guided specifically by the syntactic structure of the sentence, in addition to the particular semantic properties of the lexical items themselves, which have complex but predictable interactions with the semantic properties of other lexical items (Heim & Kratzer, 1998).

The problem of semantic composition has been addressed in great detail within linguistics, but there are two dimensions of the problem that are of interest to us here, and that do not fall within the typical purview of formal compositional semantics.

First is our interest in how lexical content interacts with the composition process. Compositional semantics standardly focuses on the general nature of the combinatorial

system, and the characterization of functional meanings, rather than on the specific content of lexical meaning and how these meanings might interact with, affect, and be transformed by the compositional system. However, the manner in which complex lexical content participates in composition is an issue with important bearing on our understanding of the human language capacity, and it ties in centrally with our overarching question of how lexical processes and sentence-level processes relate. (The necessity of solving this problem also becomes particularly salient when attempting to construct and represent sentence meaning in NLP, motivating our focus on it in Chapter 6.)

As discussed above, Pustejovsky (1991) is one example of work that explicitly addresses interaction of lexical content with the compositional system, providing an account of how this interaction can give rise to different word senses, especially during functional composition of verbs and their arguments. Another literature that addresses the problem of composing complex lexical content is that of "concept combination". Much of this literature is focused on combination of nouns, and the interpretation of resulting compounds (Costello & Keane, 1997; Estes & Glucksberg, 2000; Gagné & Shoben, 1997; Wisniewski, 1996), while others discuss adjective-noun composition (Franks, 1995; Murphy, 1988; Smith & Osherson, 1984). These theories typically involve identifying the particular strategies by which properties or relations are selected for the output representation. For instance, focusing on adjective-noun combination, Smith and Osherson (1984) argue that nouns consist of structures with dimensions that have values, and adjectives serve to select a relevant dimension in the noun representation, modify that dimension's value, and add salience to it. Murphy (1988) argues that this is insufficient—that the combination indeed involves representations containing slots that are filled, but that this is not

deterministically asymmetric in the manner argued for by Smith and Osherson (1984). Instead, world knowledge is needed to determine which words will function as slot-fillers and which will be filled.

This literature presents a useful starting point for thinking about these processes, but leaves a large majority of the issue that is yet to be understood.

Our second deviation from the standard approaches of compositional semantics is that we are concerned here with incremental construction of context representations in real time. For instance, when studying the effects of context processing on the processing of an incoming word, we are interested in the nature of the representation of "the owner knew which waitress the customer had —", as it relates to and influences the processing of the incoming word *served*. We assume that syntactic structure can, and (at least sometimes) will, be reflected in that context representation and accordingly affect the processing of the incoming word. However, because this represents an incomplete sentence fragment, we need additional considerations beyond those that assume a full structure. A literature that has tried to tackle this is that of incremental parsing (Abney & Johnson, 1991; J. T. Hale, 2014; Nelson et al., 2017; Resnik, 1992).

However, in light of the asyntactic effects described in Chapter 1, it is sensible also to consider the possibility of a less structure-based combinatorial process that could be at play, even if this applies a) only as an intermediate processing stage en route to more structured representation, or b) simply under conditions of reduced time or attention. This kind of mechanism has been proposed in a number of accounts and will be discussed in greater detail in Chapters 3-5.

This brings us back to the question discussed in Chapter 1 of whether these asyntactic lexical effects arise from an active process operating on incomplete information, or a more passive, automatic lexical-relation-based process. If the effects arise from an active but limited process, then it is sensible for us to suppose that some kind of asyntactic combinatorial process must be at play in order to arrive at an interpretation. Alternatively, if we believe that these lexical effects arise from a more passive processes, then it might be appropriate to consider these lexical effects to bypass any true combinatorial mechanism entirely, and instead to reflect a simple aggregate relational effect on the incoming word.

NLP side

How do we move from the word level to the phrase/sentence level in NLP? Unless an NLP system is receiving a sentence that unfolds in real time, there is no in-principle reason for NLP systems to be constrained by incremental processing. Nor do NLP systems have an obvious need for any unstructured combinatorial mechanism that will sacrifice sentence information. So this narrows our field to an ostensibly more straightforward question: given the lexical representation structure that we have in play, how do we compose those lexical representations to obtain phrase and sentence representations?

There are some approaches that have implemented rule-based combinatorial processes for the more structured, human-crafted lexical/concept representations described above—among these are examples that have already been discussed (e.g., Pustejovsky, 1991; Siskind, 1996), as well as examples such as Lynott, Tagalakis, and Keane (2004), who implement the constraint theory proposed in the concept combination literature for interpretation of noun-noun compounds (Costello & Keane, 1997). Another approach that

makes use of structured representations is broad-coverage semantic parsing, which trains models to map automatically from sentences to logical forms (e.g., Banarescu et al., 2013; Berant, Chou, Frostig, & Liang, 2013; R. Mooney, 2004).

As an alternative to use of explicitly-structured representations, which often involve time-consuming and subjective annotation processes, many current NLP systems instead make use of combinatory operations that are based on more efficiently-derivable representations, such as vector space representations. I will focus in this section on approaches to deriving phrase/sentence representations within the vector space framework, as this is dominant in current NLP, and will be our object of attention in subsequent chapters.

Within this framework, we find once again that despite the apparent lack of need for either incremental processing or non-syntactically-driven combinatorial operations, solutions in NLP to deriving sentence vector representations are typically characterized by both. Chapter 6 will cover NLP sentence composition models in much greater detail, so here I will simply take the opportunity to introduce the framework within which many of these models operate: recurrent neural networks.

Neural networks have been around in NLP and cognitive science for decades (McCulloch & Pitts, 1943; Rosenblatt, 1958; Rumelhart, Hinton, McClelland, et al., 1986; Rumelhart, McClelland, Group, et al., 1986). (In cognitive science, use of neural networks fits within a tradition known as "connectionism".) Mathematically, neural networks are simply complex functions that map from an input vector to an output. Often, particularly within cognitive science, neural networks are conceived of as systems of layers of nodes, which are connected to nodes in other layers, and which send activations progressively through the network after an input activates the first layer. These networks

can be trained to make mappings of a certain kind, by submitting an input and sending activations through the network, calculating the error between the network's output and the desired output, and incrementally adjusting the network's connection weights to improve the output in the correct direction.

Recurrent neural networks (RNNs) are commonly used for language, as they allow for the network to retain an evolving representation of the history of inputs (Elman, 1990). This is achieved by taking components of the internal representation(s) within the network at the most recent timestep, and retaining those components to be submitted as input at the next timestep. This means that we can, for instance, present a sentence to the network word by word, and at each point the network will process not just the current word, but also the stored information about previous words. The representation that is arrived at (at some location in the network) after the entire sentence has been input is considered to be the sentence representation.

What we can see about this approach is that it operates in a highly incremental fashion, much akin to the left-to-right processing that the human brain does in real time. Additionally, typical RNNs include no explicit role of syntactic structure to guide the composition of words. (There are exceptions to this generalization—RNNs that do make explicit use of syntactic structure—which will be discussed in greater detail in Chapter 6. Additionally, there are claims that RNNs can learn and represent structure implicitly (e.g., Elman, 1990; Linzen, Dupoux, & Goldberg, 2016), an issue that we will also explore in Chapter 6.)

At an even greater extreme of asyntactic modeling, another widespread means of combining word vector representations to obtain phrase or sentence representations is simply

to average the word representations together in "bag-of-words" fashion, sacrificing all order information completely. This is another model that we will assess in Chapter 6, and that we will in fact use to model asyntactic processes in Chapters 3-5.

The result is that the combinatory methods that currently dominate NLP are in fact closer in nature to the asyntactic, incremental processes that may arise in humans as a result of the constraints of real-time processing—rather than the time-independent syntactic processes that we would associate with structured meaning composition. This is in large part due to the significant challenges of achieving structured composition of sentence meaning, as we will discuss in Chapter 6.

We will see neural networks on both the cognitive and NLP sides in Chapters 4 and 6.

2.2.3 Facilitation

The final piece that we need to consider in modeling the human language processing results is the mechanism by which the processing/representation of context affects the processing of the incoming word.

Human side

This third dimension comes off as a bit of an outlier in this discussion: the relevance of word representations and combinatorial operations to our interest in lexical/syntactic relations should be relatively clear, but it may be less apparent at the outset why we need to characterize the state of preparation for incoming words that is generated by the preceding contexts of those words. Indeed, the motivation for inclusion of this dimension comes primarily from the fact, discussed above, that the majority of the measures that we

use in psycholinguistics—such as the N400 and semantic priming, both to be discussed below—appear to index facilitation in processing. In light of this fact, it becomes quite important to characterize the nature of the mechanisms giving rise to this facilitation, if we are to understand what this facilitation can tell us about our other dimensions of interest. In this section I will highlight three popular frameworks for conceptualizing the mechanisms underlying facilitation of incoming words (frameworks which, as I will discuss below, need not necessarily be mutually exclusive).

The first framework is probability-based: the context generates a probabilistic expectation with respect to the upcoming word, and the more probable a word is with respect to this expectation, the greater its facilitation when it arrives. This seems to be the class of mechanism envisioned by probabilistic estimation methods such as cloze (described above) and model-generated surprisal, which consists of inverse log probabilities of words given their context. Like cloze, surprisal has also been found to correlate with measures of facilitation such as reading time and N400 (Demberg, Keller, & Koller, 2013; Frank, Otten, Galli, & Vigliocco, 2015; J. Hale, 2001; Levy, 2008).

A second possibility for conceptualizing this facilitation mechanism is what I will refer to as "overlap-based": the context results in pre-activation of various conceptual features, and the greater the pre-activation of the features of the incoming word, the greater its facilitation. An overlap-based mechanism is more along the lines of what has sometimes been proposed as a mechanism for semantic priming within distributed network models (Cree et al., 1999; Masson, 1995; Plaut & Booth, 2000).

The third conceptualization is that of a spreading activation account (Collins & Loftus, 1975; Quillian, 1967), mentioned above, which would hold that words are activated by

merit of their network-based connections to words that have already been encountered in the context. This typically conceives of words as existing within a semantic network, and word-to-word facilitation to occur based on activation spreading between connected nodes.

Let us compare these different possibilities. The overlap account contrasts with the other two in that it does not require any pre-existing assessment or activation of the specific target: the probability account requires a probability estimate to exist for the target word, and the spreading activation account requires that the target word be specifically activated by connections in the network—but the overlap account simply states that features of the target have been pre-activated. As we will explore in subsequent chapters, this could simply mean that words in the context shared features with the target word, and because those features were activated by the context words, those components of the target word were also activated.

How does the notion of "prediction" fit into this hypothesis space? This of course depends on our definition of the term. If "prediction" is satisfied by the existence of some probability distribution based on expectations about the upcoming word, then what I am referring to as the probability-based mechanism would qualify as a version of prediction. Alternatively, a more stringent (and perhaps more intuitive) definition of prediction might require that a specific high-probability word or concept (or a small number of words/concepts) be selected as "predictions" based on the probability distribution. If pre-activation of a word/concept by passive spreading activation qualifies as prediction, then the spreading activation account can qualify as prediction—though alternatively, we might want prediction to be considered an active or strategic process. Finally, if pre-

diction is satisfied by simple pre-activation of features of not-yet-seen words, then our overlap-based facilitation mechanism can qualify as prediction as well. Alternatively, we can imagine a variant on the overlap-based mechanism in which features are pre-activated not simply because they are shared by context words, but because of anticipation of features of the upcoming target itself—this is a scenario which quite plausibly would qualify as prediction even by a more stringent definition of the term.

It is important to note that, as in the previous section, these classes of possibilities are not mutually exclusive. In principle, for instance, a probability estimation of some kind could be responsible for pre-activation of features of an upcoming word—perhaps if the probability exceeds some threshold. Spreading activation could also in theory be a mechanism of activating features which then generate facilitation based on overlap.

NLP side

There is no specific need for a notion of a "facilitation" mechanism *per se* in NLP, but NLP has many ways of quantifying relation of preceding context to other words, many of which have also been used for cognitive modeling. One clear point of overlap here is in the estimation of probability distributions over upcoming words in a task known as "language modeling". These models are used to assign probability distributions to sequences of words, and in doing so to help in determining which are good sequences and which are bad sequences. Such models have been used to estimate the surprisal measure introduced above, which has then been correlated with human behavioral and neural measures (e.g., Frank et al., 2015).

NLP also makes widespread use of relations encoded in network structures, analogous to the semantic networks that frame the spreading activation account. One notion with clear resemblance to spreading activation is that of label propagation (X. Zhu & Ghahramani, 2002), which uses network connections to infer labels for unlabelled data. Another network-based method relevant to the tools that we will discuss below is that of Faruqui et al. (2015), who construct a network structure based on similarity relations defined in WordNet, and use the links in this structure to modify vector space representations to better reflect these similarity relations.

Finally, quantification of relations between vector representations has given rise to frequent use of a measure known as "cosine similarity", which is based on the angle between vectors, and which is closer to 1.0 the more similar two vectors are in their direction in the space. This bears a good deal of resemblance to our overlap-based mechanism of facilitation, and it is what we will often use to model facilitation in the experiments below.

2.3 Using vector space models for modeling human processes

Vector space models will figure prominently in the coming chapters, both for modeling of human cognitive processes and for modeling of composition in NLP. Before we continue, we need to think carefully about the implications of using VSMs to do these things. What assumptions are we making by using VSMs, and what fundamental limitations do they have?

2.3.1 Fundamental suitability

The first thing that we need to discuss is the fundamental suitability of VSMs as a framework for modeling hypotheses, and to do this we need to be clear about which things are necessary assumptions and characteristics of VSMs, and which are not. It is important to separate the standards of conventional usage of VSMs from their fundamental and necessary properties.

The primary core property that I assume for VSMs—and what I consider to be the primary fundamental assumption that we make in employing them—is their use of *distributed* representations. Distributed representations are characterized by complex, non-atomic structure, with information distributed across a number of different dimensions. This can be contrasted with localist (or "one-hot") representations, which represent words by assigning a single dimension to each word in the vocabulary. The immediate advantage of distributed representations is the ability to quantify similarity between representations in a graded fashion. Distributed representations have long been exploited in cognitive science in the context of connectionism (Rumelhart, Hinton, McClelland, et al., 1986; Rumelhart, McClelland, Group, et al., 1986)—in connectionist networks, distributed representations are derived by training a neural network and extracting the internal representations that the network learns over time (as described above for RNNs). Each dimension of the vector representation corresponds to a node in the relevant representational layer of the network, and in order to represent a given item, these nodes will each take on some level of activation, specified by a continuous value. (In current NLP, it is often the case that these representations are learned directly as sets of network weights.)

If we relax our expectations of VSMs away from the specifics of conventional usage, instead focusing primarily on their exploitation of distributed representation, we afford ourselves the flexibility to address certain objections that have been raised with respect to use of VSMs as representational models.

Use of a continuous space and distance metrics Prominent among objections to the VSM framework is that of Tversky (1977), who argues not against vector space representations *per se*, but rather against the accompanying use of distance metrics within those spaces as measures of similarity, on the grounds that similarity does not follow the metric axioms of minimality, symmetry, and the triangle inequality. Tversky instead proposes a non-metric feature matching approach to similarity.

Another prominent argument comes from Griffiths et al. (2007), who compare LSA to topic models as approaches to representing words, observing that topic models are better-equipped to deal with polysemy and homophony, and that topic models outperform LSA in predicting human word-association data. These authors argue that their topic models are also compatible with Tversky's feature-based approach, with the features of a word being topics under which that word has high probability.

I see no reason to argue for VSMs as an alternative to feature structures, or for distance metrics to the exclusion of non-metric similarity measures. I consider the class of distributed representations to include and in fact embody the concept of a feature structure. Features may be defined in a wide variety of ways—whether as hand-crafted conceptual characteristics, as topics, or as dimensions within a neural network layer. Any of these options should in turn be representable within a corresponding feature space, if

we are willing to allow for continuously varying feature values (we can also have binary feature values, if we constrain our continuous values such that they are always within a small margin of zero or one). Similarly, use of VSMs need not constrain us to a particular manner of quantifying similarity, or of dealing with polysemy. There has been work using asymmetric measures of relations in a vector space, and there has been a multitude of work exploring possibilities for dealing with polysemy in VSMs (Guo, Che, Wang, & Liu, 2014; Huang, Socher, Manning, & Ng, 2012; Neelakantan, Shankar, Passos, & McCallum, 2015; Reisinger & Mooney, 2010; Tian et al., 2014). In the simulations in this thesis we will take a fairly conventional approach to these issues—we will use cosine as a measure of vector relations, and we will not attempt an explicit treatment of polysemy—but it is important to note that these are not fundamental necessities of the VSM framework.

Non-interpretability of features One common objection to equating VSM dimensions to features is that VSM dimensions often are not directly interpretable. There are two relevant points here.

First, non-interpretability of dimensions is not a commitment of VSMs, but rather a function of the particular algorithms being used. Some current distribution-based VSMs do in fact have a high level of interpretability of dimensions (Fyshe, Wehbe, Talukdar, Murphy, & Mitchell, 2015). However, a VSM could also in principle be constructed with hand-crafted features of the kind more conventionally used in theories of word or concept representation—we would simply need an adequate framework to assign values to every feature for all words in the vocabulary (or at least all words being tested). Obviously this

is a very difficult proposition—though we will see an example of this, implemented on a small scale, in Chapter 4.

Second, it is worth further examining the assumption that words must be structured in the form of features that are sensibly interpretable on a conscious level for our analytical purposes. If we suppose that these features are to correspond to the distributed representation of a word or concept in the brain, what cause do we have to assume that each of these features will align with an intuitive and identifiable category? I would argue that from a neural perspective, the notion of uninterpretable features may be entirely plausible.

That said, the problem of interpretability of models and representations will be a recurring theme in both Chapter 5 and Chapter 6.

Use of distributional information Another common objection to VSMs as representational models is their reliance on distributional information to infer the nature of the representations. (Note the important distinction between the term "*distribution*-based" which refers to the use of distributional information in text, and the term "*distributed* representation" which as described above refers to non-atomic representations). This objection is particularly relevant when considering whether these representations can count as meaning representations. Two related objections fall under this umbrella, one of principle and one of practicality. The in-principle objection is that using the distribution of a word to infer its meaning is too indirect—the meaning will surely influence the word's distribution, but the distribution is not equivalent to the meaning. The second, practical objection is based on the observation that distribution-based VSMs often capture associa-

tive relations between words—which, as we have discussed above, by many construals would not be included in word "meaning".

It is important to note first of all that the VSM format is not committed to the assumption that learning is based upon word distributions. As I establish above, one could in principle design distributed representations in a variety of ways, including using manually-designed features. In a slightly more practical direction, when learning word representations within the context of a neural network, one could learn word features while training on all kinds of learning objectives, which may or may not encourage the learned word representations to reflect distributional properties.

Having established this, it is important to note that in the following chapters we *will* however be making use of distribution-based VSMs, so it is worth thinking about what that means for the modeling enterprise. We will do this in Section 2.3.2.

Fixed-length and representing phrases and sentences Finally, a notorious objection on the NLP side to use of vectors, in particular for use in representing higher-level linguistic units like phrases and sentences, is that a fixed-length vector is not adequate for capturing all of the information encoded in a sentence. Ray Mooney has famously been quoted as saying "You can't cram the meaning of a whole %&!$# sentence into a single $&!# vector!" (R. J. Mooney, 2014).

This matter remains up for debate, but unlike the other objections above, it could in fact represent a fundamental limitation of this representation framework. One way around this objection could be to use variable-length vectors in some way, but it remains to be

determined whether continuous feature structures are an adequate way to capture sentence meaning representations.

2.3.2 Use of distributional characteristics

Although the use of distributional information is not a fundamental requirement of learning word vectors, I will be using distribution-based vectors in the subsequent chapters. Here I discuss some arguments for, and implications of, this approach.

The immediate advantage of distribution-based representation learning is one of cost-efficiency. Distributional characteristics carry a good deal of useful information—for instance, consider what we can infer about a word if we know that it always occurs near words like *eat* and *hungry* (Harris, 1968). Furthermore, distributional information is available essentially free of cost: as long as we have access to written text, we can obtain information about the distributional characteristics of words, without requiring any additional annotation of that text. For engineering considerations, this is a massive advantage.

But what good are such representations from the perspective of modeling human language? If these representations are derived based on co-occurrences and distributional profiles, and they yield word relations such that associated words like *eat* and *hungry* are considered "similar" or at least "close", what good are they?

While these vectors may not hold water when it comes to modeling word meaning *per se* (contrary to prevailing views in NLP) it should be clear from the above discussion that we do have cause to model associative relations, given our observation in human language comprehension of asyntactic lexical effects that are sensitive to lexical association.

For this reason alone, having access to representations that capture associative relations stands to be of value for the modeling enterprise. However, the connection may in fact go further: particularly if we consider these effects to be a result of passive, automatic processes, then a learning procedure much like that used by these distribution-based VSMs could serve as a plausible origin story for the existence of these effects. Specifically, one can easily imagine that over time, with repeated exposure to words in context, the brain would develop fast-acting and passively-available connections between words and/or concepts that co-occur often or are distributed similarly. Recall that this is essentially the same idea as that behind semantic networks and spreading activation—the only difference is that in the case of these vectors, these connections are encoded in the form of shared featural characteristics between words that are associated.

It should be noted that there is also an argument to be made for humans learning actual word meaning (or at least components of it) in a similar way—this is in fact the argument made by Landauer and Dumais (1997). In learning to map lexical items to meaning representations, humans too lack access to any "annotation" of the relevant meanings. Rather, human learners' primary source of information seems simply to be a great deal of experience with words in context. To be sure, the human learning and representational mechanisms that make use of this context information are most likely different from those employed by VSMs (and an important difference in humans is the involvement of perceptual information from context—for instance, part of what humans know about the meaning of *cat* is what a cat typically looks like, whereas conventional distribution-based VSMs have access only to text). Still, it is worth considering that VSMs based on word distributions

in text could in fact simulate to a reasonable extent elements of what humans pick up on and make use of during word learning.

It is also worth noting that Landauer and Dumais draw a distinction between "local conditioning or associative processes" (which would correspond to our passive mechanism above) and "global representation of knowledge"—they consider their initial transformed co-occurrence matrix to correspond to the former, and their dimensionality-reduced final product to correspond to the latter. We will draw the line a bit further along, in that we will consider even reduced-dimensionality VSMs to be a plausible model of conditioning/associative processes—and the "global representation of knowledge" to be a separate goal that we have yet to attain.

2.3.3 Using vectors to model the dimensions

Bringing everything together now: how can these VSMs be used to model aspects of the dimensions described above?

Word representations. As I have argued above, the primary commitment that I am making by using VSMs to model word representations is that those representations are distributed—complex feature-based structures—rather than atomic. I will be using distribution-based vectors, but importantly, when I use them to model asyntactic lexical processes this is not intended to make any claim about the representation of word meaning components that contribute to composition of sentence meaning—the use of vectors to contribute to actual meaning composition will not be addressed until Chapter 6. In doing this I am,

however, making assumptions about the nature of the representations that give rise to those particular lexical effects.

Combinatorial. When I use these vectors to model asyntactic lexical processes, I am primarily using an averaging operation to combine their effects. One could argue that this is in fact equivalent to the kind of combinatorial mechanism that might be operating in the active, plausibility-heuristic version of the lexical hypotheses discussed above. However, we could also think of this as a more passive summative process, without active combinatory operations involved. At the level at which these things are specified, I am not sure that we can distinguish those two possibilities with the present methods.

The one thing that we know I am not simulating when averaging vectors together is syntactic composition—and this accords well with the fact that this is the primary component that we have identified to be absent in these lexical processes that we observe in the N400. In Chapter 6 we will dig into questions of how to use vector representations to do true syntactically-driven composition.

Facilitation. How many of our facilitation mechanisms are VSMs suited for modeling? VSMs can be used to produce language modeling probabilities—Bengio et al. (2003) showed that use of distributed representations can improve the ability of a neural network to estimate a probability distribution over upcoming words. This requires that we create or learn a function from the vector representations to the probability distribution, but it is possible.

As for the overlap-based conception of facilitation: VSMs are particularly intuitively suited to modeling an overlap-based mechanism due to their straightforward use of the cosine measure to quantify the extent to which two vectors "overlap" in space, i.e., are oriented in similar directions, taking into account all of their dimensions. As I have mentioned, this is the primary mechanism that I will focus on with respect to the use of vectors.

What about the spreading activation conception of facilitation? It is less clear on the face of it how vector representations could be used to embody a spreading activation account, since this type of account is conceived of within a network structure. However, as we hinted above, we can potentially think of the network structure as analogous to a feature structure in which the features of a word/concept are relation-based, and the values of these dimensions could be thought of as the values of the connecting edges. This amounts essentially to a co-occurrence-based VSM without dimensionality reduction. If we think of the network as being instantiated in this manner, it essentially becomes an analogue to a VSM. But rather than using cosine to quantify facilitation, we would essentially be looking at a single dimension of the vector to see the extent to which the relevant target word would be activated.

For modeling facilitation based on asyntactic lexical processes, we will be making use of the overlap-based mechanism, in that we will be using simple cosine to quantify similarity in the dimension values of the vector representations.

When we move to modeling the interplay of lexical and message-level effects in Chapter 5, we will also incorporate a probability-based facilitation mechanism, which we posit to be associated with the message-level facilitation.

2.4 Conclusion

In this chapter I have delved into the relevant issues that we must consider in understanding and modeling the problems of interest to us in this dissertation, bringing together questions and tools from both human and NLP approaches to language.

First I discussed three dimensions relevant to the lexical/syntactic distinction that we are exploring: word representations, combinatory operations, and facilitation mechanisms. For each of these dimensions, I discussed some of the considerations that we have in thinking about the hypothesis space on the human side, and some of the tools used in NLP to address corresponding needs. I introduced vector space models as a tool used in NLP for representing words, and recurrent neural networks as a tool used for mapping to representations of sentences from their component words.

I then discussed the use of vector space models as a tool for modeling human language processing mechanisms. I discussed the core properties fundamental to VSMs, clarifying limitations and addressing potential objections to their use. Finally, I discussed how these models can be used to address the three dimensions discussed in the first section, and how VSMs will be used in the modeling experiments to be discussed in the following chapters.

Chapter 3

Modeling asyntactic lexical processes with vector space representations

[Section 3.2 in this chapter is adapted from Ettinger, A., Feldman, N.H., Resnik, P., & Phillips, C. (2016). Modeling N400 amplitude using vector space models of word representation. *Proceedings of the 38th Annual Conference of the Cognitive Science Society.*]

[Section 3.3 in this chapter is adapted from Ettinger, A. & Linzen, T. (2016). Evaluating vector space models using human semantic priming results. *Proceedings of the First Workshop on Evaluating Vector Space Representations for NLP, ACL 2016.*]

3.1 Introduction

In this chapter I will begin the modeling discussion by focusing exclusively on the use of vector space models (VSMs) to simulate asyntactic lexical relation-based processes, setting aside for now the syntactically-driven message-based processes from which these

lexical processes are to be distinguished. (We will return to the message-based processes in Chapters 4 and 5.)

First, in Section 3.2, I will use VSM-based models to demonstrate the potential for asyntactic lexical processes (if modeled in this way) to account for N400 patterns even in cases when lexical-associative relations between context and target are not immediately clear by inspection. Our case study will be the influential result of Federmeier and Kutas (1999), which stands as another important example of N400 amplitude deviating in informative ways from the predictions of cloze probability. The successes of our simulation of this study highlight an important advantage of using computational models such as VSMs to quantify lexical relations (particularly when quantifying associations that could plausibly arise from passive co-occurrence-based learning like that used to obtain these vectors): these models give us the potential to tap into complexities of word relations in a way that our conscious intuitions about association, or norming tasks like free association, may not have access to. For instance, one can imagine two words like *baseball* and *touchdown* having some level of facilitatory relationship in the brain (perhaps due to mutual co-occurrence with similar sports-related terms and contexts) despite the fact that these two words typically would not be considered lexical associates.

Whether the relations produced by these vectors deviate from conscious intuition in a way that is in fact neurally valid is an open question—and one that we will take a stab at exploring in Section 3.3. In that section, we will examine the correspondence between a) relations produced by several VSMs and b) magnitudes of priming effects across a large database of behavioral priming for a number of stimulus onset asynchronies (SOAs) and behavioral measures. This represents a brief deviation from our focus on the N400,

but it allows us to explore how a number of slightly different distribution-based VSMs might vary in their capturing of variance in priming magnitudes, and also to explore how differences in SOA and behavioral measure affect the extent of correspondence between the priming magnitudes and the VSM relations.

3.2 Case study: a lexical account of Federmeier and Kutas

In this section I will simulate the pattern of N400 amplitude observed in a study that has been influential due to results that deviate in an informative fashion from the predictions of cloze probability: Federmeier and Kutas (1999). What we will find is that contrary to the authors' assumptions, an explanation based on asyntactic lexical relations can account for a number of the interesting components of the N400 patterning that they observe. This presents a compelling alternative account, and highlights the need to take seriously the potential influence of asyntactic lexical effects on our observations of the N400, even when this influence is not immediately apparent.

In related work, another computational model that has been used to simulate the Federmeier and Kutas result is that of Rabovsky, Hansen, and McClelland (2016), which will be described in detail in Chapter 4. That model, a connectionist neural network model, successfully simulates the Federmeier and Kutas result—but given the opacity of the functioning of the model (an issue that will be a focus in Chapter 4), it is not entirely

clear whether that model presents an alternative to the Federmeier and Kutas account, or an implemented proof-of-concept.

Another related model is that of Parviz, Johnson, Johnson, and Brock (2011), who make use of a simple combinatorial mechanism on a distribution-based VSM, in order to generate one of several predictors of N400 amplitude (though not of this particular N400 result).

3.2.1 Federmeier and Kutas (1999)

In the study of interest to us in this section, Federmeier and Kutas (1999) investigate N400 effects to target words with varying levels of similarity to the expected word within the context. The key finding of this study is that in highly constraining contexts, not only the expected targets show reduced N400 amplitude, but also unexpected (zero-cloze) targets that fall in the same category as the expected item.

To investigate the effects of semantic category on N400 amplitude, Federmeier and Kutas construct two-sentence contexts with three possible ending types: "expected", "within-category", and "between-category". Expected targets are predicted in the context, as evidenced by high cloze probability. Within-category and between-category targets are both unexpected in the context, with cloze probability of approximately zero—however, within-category targets share a category with the expected target.[1] If N400 amplitude were to track the cloze probability of these stimuli, then we would see reduced N400 amplitude for the expected target condition, and roughly identical, unreduced N400 am-

[1] Federmeier and Kutas explain that "Categories were chosen to be those at the lowest level of inclusion for which the average undergraduate student could be expected to readily differentiate several exemplars." See Table 3.1 for examples.

Table 3.1: Sample stimuli.

Stimulus (**expected/within/between**)
He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of **football/baseball/monopoly**.
The day before the wedding, the kitchen was just covered with frosting. Annette's sister was responsible for making the **cake/cookies/toast**.
He complained that after she kissed him, he couldn't get the red color off his face. He finally just asked her to stop wearing that **lipstick/mascara/earring**.

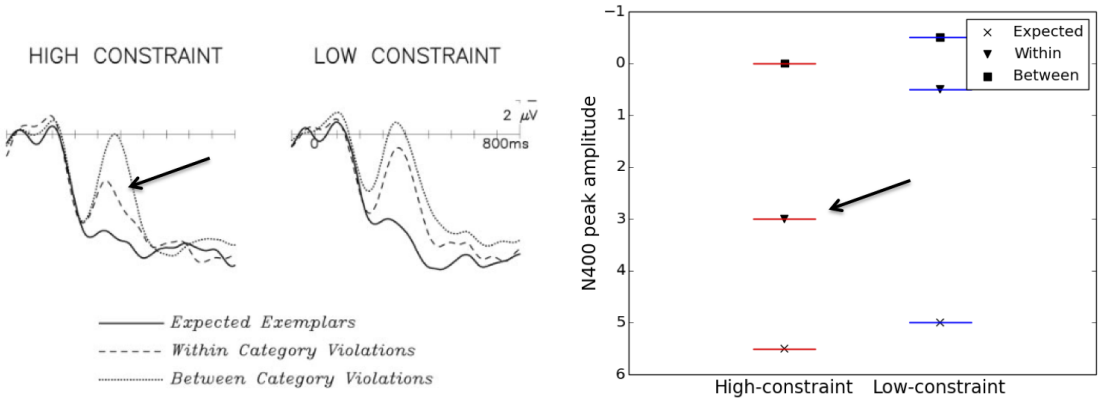

Figure 3.1: Federmeier and Kutas (1999) N400 results. Left: original results as reported by the authors. Right: results re-plotted as points representing peak N400 amplitude, for greater ease of comparison to simulation results below. Arrows indicate key facilitation in high-constraint within-category condition.

plitude for the two unexpected target types, regardless of category relationship to the expected target.

The stimuli are furthermore binned into two conditions based on the extent to which the context constrains toward the expected word: stimuli were classified as either "high-constraint" or "low-constraint", according to a median split on cloze probability of the expected target.

Figure 3.1 shows the results of Federmeier and Kutas's study. Negative voltages are plotted upward[2], with higher N400 amplitude (corresponding to a word that is less ex-

[2]This is conventional in the N400 literature.

pected or facilitated) represented by a greater negativity. In both constraint conditions, we see that the expected target has extremely low N400 amplitude, compatible with strong facilitation. Additionally, in both constraint conditions, between-category targets show very high N400 amplitude, compatible with lack of facilitation. There is also a main effect of constraint level, but the key difference emerges for within-category targets: in high-constraint contexts only, within-category targets show reduced N400 amplitude—despite the fact that within-category targets (like between-category targets) have roughly zero cloze probability. Federmeier and Kutas interpret this result as evidence that within-category targets are facilitated due to semantic overlap with features of the expected target, which are pre-activated in high-constraint contexts.

Federmeier and Kutas' account assumes feature-based representations of words, consistent with a use of distributed representations, and describes facilitation in a manner consistent with an overlap-based account—by their account, the context generates expectations about an unseen word, causing features of that word to be pre-activated, and overlap of those pre-activated features with features of the incoming within-category word is the source of the intermediate facilitation of that word. Their account does not make explicit commitments with respect to the combinatorial dimension, but their implication is that the pre-activations of the features of the expected word arise as a result of a full interpretation of the message conveyed by the stimuli, rather than an asyntactic lexically-based representation of some kind.

What about asyntactic lexical mechanisms? Could they be playing a role in this result? In the examples described in Chapter 1, our evidence for the existence of such mechanisms comes in the form of deviations from the predictions of cloze probability—

deviations that give us reason to believe that something less than the full syntactically-composed representation is informing the facilitation reflected in the N400.

In the case of the Federmeier and Kutas result, we again have a deviation from the predictions of cloze probability, and as in the above cases, it amounts to unexpected facilitation of continuations that do not fit the context. Federmeier and Kutas dismiss the possibility of a lexical-associative account (which they assume would consist of a context word priming the expected target, and the within-category target by extension) based on the fact that they have constructed their stimuli to contain lexical associates only in the first sentence, with the second sentence being equally compatible with all three continuation types. As a result, they reason that the distance is too great for a lexical-associative account to hold. Additionally, they argue that "only about one third of our context (first) sentences actually contained a word lexically associated with the expected ending".

Of course, as we have seen in Chapter 1, we have good reason to posit effects of asyntactic lexical mechanisms over longer distances than they are seen in normal priming, and as I have argued earlier in this chapter, subjective assessment of lexical relation may not capture the full complexities of the relations that affect the N400. So this brings us back to the question of whether lexical relations, when quantified by a computational model like a VSM, could in fact explain the deviation from cloze probability observed in this study.

In the simulations below, this is the hypothesis that we test: that the N400 patterning could be explained by asyntactic lexical effects arising directly from the relation between the context words themselves and the target words, without requiring a prediction based

on a structured representation of the context, and without requiring a mediating influence by the expected target.

3.2.2 Model

For this simulation, we use a VSM framework that has been influential within NLP, the word2vec model (Mikolov et al., 2013). This model learns word representations as weights within a neural network, training these weights so as to best predict words in the surrounding context. The model for this simulation is trained on approximately 2 billion words of semantically diverse web data from the ukWaC corpus (Ferraresi, Zanchetta, Baroni, & Bernardini, 2008), training vectors of 100 dimensions using word2vec's skip-gram architecture, which involves predicting context based on the current word (by contrast to an alternative architecture that involves predicting a target word given surrounding context).

Recall that once we have trained this VSM, each word of the vocabulary is represented as a vector situated within the resulting vector space. For a given sentence context, we will refer to vectors for the expected target, within-category target, and between-category target as vectors E, W, and B, respectively.

In order to model the collective effect of words in the context in an asyntactic manner, we model the preceding context as the output of a simple averaging procedure: vectors for selected context words are averaged to obtain a context vector C. This representation reflects the collective effect of the included words, without any contributing information from the syntactic structure of the sentence, or from unseen (predicted) words.

In forming our context representations, we opt not to include all context words indiscriminately; instead, we aim to operationalize the notion of "informativity" in order to isolate the top several most content-bearing words for inclusion in the average, hypothesizing that the most contentful words will have the strongest influence upon the context representation (e.g., receive the most attention). For the purposes of this preliminary simulation, we try two selection methods, referred to as *anchored* and *agnostic*.

In the anchored setting, we use relation to the expected target as a proxy for informativeness: using the expected target as an anchor, we select the four context words with highest cosine similarity to that expected target.[3] We employ a minimum cosine similarity of 0.2 (chosen by examination of context word cosine similarities in a small subset of stimuli) to further filter words bearing little relation to the target.

In the agnostic setting, we take the top four words based on negative log frequency (the least frequent words), excluding person names (e.g., *Annette*). This is equivalent to choosing words based on maximum surprisal (information content) as determined by a unigram probability model.

The modeling results suggest *prima facie* that the anchored setting is more successful in isolating the most significant words of the context. If so, this would likely be due to the fact that the frequency metric underlying the agnostic setting, while reasonable, is a rather blunt tool for assessing informativeness. That said, as will be discussed in more detail below, there are many reasons to withhold judgment with respect to adjudication

[3]One target. *polar bear*, is made up of two words; this is represented as the average of the two separate word vectors.

between the predictions generated in these settings (and as we will also discuss below, the anchored setting has its own disadvantages with respect to our modeling goals).

Within these word selection settings, we test two types of average: unweighted, and weighted inversely by linear distance. The latter average aims to instantiate the hypothesis that lexical effects, despite being carried forward longer than conventional priming effects, may still decay in strength over time, with earlier words having less influence than later words.

As described in Chapter 2, we use cosine similarity as our implementation of an overlap-based facilitation mechanism. The computation of cosine similarity (the cosine of the angle) between two vectors v and w with n dimensions is as follows

$$\frac{\sum_{i=1}^{n} v_i w_i}{\sqrt{\sum_{i=1}^{n} v_i^2}\sqrt{\sum_{i=1}^{n} w_i^2}} \tag{3.1}$$

where v_i and w_i represent the i-th dimensions of v and w respectively. The denominator of this formula is simply the product of the two vector lengths, so what this computation amounts to conceptually is a (length-normalized) dimension-by-dimension multiplication of the corresponding vector elements—which means, roughly, that we return a high cosine value to the extent that the two vectors have similar values on their various corresponding dimensions. For this reason, I consider cosine to be an apt implementation of the overlap-based facilitation mechanism: the greater the cosine similarity between two word representations, the greater the overlap, and the greater the expected facilitation. One can

easily imagine implementing and testing other feature-matching algorithms as alternative hypotheses.

For every stimulus, we take the cosine similarity between the context vector C and each of E, W, and B, and we average these cosine similarity values across stimuli within each condition, in order to simulate average N400 amplitude.

It is important to note that by assessing overlap directly between the context vector and the incoming word vector, we are testing the hypothesis that overlap of the incoming word with features of the *context words themselves* could possibly produce the pattern of facilitation observed in this experiment. So although we mirror Federmeier and Kutas in testing an overlap-based account of the facilitation, our model differs from that described by Federmeier and Kutas in that our overlap is not based upon activation of features of a not-yet-encountered, expected word, as they propose. Additionally, although Federmeier and Kutas do not explicitly specify the combinatorial mechanisms that they envision, we suspect that our averaging-based operation contrasts with the more sophisticated compositional operations that they are likely assuming.

As a simple control, we also compute cosine similarity between E and W and between E and B. This allows us to assess the model's representation of the relations between different completion words.

Federmeier and Kutas make available a sample of 40 of their experimental stimuli; we run our simulation on that sample.

3.2.3 Simulation results

Figure 3.2 shows the results of the comparison between target types E, W, and B—this test simply serves as a control, to compare the model's relation computations against those assumed by Federmeier and Kutas, and to check for confounds. In Figure 3.2 and those that follow, cosine similarity is plotted on the y-axis with the negative direction upward, to facilitate comparison to N400 plots in Figure 3.1: higher cosine similarity predicts lower N400 amplitude. Note in Figure 3.2 that the expected word vector E is at cosine similarity of 1, as this is a comparison of a vector to itself. As for the other two comparisons, we see that the model predicts on average a nearly identical level of relation between expected words and within-category words in both constraint conditions. We see a slightly greater distance between the expected word E and the between-category word B in the high- than the low-constraint condition. In both cases the model's relations are roughly consistent with the categorical relations assumed by the experimental manipulation: within-category items are represented as being (slightly) closer to the expected targets than are the between-category items. The lack of any discernible difference in the expected/within-category target relation between constraint conditions also rules out—within this simulation—the possible confound of differing relation strengths between the targets themselves.

Figure 3.3 shows the full simulations under the anchored and agnostic settings, respectively. (The right-hand side of Figure 3.1 presents Federmeier and Kutas's results in the same plotting format, for ease of comparison.) In these figures we see several things. First, we see a main effect of constraint consistently captured across settings: for each

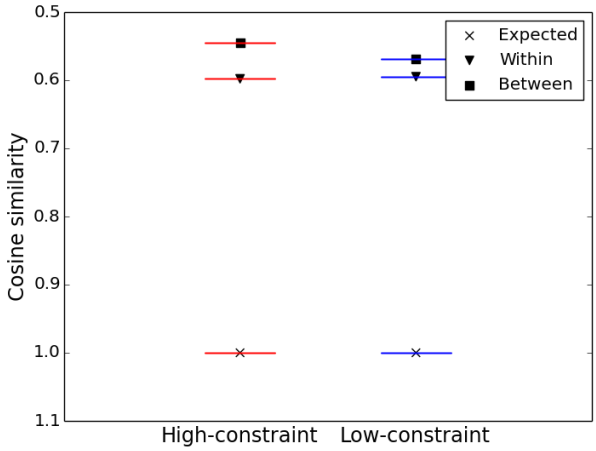

Figure 3.2: Cosine similarity to expected target

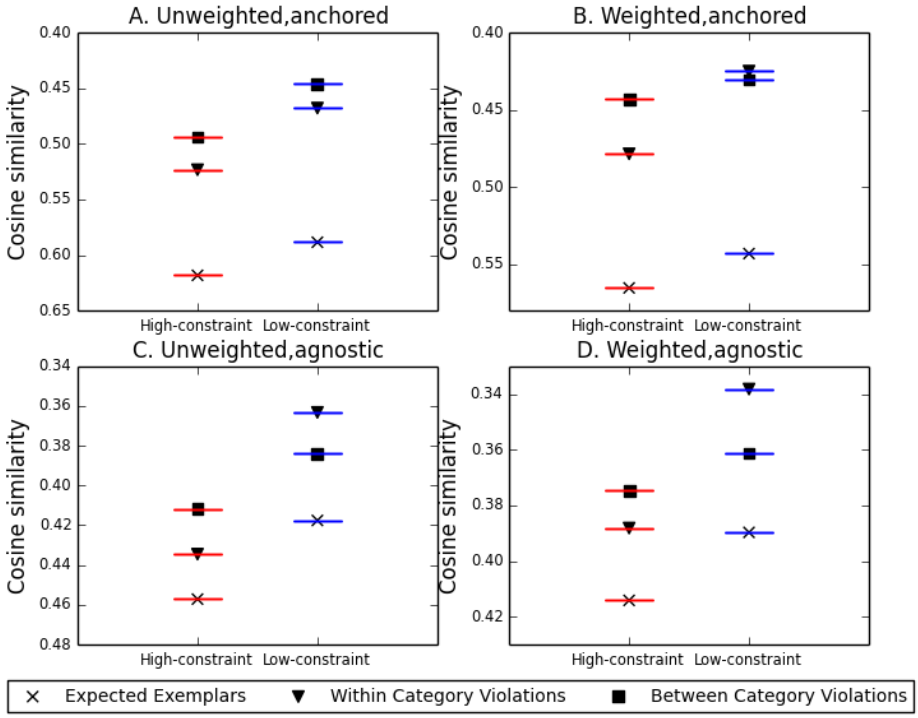

Figure 3.3: Simulations in four settings. A) Context average unweighted by linear distance and words selected with expected target as anchor. B) Context average weighted by linear distance and words selected with expected target as anchor. C) Context average unweighted by linear distance and words selected by low frequency. D) Context average weighted by linear distance and words selected by low frequency.

ending type, average cosine similarity to context is higher in the high-constraint condition, corresponding to greater facilitation (lower N400 amplitude). This is consistent with the main effect observed in Federmeier and Kutas's N400 results.[4]

In addition, we see that for the most part, looking independently at the high- and low-constraint conditions, the three ending types pattern as the experimental paradigm predicts: expected targets are most facilitated by the context, while within- and between-category targets are less facilitated. We also see that under all settings, in the high-constraint condition the within-category target falls at an intermediate position between the other two target types. In the low-constraint condition, however, three of the four settings have within- and between-category conditions in reversed or roughly identical positions. The fact that between-category targets in the low-constraint condition fail to fall farthest from the context, often switching with within-category targets, could in theory reflect similar factors to those that lead to within- and between-category conditions having statistically indistinguishable N400 amplitudes in Federmeier and Kutas' results.

We see in Figure 3.3 that these models—under both anchored and agnostic word selection settings—do predict greater facilitation of within-category targets in the high-constraint as compared to the low-constraint condition, suggesting that direct overlap with words of the context could offer a valid explanation for the increased facilitation of within-category items observed in the high-constraint condition.

[4]Having access only to 40 items of the original Federmeier and Kutas study, we are not making claims of statistical significance for this pattern of results in the models.

3.2.4 Discussion

In this simulation we find that using a distribution-based VSM, a simple averaging-based combinatorial mechanism with heuristics for selecting contentful words, and cosine similarity as a measure of direct overlap between context words and the incoming word, we are able to simulate key aspects of Federmeier and Kutas's N400 results: the basic patterning of item types within constraint conditions, as well as the main effect of constraint. Our model accounts for the deviation from predictions of cloze probability in the high-constraint condition, suggesting an alternative possible account for this facilitation, in terms of asyntactic lexical mechanisms.

At face value, if we assume a linear relation between cosine similarity and N400 amplitude, then Figure 3.3B shows the most faithful simulation of the observed Federmeier and Kutas results. We could in theory take this as evidence supporting a cognitive model in which the N400 amplitude in this study is driven by lexical relations from informative words (with relation to expected target being a better proxy for informativeness), with a given word's influence decaying over time.

However, the drawing of this type of conclusion is not intended as an outcome of this demonstration, for a couple of reasons. First, we are modeling only six datapoints (based on averaging with only 40 items), without claims of statistical significance. This is simply not enough data to adjudicate in a fine-grained manner between different models within the hypothesis space.

Second, we are for the moment assuming a linear relation between cosine similarity and facilitation observed in N400 amplitude, but this is likely an oversimplification.

Consider, for instance, ceiling and floor effects, which are understood to influence N400 amplitude. Floor effects, at very least, are likely a factor in Federmeier and Kutas's results, given that the study finds no significant effect of constraint on N400 to expected targets—despite the fact that high- and low-constraint contexts are defined precisely by how predictive they are of the expected target. The fact that our cosine similarity measure does reflect an effect of constraint on expected targets suggests that we are capturing important aspects of the context-to-target relation with this measure. However, it also suggests that we may need some kind of nonlinear linking hypothesis to predict the N400 with more precision. This means that we should not be quick to dismiss the three other settings in Figure 3.3, as they could ultimately prove to be the more accurate simulations once we identify the proper linking hypothesis.

An important question that arises is the extent to which our model could in fact be tapping into the same factors described by Federmeier and Kutas' account: what if the vectors of the context words, when averaged, arrive at a representation that overlaps significantly with that of the expected target, such that the priming from the context vector is essentially approximating the overlap from the expected target? This concern is greatest in the anchored setting, when the context words are selected based on similarity to the expected target.

Two things argue against this option. First, Figure 3.2 shows the relations of the target words to the expected targets themselves, and this pattern is quite different from those seen in Figure 3.3. Second, even when context words are selected based on inverse frequency, the major components of the pattern—main effect of constraint, greater facilitation of within-category high-constraint items than within-category low-constraint

items—are retained. So while we cannot fully rule out the possibility that there is more similarity between the accounts than meets the eye, the key point to make here is that when we quantify lexical-associative relations based on a VSM in this manner, the assumption that such relations cannot account for the effects seen in this study is called into question, and we see good reason to take our asyntactic lexical mechanisms seriously as influences not just in the cases where they are obvious, but also more generally.

A final comment: as we will see repeatedly in this dissertation, there are bound to be many possible explanations for a given set of observed results. Our simulation does not rule out Federmeier and Kutas' interpretation, and of course, we do not intend to suggest that asyntactic lexical processes could explain all of human comprehension. What this simulation does do is offer an alternative account of this particular result, which widens the space of possibilities left in play by the observed N400 data. Again, the restriction to six datapoints is a major limitation on the strength with which we can draw conclusions for these particular results. To take full advantage of this method, we will need to model larger sets of data in order to adjudicate with greater confidence and precision between differing hypotheses. We turn now to modeling of larger datasets in the next section.

3.3 Modeling semantic priming with VSMs

Above we have demonstrated how, using a distribution-based VSM for representing words, we can make a case for influences of asyntactic lexical effects even in circumstances where those relations are not subjectively obvious.

In this section we aim to look closer at the correspondence between VSM-based lexical relations and lexical-level facilitation effects by testing VSMs' success in simulating single-word semantic priming. When we study lexical effects in sentence contexts, there are many interacting factors at play, and it is difficult to be sure what contributions are made by lexical processes *per se* (a problem that we will tackle in Chapter 5). As we have discussed above, there is an extent to which we expect the nature of lexical effects in sentence contexts to be different from normal semantic priming, as evidenced by the fact that they appear to be longer-lasting. However, we still stand to gain some insight by filtering out the sentence context and examining the performance of VSMs in single-word contexts. So, bearing in mind the important distinctions between these contexts and sentence contexts (and distinctions between the N400 and behavioral measures), we will now examine how well the relation strengths indexed by different VSMs accord with the relations suggested by semantic priming magnitudes in single-word contexts.

3.3.1 Single-word contexts: semantic priming

Semantic priming refers to the phenomenon in which, when performing a language task such as lexical decision (deciding whether a string of letters is a word or not) or naming (naming a word aloud), language comprehenders show speeded performance if the word to which they are responding is preceded by a semantically related word (McNamara, 2005; Meyer & Schvaneveldt, 1971). In the terms of our discussion above: semantic priming represents facilitation by a one-word context. What semantic priming provides

us with, then, is a measure of facilitation between individual words, without involvement of the combinatorial dimension.

For word representations in these simulations, we used two different VSM frameworks popular in NLP: word2vec (Mikolov et al., 2013), the model used above, as well as GloVe (Pennington, Socher, & Manning, 2014). Both of these models make use of distributional information to train their vectors. By comparison to word2vec (which, as described above, trains word vectors as weights within a neural network optimized to predict surrounding words) the training of a GloVe VSM involves learning vectors based on global co-occurrence statistics from a corpus, optimizing vectors with a weighted least-squares model inspired by the intuition that word meanings can be usefully inferred from ratios of co-occurrence probabilities, to distinguish relevant from irrelevant words.

For each of these model types we trained two different models differing based on the size of the relevant context window around a word—using window sizes of 5 and 15 words.[5] The idea behind using different context window sizes is that this parameter will in theory influence the nature of the similarity relations reflected in the VSM. Using a very large window is likely to produce more topically-guided relations, since words will have similar distributions as long as they occur in the general vicinity of similar other words. A smaller window is likely to encourage more syntactically-mediated relations, since there will be a more restricted definition of what it means to be similarly distributed.

We may reasonably consider these four VSMs to occupy close but non-identical positions on the word representation dimension of our hypothesis space. Since the interpreta-

[5]We again used the skip-gram architecture of word2vec. All models were trained on a concatenation of English Wikipedia and English GigaWord using their default parameters.

tion of the individual dimensions in these vector representations is more or less opaque, the hypotheses being tested with these models are more oriented toward the information that goes into the learning of these representations, and the mechanisms by which that information is processed.

Although these mechanisms and information sources do differ between the four models, in light of the full extent of possibilities for the structure of word representations, these four models fall comparatively close to one another within our hypothesis space. Testing similar models in this way, then, is less a means of adjudicating between significantly differing hypotheses, and more a means of testing the stability of our results between comparatively similar models. The testing of substantially different word representation hypotheses is left to future work.

Several previous studies have already shown correspondence with semantic priming using distribution-based VSM representations and a variety of facilitation computations (Herdağdelen, Erk, & Baroni, 2009; M. N. Jones, Kintsch, & Mewhort, 2006; Lapesa & Evert, 2013; Mandera, Keuleers, & Brysbaert, 2016; McDonald & Brew, 2004; Padó & Lapata, 2007). This suggests at the very least that VSM word representations based on distributional information are to some extent able to capture what is reflected in the neural word representations involved in priming. As in our current simulation, most of these studies limit themselves to fairly similar distribution-based VSMs, rather than testing significantly diverging cognitive hypotheses.

While many of these previous studies utilize fairly small priming datasets and primarily test whether a priming effect can be captured at all (that is, whether facilitation by related versus unrelated primes can be differentiated), Mandera et al. (2016) take advan-

tage of the large online database of the Semantic Priming Project (SPP), which compiles priming data from 768 subjects for over 6000 word pairs (Hutchison et al., 2013), and uses the size of this database to enable a regression-based analysis of the capacity of various distribution-based VSMs to predict degree of facilitation between primes and targets.

Because our ultimate intention is to be able to adjudicate as precisely as possible between the simulation capacities of a large space of different models, a large dataset like the SPP, enabling regression-based prediction of facilitation, is ideal for our purposes (Keller, 2010). We therefore follow Mandera in using the SPP for regression-based assessment of the predictive power of our models. For the purposes of this demonstration, like Mandera we also limit ourselves to readily-available distribution-based VSMs.

Unlike Mandera, we do some exploration with additional finer-grained analyses made possible by the SPP. Specifically, the SPP contains priming results for four experimental conditions: two tasks, lexical decision and naming, crossed with two stimulus onset asynchronies (SOA), 200ms and 1200ms. SOA refers to the amount of time between the onset of the prime word and the onset of the subsequent target word. It is not clear *a priori* whether we should expect facilitation under all of these conditions to be the same—of the same magnitude, or even necessarily produced by the same mechanisms. So it is worthwhile to compare the predictive capacities of our models under the different conditions.

We assessed the predictive power of our four selected models on each of the four priming datasets, by fitting linear regression models to the human response times, with cosine similarity between prime and target as the predictor of interest. Since we are predicting response times rather than priming magnitudes, we use a simple baseline regression model with word frequency as a predictor, and assess the extent to which the VSMs are

able to account for variance in the response times over and above the predictions of frequency. Word frequency is widely recognized as a strong predictor of reaction time in language tasks (Rubenstein, Garfield, & Millikan, 1970), and while it is only one of the factors known to affect the speed of word recognition (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004), it is the most significant, and unlike factors such as word length, it is represented in many vector space models (Schnabel, Labutov, Mimno, & Joachims, 2015), making it very relevant to control for here.

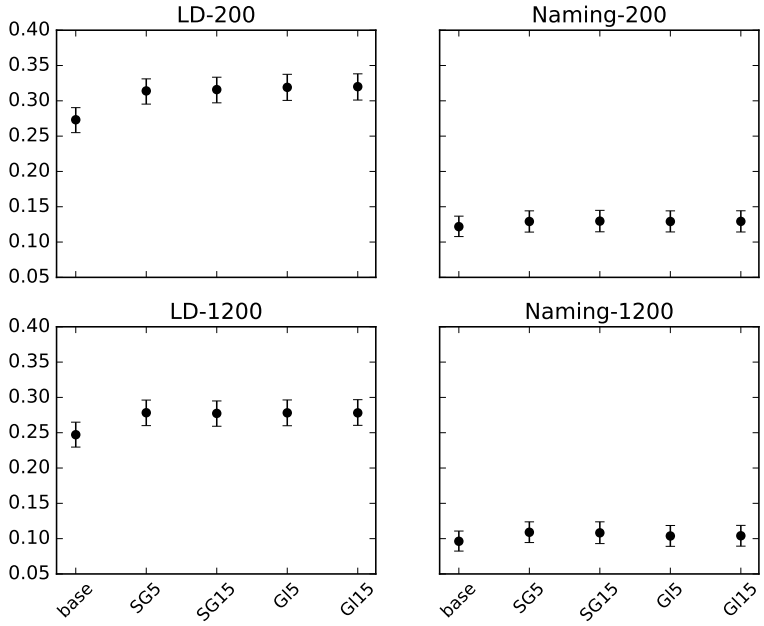

Figure 3.4: r^2 values for linear models fit to priming results in full SPP dataset, under different priming conditions. Baseline model ("base") contains only frequency as a predictor, while other models contain cosine values from the indicated VSMs. Error bars represent bootstrapped 95% confidence intervals.

3.3.2 Results and discussion

Figure 3.4 shows the r^2 values, which quantify the proportion of the variance explained by the regression model, for the baseline model and for models including both frequency and cosine similarity between prime and target, calculated using each of the four VSMs.

A number of things can be seen in these results. First, the four distribution-based VSMs, despite the differences in their training and the sizes of the relevant co-occurrence windows, perform indistinguishably from one other in terms of the amount of priming variance that they capture. This suggests that to the extent that we are to use off-the-shelf distribution-based VSMs for word representation, we don't have a strong reason to choose one over the other—keeping in mind, of course, that the measure being simulated here differs from the N400, and single-word contexts differ from sentence contexts.

Second, we see that the models are indeed capable of accounting for significant priming variance over and above that accounted for by frequency—but this is only the case for priming latencies in the lexical decision task (LD), by contrast to the naming task. Additionally, there is a more substantial margin of improvement over the baseline in the 200ms SOA (recall that this is the time between onsets of the prime and target) than the 1200ms SOA. This suggests that what these distribution-based VSMs do capture is more apt for simulating facilitation effects within the context of a lexical decision task, and when the delay between words is brief.

The significant difference in the capacity of any of these variables, including frequency, to predict response latencies in lexical decision as opposed to naming is interesting—it suggests that the different tasks involve a non-trivial difference in the pipeline of mech-

anisms ultimately giving rise to the facilitation measure. This is not surprising, as we would imagine that lexical decision would involve more substantive processes of lexical access—or at least consultation of lexical knowledge—than would a naming task.[6] Furthermore, it seems that the representations that these VSMs provide have more explanatory power when it comes to the pipeline involved in the lexical decision task, suggesting a legitimate connection between the behaviors of our VSMs and the processes involved in accessing lexical knowledge.

The closer correspondence of the VSM relations to the shorter SOA is also suggestive of an account in which the shorter SOA is more reflective of passive, automatic facilitation processes, while the longer SOA may reflect more strategic processes. This would be consistent with our discussion in Chapter 2, speculating that distribution-based VSMs may in fact be appropriately-suited to modeling more passive, automatic associative relations that arise from experience of co-occurrence statistics. This also accords well with findings, described in Chapter 5, suggesting that asyntactic lexical processes have more influence with less processing time, and less influence with more processing time.

Although our VSMs account for a non-trivial portion of the priming variance, it is also clear that a substantial amount of the variance remains unaccounted for by these VSMs. We can expect this to be caused by a combination of factors: imperfection of our representations relative to neural lexical representations, and additional variables affecting the variance above and beyond even perfectly-represented lexical relations. Since we certainly do not expect to model lexical representations perfectly—even if we are aiming

[6]Even if naming is simply noisier, this would still seem to reflect some kind of mechanistic variation such that patterns of response times end up noisier than in lexical decision.

only to capture the associative components of those representations—we will content ourselves for now with the conclusion that these distribution-based representations do capture a portion of the effects of lexical relations on semantic priming. Furthermore, we are encouraged by the fact that the VSMs simulate most successfully the priming in the 200ms-SOA lexical decision task—this bodes well for our use of VSMs to model processes that we construe as fast, (fairly) automatic, and connected to the access of lexical knowledge.

3.4 Conclusion

In this chapter we have explored the use of VSMs for modeling asyntactic lexical relations in two contexts. First, we used a VSM to simulate the result of Federmeier and Kutas (1999), offering an alternative account of that result in terms of asyntactic lexical effects, and in doing so demonstrating the capacity of these models to capture lexical-associative relations that may not be subjectively apparent.

Second, we tested the use of VSMs for simulating semantic priming from single-word contexts, to further explore the correspondence between relations produced by a VSM, and relations indexed by facilitation measures—in this case response latencies. We tested four distribution-based VSMs, finding that the response time variance accounted for is relatively stable across these four models, and that a non-trivial amount of variance is explained over and above that explained by frequency, with the variance best accounted for in short-SOA lexical decision results. This suggests that these models do capture

aspects of lexical relations to a non-trivial extent, and that they are best suited to modeling fast processes connected to lexical access.

In the next chapter, we will move to exploring the role reversal phenomenon, which presents a useful testing ground for understanding not just asyntactic lexical effects, but the interplay of those effects with syntactically-constrained message-level effects.

Chapter 4

Role reversals: background and probing an existing model

4.1 Introduction

In this chapter we will shift to discussing our understanding of the interaction between the kinds of asyntactic lexical processes that we have been discussing in the previous chapters, and the more sophisticated, syntactically-constrained processes with which we have contrasted them. We use as a case study the role reversal phenomenon, which has raised critical questions about this interaction, and which has inspired numerous accounts and models attempting to address that interaction.

The basic nature of what I refer to as the "role reversal phenomenon" is the observation that the N400 fails to reflect sensitivity to the anomaly of continuations that represent reversals of canonical thematic roles—for instance, the N400 shows facilitation in response to "the restaurant owner forgot which waitress the customer had served", despite the fact

that *served* here is anomalous (because customers don't typically serve waitresses). As we discussed in Chapter 1, this is one of the classic types of result that motivates our exploration of asyntactic lexical effects on the N400.

Why is this a good case study of the interaction of lexical and syntactically-constrained processes? After all, given our assumption of asyntactic lexical processes, the facilitation observed in the N400 in these examples lends itself to a straightforward lexical explanation.

We will see two versions of an answer to this question. In this chapter, when we review the core canon of role reversal studies, we will see sensitivity to syntactic information manifesting in the form of the P600 ERP[1] component, which is a positive-going deflection of the signal peaking around 600 milliseconds after the onset of the word—and which, by contrast to the N400, does show an effect of the role reversal.

In the following chapter, however, we will review newer results that show role reversal sensitivity not just in the P600, but in the N400 itself. This will provide us with an ideal testing ground for modeling the interaction of lexical and syntactic processes, as it allows us the opportunity to test their interaction in the context of a single ERP component.

In this chapter I will set us up for that discussion by first introducing the classic set of role reversal results, which kicked off rethinking of the functional roles of both the N400 and the P600. In Section 4.2 I will review these studies and the resulting theories, and I will discuss the relationship of the theories to the considerations that we have prioritized

[1]Recall that this is the event-related potential technique, introduced in Chapter 1, which is the measure by which we observe the N400.

in the previous chapters. I will also introduce two computational models that have claimed to account for these results.

In Section 4.4 I will then zero in on one of these computational models, which claims to account for both the N400 and P600 within a single-stream neural network model. As is often the case with such models, the precise mechanisms by which the network manages to capture these ERP patterns are less than transparent. To address this, we replicate and probe the model to better understand what is driving its apparent success. I report the results of this replication and analysis, and discuss the implications.

This will set the scene for our modeling of the more complex picture of the N400 in Chapter 5.

4.2 Background: the role reversal phenomenon

This brings us to the phenomenon of role reversals. These fall within a category sometimes referred to as "semantic illusions", in which a semantically anomalous target shows no N400 effect, but instead shows an effect in the P600.

Hoeks et al. (2004) present one of the first influential examples of this type of result. Setting out specifically to clarify the interaction between message-level and lexico-syntactic effects on the N400, they construct stimuli in Dutch as shown in Table 4.1, crossing cloze probability of target words ("message strength") with association strength between the preceding lexical content and the target ("fit"). Their message strength manipulation makes use of a role reversal contrast, holding constant the nouns that serve as arguments to the verb, but reversing the roles that these nouns fill. As we see in Ta-

ble 4.1, three of the conditions in this experiment result in anomalous sentences, while the good-fit strong-message condition results in non-anomalous sentences.

The results of this study show an N400 contrast, as expected, for both of the "poor fit" anomalous conditions, by comparison to the non-anomalous sentence condition. However, the third anomalous condition—that with a good lexical fit between the target and the preceding nouns, shows no N400 contrast, suggesting facilitation of this verb despite its incongruence with the message. This is our classic role reversal pattern: a verb showing N400 facilitation under conditions of lexical association, despite the fact that the thematic roles of the preceding nouns make it an anomalous continuation.

By contrast, the results show P600 effects such that all three anomalous conditions have greater positivity as compared to the plausible condition, suggesting that the P600 reflects sensitivity to the anomaly caused by the reversal of thematic roles.

This set of results will be the focus of the simulations of the model replicated and probed in Section 4.4.

Good fit, strong message De speer werd door de atleten geworpen lit. The javelin was by the athletes thrown	—
Good fit, weak message De speer heeft de atleten geworpen lit. The javelin has the athletes thrown	P600
Poor fit, strong message De speer werd door de atleten opgesomd lit. The javelin was by the athletes summarized	N400 and P600
Poor fit, weak message De speer heeft de atleten opgesomd lit. The javelin has the athletes summarized	N400 and P600

Table 4.1: Hoeks, Stowe, and Doedens (2004) stimulus conditions.

Kuperberg, Sitnikova, Caplan, and Holcomb (2003) also introduce anomalies involving thematic role violations, using English stimuli and comparing three conditions: normal sentences, thematic role animacy violations, and non-thematic role pragmatic violations. Stimulus examples are shown in Table 4.2. The authors find that the pragmatic violation condition (c) shows a significant N400 effect relative to the normal condition (a), but that the thematic role violation (b) does not. Instead, the thematic role violation condition shows a significant P600 effect (which the pragmatic violation lacks).

Normal
(a) For breakfast the boys would only eat ...
Thematic role animacy violation
(b) For breakfast the eggs would only eat ...
Non-thematic role pragmatic violation
(c) For breakfast the boys would only bury ...

Table 4.2: Kuperberg, Sitnikova, Caplan, and Holcomb (2003) stimulus conditions.

Kolk, Chwilla, Van Herten, and Oor (2003), too, find that role reversal anomalies show no N400 effect, but rather a P600 effect. Example stimuli from their study (conducted in Dutch) are shown in Table 4.3.

Normal
De stropers die op de vos joegen slopen door het bos
the poachers who at the fox hunted stalked through the woods
lit. The poachers who hunted the fox stalked through the woods
Reversal
De vos die op de stropers joeg sloop door het bos
the fox that at the poachers hunted stalked through the woods
lit. The fox that hunted the poachers stalked through the woods

Table 4.3: Kolk, Chwilla, Van Herten, and Oor (2003) stimulus conditions.

A. Kim and Osterhout (2005) also compare conditions involving manipulation of the typical role of the argument and the verb, as shown in Table 4.4. They find that their attraction violation condition (a), which involves lexical association but a reversal of canonical argument roles, produces a P600 effect relative to the control (b), but no N400 effect. The non-attraction violation (c), by contrast, shows an N400 effect rather than a P600 effect.

Attraction violation
(a) The hearty meal was devouring ...
Passive control
(b) The hearty meal was devoured ...
Non-attraction violation
(c) The dusty tabletops were devouring ...

Table 4.4: A. Kim and Osterhout (2005) stimulus conditions.

Finally, Chow et al. (2015) test subjects on sentences as shown in Table 4.5. In line with the results above, they find a P600 effect but no N400 effect for the role reversal contrast, while they find both N400 and P600 effects for their argument substitution contrast.

Role reversal
The restaurant owner forgot which customer the waitress had served ...
The restaurant owner forgot which waitress the customer had served ...
Argument substitution
The superintendent overheard which tenant the landlord had evicted ...
The superintendent overheard which realtor the landlord had evicted ...

Table 4.5: Chow, Smith, Lau, and Phillips (2015) stimulus conditions.

4.2.1 Existing accounts

Because the N400 had historically been associated with semantic anomaly and the P600 with syntactic anomaly, these results spurred a flurry of theories attempting to recon-

cile the observed patterns with previous conceptions of the mechanisms underlying these components.

Hoeks et al. (2004) propose that these results arise because at times it is not possible for the processor to arrive at a completely specified message representation—and as a result, an underspecified message representation drives the N400, leading to a temporary "semantic illusion". As for the P600, they propose that this effect reflects more effortful syntactic reanalysis that is subsequently needed in order to achieve a coherent message representation.

Kuperberg et al. (2003) propose two potential accounts. On one hand, they propose that their result could be explained by a model in which the N400 reflects integration difficulty—and when a thematic anomaly is established, no attempt at integration is made, so no integration difficulty is generated. As an alternative, they speculate that this N400 result could simply be modulated by the semantic relationship between the verb and the subject. For the P600, they propose that this effect is caused by the thematic role violation requiring syntactic restructuring to repair the interpretation.

Kuperberg (2007) further elaborates with a theory proposing that three streams of computation operate in parallel: a stream concerned with lexical-semantic memory-based relations, a combinatorial stream concerned with morphosyntactic relationships, and a stream concerned with lexical-thematic relationships. Within this framework, Kuperberg posits that the N400 behavior is sensitive to computations in the lexical-semantic stream—which could be some kind of continued comparison of semantic relationships, or a plausibility heuristic. By contrast, the P600 reflects continued analysis in the combi-

natorial stream, such that a greater P600 amplitude will be generated if different streams result in conflicting outputs.

Kolk et al. (2003) propose a monitoring function for the P600 (checking for veridicality of an unexpected event), and suggest that when a veridicality check is executed, the processor does not attempt to integrate the event, so the N400 is absent. Van Herten, Kolk, and Chwilla (2005) elaborate on this theory, positing that the processor first considers an interpretation that best fits world knowledge, such that there is no integration difficulty and no N400. This plausibility heuristic is independent of and parallel to the syntactic analysis. When the syntactic analysis results in a different interpretation, the conflict between the two interpretations leads to the P600.

A. Kim and Osterhout (2005) propose a semantic attraction account, by which the canonical interpretation (of a meal being devoured) is so tempting that the processor pursues it even in the face of contradictory syntactic information. This results in syntactic processing difficulty that generates the P600. The authors use this as evidence for independent semantic and syntactic processing.

Bornkessel-Schlesewsky and Schlesewsky (2008) propose an account based on their extended Argument Dependency Model (eADM). In this framework, thematic interpretations are made based on "prominence" ranking for arguments, along with a "linking" computation based on lexical requirements from the verb. This operates separately from and in parallel to the processing of plausibility information, after which there is a "generalized mapping" phase that links the two sources of information. The plausibility step can be blocked by the core argument computations (prominence and linking), and the

generalized mapping step can be blocked by a problem in either of the two other steps that it links.

Within this framework, the absence of the N400 is due to semantic association between the arguments and the verb, leading to no increased plausibility processing. The P600 occurs due to a mismatch of role assignments coming from the linking and plausibility steps.

Kos, Vosse, Van Den Brink, and Hagoort (2010) argue for a processing competition account, which assumes that semantics and syntax proceed interactively. Within this framework, syntax and semantics are independent systems which are nonetheless considered concurrently and can have mutual influence. In particular, stronger cues at one level can increase processing cost in the other level. They account for the presence of a P600 rather than an N400 in the above cases by the fact that there is a plausible interpretation that arises based on the semantic stream (the combination of the content words) which therefore places the processing burden on the syntactic steam.

Chow et al. (2015) hypothesize a "bag of arguments" account, such that the processor quickly picks out the arguments of a verb and consults event knowledge based on the identities of these arguments—but it cannot immediately incorporate role information into predictions. In a subsequent reply paper, Chow, Momma, Smith, Lau, and Phillips (2016) suggest a number of possibilities for the slow effect of role information, speculating among other things that that there may be an in-principle constraint against directly querying event memory with argument+role cues, as role-tagged entities may not be encoded in event memory.

Kuperberg (2016)

In a reply to the Chow et al. (2015) proposal, Kuperberg (2016) further elaborates an account of role reversals in which different types of cues influence the N400 in different circumstances. I give particular attention to this account because, as we will see in the next chapter, it can be adapted fairly straightforwardly to serve as a potential account for newer results that complicate the role reversal picture, and consequently it will serve as a foundation for some of the modeling to be described in that chapter.

Kuperberg bases this account on a framework in which the goal is to infer the cause of the inputs—the event or message being described. Within this framework, Kuperberg explains the pattern of effects observed by Chow et al. (2015) in terms of reliability of evidence for a given hypothesis. In the case of the N400 effect observed for the argument substitution contrast, Kuperberg argues that in these sentences ("the superintendent overheard which [tenant/realtor] the landlord had evicted"), the comprehender has various hypotheses about the event described, and these hypotheses are based on all information encountered in the context. This information provides reliable evidence to support beliefs about the likely roles that these entities play, and as a result the sentences lead to a strong hypothesis of an <evict> event when the object is *tenant*, and a much weaker hypothesis of an <evict> event when the object is *realtor*.

As for the lack of N400 effect in the role reversal contrast, Kuperberg argues that in this case the comprehender has inferred a <serve> event with high probability not just when *waitress* is the subject, but also when it is the object. When *waitress* is the subject (making the *serve* continuation good), Kuperberg posits that this inference is once again

made based on all information in the context. However, in the case of "... which waitress the customer had ..."', (when the *serve* continuation is anomalous), Kuperberg claims that the inference is made based only on a subset of cues: "the combination of 'waitress' and 'customer', in that linear order, and following a clause that established a restaurant schema". She claims that the subset of cues is used in this case because "this subset of cues offered the comprehender more reliable evidence to support her hypothesis that the event ... was, <waitress served customers>, than the full set contextual cues provided for any alternative."

It is not immediately clear why this one condition should be singled out to use only a subset of cues in making an inference, when all other conditions use the full set of cues—but we will return to this question when we run a simulation building on this account in Chapter 5.

Lexical effects in these theories

Each of these theories incorporates a component aimed at explaining N400 facilitation of the role-reversed sentences. How do these different mechanisms fit with our discussion of asyntactic lexical effects?

A majority of these proposals (Bornkessel-Schlesewsky & Schlesewsky, 2008; Hoeks et al., 2004; A. Kim & Osterhout, 2005; Kos et al., 2010; Kuperberg, 2007; Van Herten et al., 2005) involve some kind of processing stream or mechanism that produces a message interpretation driven by lexical content. This is sometimes referred to as a semantic stream, or a plausibility heuristic (or in the case of Hoeks et al., a failure to arrive at a fully specified message representation). Underlying all of these approaches is the assumption

that the N400 behavior is explained not by lexical relations alone, but by the use of lexical content to produce a non-syntactically-constrained interpretation. This aligns with our "active" version of the asyntactic lexical mechanism described in Chapter 1.

The account of Kuperberg (2016) also falls in this category, in the sense that it assumes N400 facilitation to arise from an active inference framework. It is worth noting, however, that Kuperberg's formulation of this account attributes the role reversal facilitation to a set of cues including not just lexical identities, but also the linear order of words, and the schema within which they occur. We will deviate from this assumption in our modeling in the next chapter.

Fewer of the accounts propose that the result could be based on simple word-to-word relations (our "passive" version of the lexical effects). Among the theories above, it is perhaps only Kuperberg et al. (2003) and Kuperberg (2007) who explicitly acknowledge the possibility that the observed patterns could be attributable directly to lexical relations. Other accounts acknowledge a role of lexical association, but do so within the broader context of a processing stream that produces an active interpretation.

The Chow et al. (2015) account occupies somewhat of an intermediate position between these categories. On one hand, it proposes that verb candidates are activated based on their association with the nouns that serve as arguments, and that the activation of those associated verbs results in the N400 facilitation. This has the flavor of a spreading activation account, and might be conceived of as operating within a network architecture as visualized in Figure 4.1.

On the other hand, the Chow et al. (2015) account is also framed as an active querying of event memory, suggesting that it is conceived of as an active rather than a passive pro-

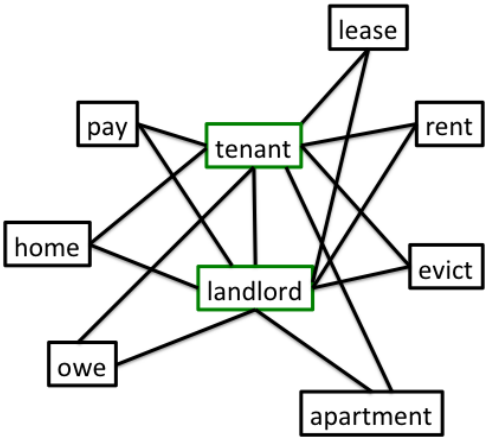

Figure 4.1: Visualization of Chow, Smith, Lau, and Phillips (2015) account within semantic network framework. This visualization is our own interpretation, and not that of those authors.

cess. Additionally, another distinguishing characteristic of the Chow et al. (2015) account is that it cannot be classified as asyntactic, given that it assumes a parse to be available to identify the arguments of the upcoming verb. Although we established in Chapter 1 that the "passive" account may reasonably have non-zero syntactic influence, at least in potentially incorporating sensitivity to the sentential nature of the context, the Chow et al. (2015) account goes further in that it assumes only role information to be omitted from the computation, with other syntactic information retained. This has the flavor of an active interpretation account in which select information is simply omitted.

Falling outside either of these categories are two of the older theories, which suggest that the lack of N400 effect is due to the processor not attempting integration at all (Kolk et al., 2003; Kuperberg et al., 2003). These are more difficult to square with our concept of asyntactic lexical processes, given that they rely on the absence of a process rather than the presence of one. However, because in both cases the relevant authors have later moved on to alternative theories that better align with our assumptions here—and because these

no-integration theories are unable to account for the role reversal sensitivities that we will review in the next chapter—we will not dwell on them.

4.2.2 Computational models

We will now move to discussing existing computational models that have been proposed as accounts of the role reversal phenomenon. Computational modeling of the role reversal phenomenon has primarily been implemented within the connectionist framework, introduced in Chapter 2, which makes use of neural networks to simulate processing.

There are two primary connectionist models that have simulated these results. I will introduce each of these models in some detail, and then I will discuss relevant considerations in determining what to take away from their performance.

Rabovsky et al. (2017)

The first model, presented by Rabovsky et al. (2016), aims to capture N400 behavior, testing the hypothesis that the N400 component reflects "semantic surprise". This model successfully simulates role reversal results of the kind found by Kuperberg et al. (2003), in that it produces only a slight increase in the simulated N400 for sentences such as "For breakfast, the eggs eat ..." as compared to "For breakfast, the boys eat ...", while the increase for sentences such as "For breakfast, the boys plant ..." is larger.[2]

[2] As we have mentioned in Chapter 3, Rabovsky et al. (2016) also use this model to simulate the Federmeier and Kutas result. They do this by training the model such that for each semantic category that they have created, one member of that category is never seen in the same contexts as other category members, in order to create a set of unexpected items sharing a category with expected items. With this training regimen, they find that they are able to simulate the gradual and significant increase in N400 amplitude between the different categories.

The Rabovsky et al. (2016) model is based on a connectionist model known as the Sentence Gestalt model (McClelland, St. John, & Taraban, 1989), illustrated in Figure 4.2. This model first takes as input localist word representations, performing two layers of computations which lead to the "Sentence Gestalt" (SG) layer—this layer is intended to serve as a representation of the sentence. The network is a recurrent neural network (RNN) as described in Chapter 2, in the sense that the representation from the SG layer at each timestep feeds back into the second layer of computation at the next timestep, to preserve a "memory" of the already-processed parts of the sentence.

In the second part of the model, the output of the SG layer, along with a probe representation, are fed into two further layers of computation, leading to an output prediction. The output prediction is trained to reflect an aspect of the meaning of the unfolding sentence, given the input probe (which can consist of a role or a filler).[3] For instance, for a sentence "the cate ate the mouse" and a role probe of "agent", the correct output would correspond to "cat". For a sentence "the dog chased the cat" and a filler probe of "dog", the correct output would correspond to "agent". All possible probes are presented to the network at all timesteps, thus training the network to predict sentence meaning before it has seen the full sentence.

Notably, for the fillers, Rabovsky et al. (2016) use hand-crafted feature-based semantic representations designed such that members of semantic categories share semantic features in a hierarchical manner.

[3]Role probes have localist representations, while fillers have manually-designed feature-based representations.

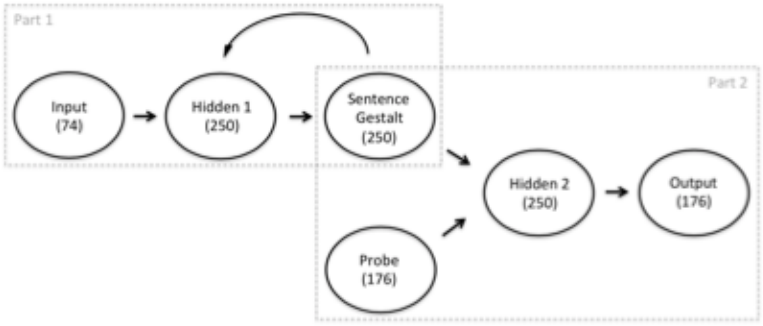

Figure 4.2: Sentence Gestalt model architecture (credit: Rabovsky et al. (2016))

The model is trained on synthetically-generated sentences, with the correct output responses serving as training signal. The authors train the model within a data environment that is generated to reflect relevant probabilistic properties. Training data consist of sentences such as "At breakfast, the man eats eggs", paired with the corresponding event information. The training sentences are generated online based on pre-determined probabilistic constraints.

For simulating N400 amplitude, the authors take the change in the SG layer of the model from one word to the next, using a cross entropy measure. This is intended to represent the change in the inferred sentence representation caused by the arrival of the target word.

Brouwer et al. (2017)

Brouwer, Crocker, Venhuizen, and Hoeks (2017) present another connectionist model, this time a model of both the N400 and the P600. The model is designed to instantiate the Retrieval-Integration account of these components (Brouwer, Fitz, & Hoeks, 2012), which proposes that the N400 reflects retrieval of lexical-semantic information, given

the input word and the preceding context, and the P600 reflects integration of the word meaning into the utterance representation. The model successfully simulates the N400 and P600 patterns observed by Hoeks et al. (2004), which we have described above.

The Brouwer model consists of a connectionist network with two different modules — corresponding to Retrieval and Integration — as illustrated in Figure 4.3. Input at the first layer of the model is a localist word representation, and output at the final layer is a thematic-role-assignment representation: a 300-dimensional vector consisting of three 100-dimensional slots filled by the word vector representations of the agent, action, and patient, respectively. As with the model above, this model is a recurrent network, with the output of the "integration" layer feeding back after each timestep into both the "retrieval" layer and the "integration" layer for the next timestep. Also like the model above, the model is trained to predict the meaning components of the full sentence after every word, thus training it to make predictions of sentence meaning as the sentence unfolds.

Like Rabovsky, the authors of this model also train on synthetically-generated sentences with controlled probabilistic properties. I will go into greater detail about this training environment in the following section, which will focus on better understanding the functioning of this model.

Using this synthetic data, the authors train the network in an unusual regimen: they train the integration module first, using as input pre-trained word vector representations, in the form of 100-dimensional binary representations (COALS: Rohde, Gonnerman, & Plaut, 2006). They then freeze the weights of the Integration module, add the Retrieval module to complete the network, and retrain the full network, now with localist word

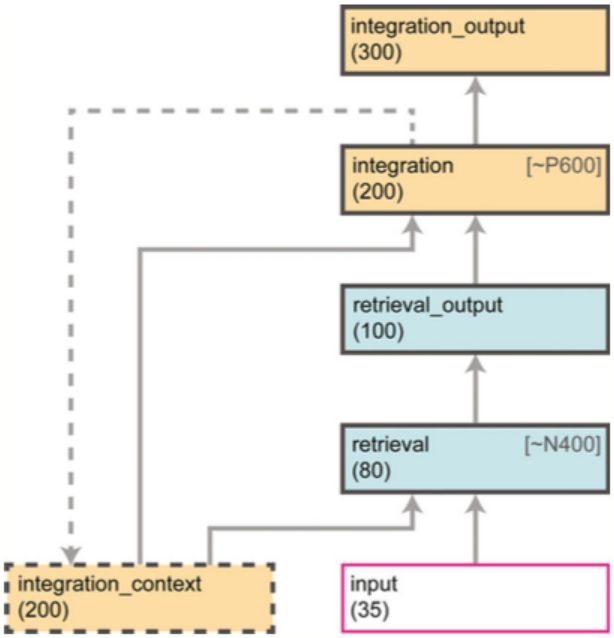

Figure 4.3: Brouwer model architecture (credit: Brouwer et al. (2017))

representations as input, and with training updates being applied only within the Retrieval module.

The intention of this training approach is to guide the model such that the retrieval_output layer is encouraged to produce outputs that are similar to the pre-trained word representations—resulting in the retrieval module being constrained to serve the intended retrieval-like function.

For the N400 amplitude simulation, the authors use the change in the activity pattern of the "retrieval" layer between one word and the next, and for the P600 amplitude simulation they use the change in the activity pattern of the "integration" layer between one word and the next. To quantify the change in these layers, they take the cosine of the activation vectors.

Interpreting these models

What should we take away from these models' results? To what extent can we interpret the Rabovsky et al. simulation as evidence for a semantic surprise hypothesis (as those authors claim)? Or the Brouwer et al. simulation as evidence in favor of the Retrieval-Integration hypothesis (as those authors claim)? We will examine the Brouwer model and the robustness of its performance in more detail in the next section, but first let us consider a comparison of these models, and what we can infer about their claims.

There are a number of notable similarities between these models. They both employ meaning-based training objectives, with each model trained to make predictions of the full sentence meaning at every word. Both are recurrent networks that feed the activations of one of the internal layers back as inputs at the next time step. They also both take N400 amplitude (and, in the case of Brouwer, P600 amplitude) to be represented by the change in activation at an internal layer in the network.

As for differences, we can say with confidence that the models differ in the location of the "N400 layer" within their respective networks. However, our ability to compare the models' claims about the N400 breaks down rapidly as we try to determine exactly what those respective N400 layers should be construed to represent. This difficulty arises because neural networks are notoriously opaque with respect to the nature of the internal representations and computations that they settle on during the course of training—a challenge that we will address in Chapter 6. So while we can reason about the types of information that might be useful for the networks to represent at those points in their architectures, given the various mappings involved (for instance, Brouwer's N400 layer

directly precedes the layer that is roughly constrained to resemble word meaning vectors) we cannot be certain what functional roles those layers actually serve, or what information they actually represent.

As we will see in the next section, another challenge for interpreting the claims of these models is that the internal computations and representations that the networks settle on are not determined solely by the nature of the architecture or training objective—which are typically used to instantiate the hypotheses of interest—they are also influenced significantly by the particular properties of the training data and the types of distinctions that the network is forced to make during training. This further complicates our ability to map confidently between the structure of the model and the corresponding cognitive claims.

This brings us to a final point: both of these models rely on the use of synthetic data environments, which have controlled probabilistic properties determined by the modelers. What this means is that the models can only claim to have explanatory power to the extent that the properties of the stimuli in the synthetic training environment are representative of the properties—the *relevant* properties—of the real stimuli. Otherwise, the results are simply demonstrations of how these particular models respond to these particular probabilistic properties. In the next section, we will see that in the case of the Brouwer model, the assumptions of the training environment prove to be both critical to the model's performance, and problematic for drawing the desired conclusions.

In Chapter 5, we aim to address some of these limitations by making use of models with fewer uncontrolled parameters, and by working with real experimental stimuli and non-synthetic quantification of the properties of those stimuli.

4.3 A closer look at the Brouwer et al. (2017) model

The Brouwer model makes an impressive claim: a full computational model of N400 and P600 patterning for the kinds of role reversal anomalies and standard semantic anomalies compared in the classic Hoeks et al. (2004) study. The model also makes a notable mechanistic claim: by contrast to most of theories reviewed above, which appeal to multiple parallel streams of processing in order to account for the N400 and P600 patterns (Bornkessel-Schlesewsky & Schlesewsky, 2008; A. Kim & Osterhout, 2005; Kos et al., 2010; Kuperberg, 2007; Van Herten et al., 2005), Brouwer et al. (2017) offer their model of these components as proof-of-concept for a "single-stream" account, in which the output of the process that gives rise to the N400 is fed directly into the process that gives rise to the P600. Specifically, as described above, the model is intended to instantiate the Retrieval-Integration (RI) theory (Brouwer et al., 2012), which holds that the N400 reflects lexical retrieval, while the P600 reflects integration of lexical content into the sentence meaning representation.

The success of this RI-based computational model in simulating divergent role reversal effects with a single-stream architecture is significant and worth examining in greater detail—particularly in light of the opacities discussed above, which make it difficult to assess precisely what is driving the model's success. Here we present a series of additional simulations in which we rebuild the Brouwer model, replicate its original result, and then probe it further by examining the factors that influence its performance. We find that the success of the model in capturing divergent sensitivity to reversal anomalies depends on

potentially unrealistic properties of the training data, as well as the (more concerning) presence of the test data in the training set.

4.3.1 Simulated experiment

Brouwer et al. simulate the result reported by Hoeks et al. (2004), described above. Brouwer et al. reframe the four conditions of the Hoeks et al. experiment as shown in Figure 4.4. Recall that in that experiment, as Figure 4.4 shows, all three anomalous conditions elicited P600 effects, while N400 effects were present only for the "poor fit" conditions (what Brouwer et al. call "mismatch anomalies"), with the N400 showing no sensitivity to the incongruity when there was good lexical fit (what Brouwer et al. call the "reversal" condition).

Item	Condition	Effect
De speer werd door de atleten geworpen *The javelin was by the athletes thrown*	Control (Passive)	—
De speer heeft de atleten geworpen *The javelin has the athletes thrown*	Reversal (Active)	P600
De speer werd door de atleten opgesomd *The javelin was by the athletes summarized*	Mismatch (Passive)	N400/P600
De speer heeft de atleten opgesomd *The javelin has the athletes summarized*	Mismatch (Active)	N400/P600

Figure 4.4: Hoeks et al. (2004) effects (credit: Brouwer, Crocker, Venhuizen, and Hoeks (2017))

4.3.2 Training data

In each phase of training, Brouwer et al. train on data composed of two halves: eight thousand "stereotypical" sentences, and eight thousand of what we will call "all-combinations" sentences. The stereotypical half includes active and passive sentences reflecting each of

ten non-anomalous agent-action-patient triplets, such as ⟨*player scored goal*⟩ and ⟨*lawyer sued company*⟩. These triplets are the basis of the passive control condition in the test simulation. Each stereotypical triplet is represented 800 times in this half of the data (400 active, 400 passive).

By contrast, the all-combinations data represents every possible agent-patient-action combination from among the 20 nouns and 10 verbs ($20 \times 20 \times 10 = 4000$) in the training vocabulary, each with an active and passive form (8000). The model thus trains on every possible combination of agents, patients and actions (e.g., ⟨*goal sued lawyer*⟩), but it sees the non-anomalous combinations with much greater frequency (at a ratio of 401:1). All of the test sentences—those that will be used during the simulations—are also seen during training.

4.3.3 Replication

We implemented the RI model architecture, using the same two-phase training and the same number of training epochs reported by Brouwer et al. All simulations were run on Dutch data from Brouwer et al.'s Simulation 1.[4] Figures 4.5 and 4.6 show all simulation results.

The "Replication" setting of Figures 4.5 and 4.6 shows our results when training on the data described in Section 4.3.2. While the relative values of the mismatch conditions differed slightly from those reported in Brouwer's simulation (we find that these fluctuate somewhat between runs), we have a consistent replication when it comes to

[4]We thank Harm Brouwer for generous discussion, clarification, and provision of the original Dutch word vectors.

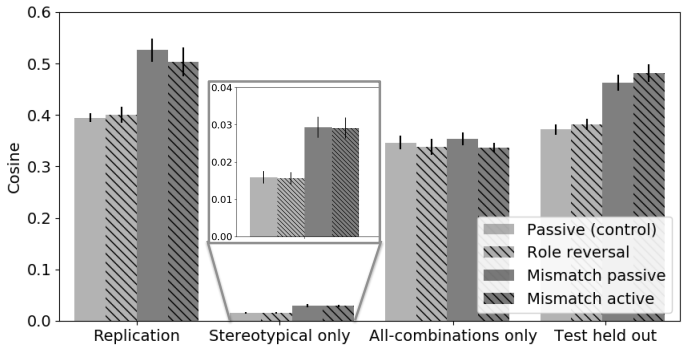

Figure 4.5: N400 simulation results across different training settings

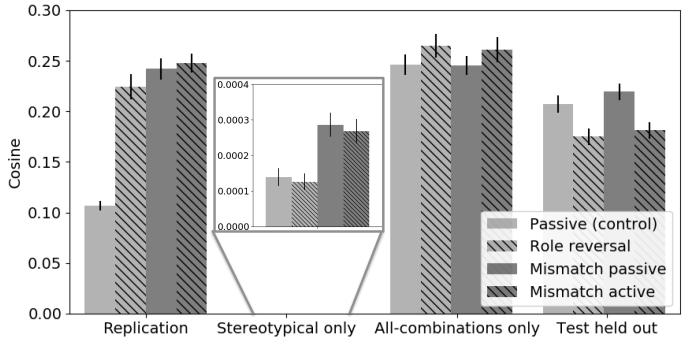

Figure 4.6: P600 simulation results across different training settings

the critical reversal condition: the N400 layer shows insensitivity to role reversal—in that the reversal condition patterns with the non-anomalous passive condition—while the P600 layer shows sensitivity to the reversal. Replicating Brouwer, our model also reaches 100% "comprehension" performance for output, where success (as defined by Brouwer) is achieved when the output sentence vector has the highest cosine similarity with the target sentence vector as compared to all other training sentence vectors.

4.3.4 Examining effects of training data

Let us now think about the details of the environment within which this model is operating. In probing the model, we will focus on the extent to which its success is driven by the particulars of the data on which it is trained.

Upon closer inspection, the data used to train this model is potentially problematic for two reasons. First, due to the all-combinations data, the training contains many implausible sentences that humans are unlikely to have encountered prior to an experimental context. These implausible sentences include the anomalous meanings used in the simulations of the Hoeks et al. experiment, such as "the goal served the player", as well as a host of other implausible meanings, such as "the meal sang the painting". The authors justify the all-combinations data as teaching the model that any noun can serve either an agent or a patient role.

A second, and more pressing, concern is that all of the sentences on which the model is tested in the critical simulations are also included in the training data, such that the model is not required to generalize at simulation time. This means that the model's performance could in truth be contingent on the distribution of test sentences themselves within the training data—an unrealistic assumption for modeling of human cognition, given that humans frequently process sentences that they have not previously encountered. (As we will discuss in greater detail in Chapter 6, a critical principle adhered to in machine learning is the separation of training and test data, to ensure that models are able to generalize to new data.)

Isolating effects of each half of the training set

To isolate the contributions of each training data component, we ran separate simulations with each half. The "Stereotypical only" and "All-combinations only" settings in Figures 4.5 and 4.6 show simulation results when the model has been trained only on the corresponding half of the data.

The stereotypical-only simulation is of particular interest, as it allows us to test the model's performance when implausible sentences have been excluded from the training data. It is also the first setting in which the model is tested on partly unseen data (specifically, in this setting the sentences from the non-anomalous passive condition are seen in the training data, but the sentences from the three anomalous conditions are not). When we train the model on this more restricted dataset, we find that comprehension accuracy drops from 100% to 25%—indicating that in this training setting the model does not generalize to comprehension of untrained sentences. Inspection verifies that the 25% of simulation sentences with correct output are those in the non-anomalous passive condition seen in training.

As for the N400/P600 simulation, we see that with stereotypical-only training, the model's N400 layer still captures the desired pattern, with attenuated N400 amplitude for reversals. The P600 layer, by contrast, no longer shows the desired sensitivity to all three anomalies, instead resembling the N400 layer.

With all-combinations training data only—when implausible sentences are included in training and all sentences are encountered with equal probability—simulation comprehension accuracy returns to 100%.[5] However, both the N400 and P600 layers now fail to simulate the desired effects, as all conditions pattern roughly together at both layers.

These results suggest that the model's simulation of the N400 pattern is driven by the relative training frequency of particular noun-noun-verb (NNV) combinations, without regard to role. This conclusion emerges from the fact that the model produces N400 fa-

[5] We did require more training epochs than other simulations in order to reach 100% comprehension for this setting.

cilitation for the reversal "the javelin has the athletes thrown", when it has only seen the verb *thrown* in the context of the canonical active and passive sentences "the javelin was by the athletes thrown" and "the athletes have the javelin thrown". We can say, then, that this layer indeed appears to be behaving in a manner that is sensitive to lexical content but not to syntactically-constrained role information (characteristic of the kind of asyntactic lexical mechanism we have been discussing). This is consistent with the general observation that this is the type of information that existing models such as neural networks are best able to capture—but as we have seen in Chapter 3, we do not need a complex model to capture this pattern of sensitivity, when simply averaging word representations together has this effect as well.

By contrast, the results suggest that the model's successful simulation of the P600 effect requires the distribution provided by the full training dataset, as neither half independently produces the desired P600 pattern. More to the point, what these separate simulations suggest is that the P600 layer appears to reflect training frequency of particular sentences or role triplets—but only when trained with the combined stereotypical and all-combinations data. When trained on the stereotypical-only data, it seems to reflect instead the same sensitivity pattern as the N400 layer. As we will see in the next simulation, this pattern of behavior at the P600 layer is likely driven by a combination of the granularity of the problem being solved in a given training setting, and the model's sensitivity to the distribution of specific surface characteristics of the training sentences.

Testing with held-out data

An important possibility is that the substantial failure of comprehension generalization in the stereotypical-only setting could be due to the fact that the stereotypical sentences are distinguishable by very coarse features, and as a result, in this training setting the model may simply not be forced to arrive at any kind of sophisticated solution during training. To rectify this, we created a new dataset with sentences removed based only on the simulation data: the stereotypical-only half was converted to active sentences only, to avoid occurrences of the simulation's passive controls, and from the all-combinations half we removed all sentences reflecting role triples from the test data, both active and passive. The resulting training data has the original degree of skew toward stereotypical meanings (albeit only instantiated in active form), contains both actives and passives (since we only exclude passives that will be seen in the simulation), and embodies the principle that any noun can be an agent or a patient.

In this setting, the simulation comprehension returns to 100%, indicating that when trained on fine-grained data, the model's comprehension performance does generalize to untrained sentences. However, we see in the "Test held out" setting in Figures 4.5 and 4.6 that the N400/P600 patterns now change, and in particular that the P600 layer, rather than reflecting sensitivity to whether the meaning of the sentence is anomalous, appears instead to reflect the skew of the training data toward active sentences.[6]

[6]Note for this simulation: we find that occasionally the training loss spikes in the final epoch, an occurrence which is often accompanied by noisier N400 and P600 results—for most simulations we were able to avoid this, but due to erratic loss in this hold-out simulation, we used a diminishing learning rate as employed by Brouwer. Other simulations used a constant learning rate.

4.3.5 Explanation

These simulations allow us to zero in on logical explanations of the model's behavior, in terms of optimal solutions that the model is likely to arrive at during training. More specifically, these simulations suggest that the behavior of a given module of this model essentially comes down to the type of prediction that that module is mapping to, and what cues the model needs to remember in order to make that prediction based on the training data. In this section we will walk through the relevant reasoning.

Consider first the Retrieval module. Due to the nature of the training regimen, this module arrives at an optimal solution by mapping to (a rough approximation of) the vector representation of the current word. The N400 estimate arises based on how much the intermediate layer preceding that output changes from one word to the other.

We can come to a straightforward explanation of this layer's behavior if we simply consider which cues are useful for this prediction task, and what those cues predict at a given time. For outputting the word vector, the word identity (the one-hot input) will be perfectly predictive, but because of the way that the training environment is constructed, clusters of content words will also be predictive of each other. For instance, *athlete* and *javelin* occur far more often in the training sentences "the athlete has the javelin thrown" / "the javelin was by the athlete thrown" than either word occurs in any other context—so those two words are highly predictive of each other, and they are also each predictive of *thrown*. By contrast, the prediction of the word vector will not be aided by attending to function words in the sentences (though some information about function words needs to be retained for the subsequent module).

Based on these facts, it is quite sensible that the Retrieval module would show sensitivity based on how well the verb fits with the preceding nouns—as long as the training environment is such that certain pairs of nouns are heavily predictive both of each other and of their good-fit verb. (Alternatively, if there is no such skew, we can expect to see a lack of this pattern, as we do in the all-combinations setting). Specifically, the nouns are the only context information relevant to aiding the verb vector predictions, and the states involved in predicting the nouns and the good-fit verb are similar (because each noun predicts both the other noun and the verb). So it is logical that there would be less of a change of state in the N400 layer when predicting the good-fit verb (*thrown*) than when predicting the poor-fit (mismatch) verb (*summarized*).

The Integration module, on the other hand, must arrive at a representation of event structure, such as <athlete throws javelin>, which requires differentiating the roles filled by the nouns. This in theory requires sensitivity to syntactic words—but only to the extent that those words are needed to identify the most probable outcome.

For the original training setting, the model sees stereotypical sentences like "the javelin was by the athlete thrown" and "the athlete has the javelin thrown" at a very high frequency compared to other sentences with *athlete*, *javelin*, or *thrown*. At the point of "the javelin was by the athlete —", the event <athlete throws javelin> is by far the most probable outcome, so the predicted output is not expected to change much with the arrival of *thrown* (and by extension the intermediate mapping layer may not change much in that circumstance). By contrast, if there is a continuation of *summarized* (mismatch passive condition), the predicted mapping must change substantially. Similarly, if the sentence is "the javelin has the athlete —", there is no specific high probability outcome,

so the arrival of the verb (whichever verb it is) will logically cause a greater change in the predictions, and in the intermediate mapping to those predictions. In all cases, the nouns alone cannot allow for accurate predictions, so we must assume that the network is retaining information about the syntactic words and using them in the final output.

In the case of the stereotypical-only setting, by contrast, the model only sees the words *athlete*, *javelin* and *thrown* in the sentences "the athlete has the javelin thrown" and "the javelin was by the athlete thrown", both of which map to the event output of <athlete throws javelin>. This means that upon encountering either *athlete* or *javelin*, the model can confidently predict a single output, such that it has no need to make use of syntactic words, and it will logically arrive at a pattern of behavior that mirrors that of the Retrieval module, such that only the nouns appear to be relevant.

In the "Test held out" setting, the model has seen active stereotypical sentences such as "the athlete has the javelin thrown" in very high frequencies, but the strings "the javelin has the athlete —" and "the javelin was by the athlete —", which occur in the simulations, do not have any strongly-predicted outcomes. This can help us to explain the fact that the model seems equally "surprised" regardless of the verb, and instead is more surprised in passive sentences as compared to active sentences. Because none of the simulation strings have any particularly predicted verbs, it is likely that the model simply is not making any strong predictions about the verb component of the output until the verb itself arrives. The smaller change for the active sentences could be explained if the model is simply making more confident predictions about the agent and patient assignments within the active sentences—because it has seen more active sentences—such that the mapping at the second noun is closer to the final mapping for active than passive sentences.

How cognitively interesting is the model's behavior in light of this explanation? Essentially, the model works because it designates particular probabilistic tasks for different modules to solve, and it arrives at optimal solutions to these tasks based on the predictiveness of cues in the training. At a general level, this is a reasonable hypothesis for predictive processing in humans.

Notably, however, we see from these simulations that the model is not making use of actual event frequencies—for instance generalizing from "the athlete has the javelin thrown" to the unseen sentence "the javelin was by the athlete thrown". Instead, the model appears to be dependent on specific word strings, with no evidence of abstracting away to meaning-based predictions. Along this line, the model's performance relies critically on the specific makeup of the training environment, which almost certainly does not accurately reflect reality.

4.3.6 Takeaways

These simulations allow us to replicate and confirm the behavior of the model reported by Brouwer et al. (2017), and to examine more closely the factors that drive the performance of that model. We learn from these simulations that the model's success is seemingly driven in large part by the skew in the distribution of training sentences, and the effects of those distributions on the optimal solution for solving the target problem for which the model is trained. Furthermore, rather than reflecting abstract generalization based on sentence meaning, the model appears to pattern with relative training frequencies of the word strings in the simulation sentences. This is problematic for interpreting

the model's performance in terms of the RI theory that it claims to support, as the performance appears to be driven not (solely) by the RI-based architecture, but rather by properties of training which are neither included in the claims of the theory nor sufficiently cognitively plausible to be innocuous assumptions.

Importantly, we see here that the model *does* manage to capture the desired patterning of the N400 component, as long as the particular combination of nouns and verb has been seen with a comparatively high frequency. As we have observed above, this is a simple type of pattern to capture even by simple vector averaging, so the Brouwer model's approach to capturing it is almost certainly more complex than necessary—but it is worth acknowledging that the Brouwer model does capture this pattern more robustly than it does the P600.

Unsurprisingly, it is the more sophisticated message-based facilitation—requiring sensitivity to syntactic cues and anomaly of the resulting sentence meaning—that turns out to be out of reach for this model in terms of robustness and generalizability. As we discussed in the beginning of this chapter, in the experiments reviewed and simulated here it is the P600 pattern that shows this type of sensitivity. However, we will see in Chapter 5 that the N400 too can show sensitivity to these more sophisticated cues, even in the context of a role reversal paradigm. At that point we will be faced with two challenges: 1) how do we model a more complex N400, which at times shows more sophisticated sensitivities than either the Brouwer model or the averaging model can capture, and 2) how do we accurately model the more sophisticated effects, when existing computational models (even those more recent or sophisticated than the Brouwer model, as we will see in Chapter 6),

currently fall short of capturing this aspect of human cognition? These are questions that we will address in Chapter 5.

Finally, this model is advertised as a single-stream model, by contrast to multi-stream models embodied by many of the verbal theories reviewed above. How does the single-stream / multi-stream distinction fit into the lexical/syntactic distinction that we have been discussing?

We have not made any strong commitments with respect to a single-stream versus a multi-stream understanding of our lexical/syntactic distinction. As we discussed in Chapters 1 and 2, we could reasonably understand our asyntactic lexical effects to arise as a simple byproduct of the lexical access process, and depending on our understanding of the composition process and how it serves to filter inputs, it is certainly possible that the same lexical representations that produce associative effects would also be fed as input to the compositional processes that produce syntax-sensitive effects. This is consistent with a single-stream architecture. Alternatively, it is possible that the asyntactic lexical effects arise from an entirely parallel process that does not feed directly into the compositional processes. We are only committed to the existence of some process or processing stage that gives rise to asyntactic lexical effects in some way.

However, given that Chapter 5 will show syntactically-constrained sensitivities in the N400, we cannot be satisfied with a single-stream theory if, like Brouwer's, it reserves more sophisticated, syntactically-constrained anomaly sensitivity exclusively for the P600.

4.4 Conclusion

In this chapter I have reviewed the core canon of results pertaining to the role reversal phenomenon, along with theories that have been proposed to account for this phenomenon. I have reviewed two connectionist computational models that have claimed to simulate the role reversal phenomenon, and discussed some implications and limitations of these models. Finally, I have proceeded to replicate and further probe the behavior of one of these models: that reported by Brouwer et al. (2017). We see in these experiments that this model's performance is driven not (solely) by the Retrieval-Integration hypothesis that it aims to support, but by the particular distributions of the sentences in its training data. In particular, its capturing of the P600 pattern, which is the correlate in these experiments of syntactically-constrained message-based anomaly patterns, appears to reflect overfitting to the surface properties of the test sentences within the training data, rather than a generalizable effect based on sentence meaning.

Chapter 5

Modeling parallel interaction of lexical and syntactic processes

5.1 Introduction

In this chapter I will review a newer set of results (Chow, Lau, Wang, & Phillips, 2018; Ehrenhofer, 2018), which indicate that under some circumstances—by contrast to the above results showing only the P600 to reflect role reversal anomaly—the N400 too can show role reversal sensitivity. One of these experiments, Ehrenhofer (2018), is structured nearly identically to that of Chow et al. (2015) but uses different stimuli—presenting us with a particularly useful testing ground for zeroing in on the factors that could drive this contrast in outcomes.

I will discuss the implications of these results for the issues that we have been discussing, and for the above existing theories. I will then report the results of a computational simulation that combines components of the Kuperberg (2016) account with ad-

ditional factors suggested by the characteristics of the Ehrenhofer (2018) experiment, to produce a fully specified model able to produce the observed complexities in N400 behavior. Like the Brouwer and Rabovsky models, this first model operates within a synthetic environment, and thus relies on the assumption that the characteristics of that environment are representative in the relevant respects. However, this nonetheless serves as a useful initial exercise for demonstrating how these mechanisms could be fleshed out and used to account for the observed results, if the given assumptions about the environment are true.

Finally, I will introduce analyses of the real experimental stimuli used for the Chow et al. (2015) and Ehrenhofer (2018) experiments, aimed at better understanding what could have driven the different outcomes between these experiments. Based on these analyses, I will then introduce a set of new models and simulations which test a number of hypotheses about the relative roles of lexical and syntactically-constrained processes in generating N400 facilitation, making use of the finer-grained characteristics of the individual experimental items used in these studies. This allows us to move beyond the broader strokes of the intended manipulations of these experiments, and instead to test how the dynamics of the specific experimental stimuli play out in aggregate under different models.

5.2 The "smart" N400

In the classic role reversal studies reviewed in Chapter 4, the striking finding is that the N400, which has traditionally been associated with sensitivity to semantic anomaly, fails to show sensitivity to a role reversal anomaly, with the P600 instead showing that sensitivity.

In this section I will now discuss two recent results that complicate the picture, by employing the same role reversal paradigm, but yielding results in which the sensitivity of the N400 to the semantic anomaly appears to be restored.

5.2.1 N400 sensitivity after a delay

The first of these results is reported by Chow et al. (2018). In a study conducted in Mandarin Chinese, the authors test the role reversal paradigm, but introduce an additional manipulation in the duration of the delay between the critical preceding arguments and the target verb.

No delay
last week police BA suspect arrest ... (lit. "Last week the police arrested the suspect")
last week suspect BA police arrest ... (lit. "Last week the suspect arrested the police")
With-delay
police BA suspect last week arrest ... (lit. "Last week the police arrested the suspect")
suspect BA police last week arrest ... (lit. "Last week the suspect arrested the police")

Table 5.1: Chow, Lau, Wang, and Phillips (2018) stimulus conditions. Note that these are translations from Mandarin, and "BA" refers to a syntactic particle that does not have an equivalent in English.

In the no-delay condition, as in the previously reported role reversal results, the authors observe a P600 effect but no N400 effect. However, in the case of the additional delay between arguments and verb, they find a significant N400 effect (in addition to a P600 effect).

This result follows up on a previous experiment, reported in the same paper, which combines the delay with a manipulation of contextual constraint, and shows that the N400 effect is obtained with the delay, but only in the case of high-predictability contexts. For

lower-predictability contexts (for example: *Mr Liu BA parrot that summer train*), even with the intervening delay, the authors observe the typical response of a P600 effect but no N400 effect.

For the sake of clarity, because this is the second Chow et al. study that we will be discussing, I will from this point refer to this new study as "Chow18-Delay", and the Chow et al. (2015) study, which is the original English role reversal study described in Chapter 4, as "Chow15-Reversal".

5.2.2 N400 sensitivity with differently-constructed stimuli

The second result of interest to us in this chapter is reported by Ehrenhofer (2018). Intending to replicate the Chow15-Reversal study and test additional hypotheses about the comparative roles of arguments and verbs in online processing, Ehrenhofer created a stimulus set modified from the original Chow15-Reversal stimuli. Though these stimuli use the same stimulus structure and role reversal manipulation, they differ in two primary ways. First, Ehrenhofer includes as targets the high-cloze continuations for both argument orders, as shown in Table 5.2. By contrast, Chow15-Reversal included the high-cloze continuations for only one of the argument orders.

Argument order 1 = high cloze
the restaurant owner forgot which customer the waitress had served ...
the restaurant owner forgot which waitress the customer had served ...
Argument order 2 = high cloze
the restaurant owner forgot which customer the waitress had tipped ...
the restaurant owner forgot which waitress the customer had tipped ...

Table 5.2: Ehrenhofer (2018) stimulus conditions.

The second difference is that, as part of this symmetric inclusion of stimuli, Ehrenhofer selected these stimuli so as to ensure that for every item, both argument orders yield a relatively high-cloze continuation.

With these modified stimuli, Ehrenhofer (2018) finds that rather than replicating the commonly-observed N400 insensitivity to the role reversal, she instead finds a significant difference in N400 amplitude between high-cloze canonical sentences ("... which customer the waitress had served") and low-cloze role reversals ("... which waitress the customer had served"').[1]

5.3 Implications

What can we take away from these N400-sensitivity results? In combination, the results produce an interesting picture.

The Chow18-Delay result suggests that the N400 itself can vary in the sophistication of information that it reflects, depending on the amount of time for processing of key context elements. This suggests that different types of information influence the facilitation reflected in the N400 at different amounts of delay. More specifically, the no-delay N400 appears to reflect what we have been referring to as asyntactic lexical processes, while the with-delay N400 appears to reflect the syntactically-constrained information needed

[1] It should be noted that the original report of these results described a significant N400 effect for this experiment as well as for the associated NVN experiment (which will be described below). This is the report on which we base the discussions and model assessments below. Subsequent reports from that study clarify the finding as a main effect of cloze across the NNV (role reversal) and NVN (see below) context types, with no interaction of context type and cloze—this is consistent with there being significant effects within each context type, but does not confirm it. Pending follow-up analyses to confirm the individual effects, we will operate under the assumption that the originally-reported effects hold. In the event of a change, this will simply result in a reorganization of the target significance patterns and our partition of model performance.

to show sensitivity to the reversal. So we might suggest that the no-delay N400 has access only to asyntactic lexical processes, while the with-delay N400 has access to the syntactic and world-knowledge information needed to recognize the reversal anomaly.

The Ehrenhofer (2018) result, in combination with the Chow15-Reversal result, suggests that holding constant the basic contextual structure, experimental manipulation, and delay between relevant context and target verb, we can see in one study the classic lack of N400 effect, and in another study the presence of an N400 effect. This presents us with a valuable opportunity to gain finer-grained insight into the factors driving the absence or presence of an N400 effect in these sentences. More specifically: no account based solely on the role reversal manipulation can account for both the Chow and the Ehrenhofer results. Instead, we can look closer at the characteristics of the stimuli to learn what additional factors might explain these differences.

5.3.1 Revisiting existing theories

How do the existing theories hold up to these results? The immediate challenge for the majority of existing theories is that, like the Brouwer model, they aim to explain the N400 and P600 in such a way that the N400 will lack sensitivity to the reversal of arguments preceding the verb, and they do not explicitly account for any re-emergence of such sensitivity with different verb-argument combinations, or with a delay in target onset.

For many of the theories, we can imagine a general accommodation of these results. To accommodate the timing manipulation results, we can imagine that Hoeks et al. (2004) could claim that the underspecified message representation is able to become more spec-

ified given additional time, reducing the resulting integration cost. Theories that posit contribution by a plausibility-based or semantic processing stream could claim that the dominance of this stream reduces over time, resulting in a more syntactically-driven influence on the N400. Accounts that seemingly cannot accommodate the timing manipulation results include those that assume detection of a thematic anomaly to block any attempt at integration (thus disallowing any N400 effect), or those that assume the N400 to be driven solely by the asyntactic plausibility-based or semantic streams.

As for the Ehrenhofer results, we can again imagine (in the absence of more concrete knowledge of differences between the two stimulus sets, to which we will shift below), general ways in which the theories might accommodate this outcome. Again taking Hoeks et al. (2004) as an example, we could imagine this account claiming that for some reason the message interpretation is able to become more specified in the case of the Ehrenhofer stimuli. Those that posit a plausibility or a semantic stream could claim that something about the Ehrenhofer stimuli causes the weighting of that stream to be reduced. The primary commitments that truly cannot be accommodated are again those that assume role reversals to block the N400 entirely. (One might think that a lexical-only account of the N400 would also be unable to accommodate the Ehrenhofer result, but as we will see below in our analysis of the stimuli, this is not quite true.)

Chow et al. (2016) do have an explicit theory of the delay results: they suggest that there may be an in-principle limitation on how quickly role information can be used to query semantic memory. They describe a framework within which candidates for the relevant event/verb to be predicted are first generated by querying semantic memory based on argument identities alone, and only after more time can a candidate or candidates can

be identified using role information. This account appears at face value to encounter some difficulty with the Ehrenhofer (2018) results, which show role sensitivity in the absence of a delay manipulation. However, as we will discuss below, we might reasonably consider the Ehrenhofer (and Chow15-Reversal) stimuli to in fact align more appropriately with the "with-delay" condition of Chow18-Delay.

One interesting direction for addressing the Chow15/Ehrenhofer contrast comes from the Kuperberg (2016) hypothesis, which suggests that different cues can underlie event predictions that drive the N400, and the nature of the cues used at a given time depends on the strength of the predictions that arise from those cues. This account does not have any particular explanation for the evolution on the timing dimension, but it lends itself to a reasonably straightforward potential explanation of the divergence between the Chow15-Reversal and Ehrenhofer results: perhaps in the Ehrenhofer experiment, syntactic/role information led to stronger predictions than did the simpler lexical information, so the predictions in that experiment were based on the role information. By contrast, in the Chow15-Reversal experiment, simpler cues like lexical information led to stronger predictions than did the role information, so the lexical information drove the predictions in that case.

(Or, to be more precise: syntactic cues drove the predictions on *sufficiently many* of the Ehrenhofer stimuli to show an N400 effect for that experiment, and syntactic cues drove the predictions on *sufficiently few* of the Chow15-Reversal stimuli to produce a lack of N400 effect in that experiment.)

In the next section we will run a simple computational simulation demonstrating how this basic picture could work.

5.4 Computational simulation: Kuperberg(-esque) account of Chow/Ehrenhofer contrast

In this section we will walk through a concrete example of how a version of the Kuperberg (2016) hypothesis can be implemented and further specified, in order to produce the contrast in effects observed across the Chow et al. (2015) and Ehrenhofer results. We will do this by running a brief computational simulation.

We will make some adjustments and additions to the Kuperberg account. First, as we noted in Chapter 4, Kuperberg's account of the reversal-insensitive N400 is stated in terms of cues that include not just lexical content, but also words' linear orders as well as associated event schemas. We will simplify this picture for the present simulation by assuming just two different classes of cues: lexical-only cues, and syntactically-constrained cues.

We will also fill in a couple of missing blanks from the Kuperberg account. The first of these blanks is the question that we raised and put on hold in Chapter 4: why the less sophisticated subset of cues should be utilized only in the role reversal contexts ("... which waitress the customer had served") when the full set of available cues is proposed to be used in all other conditions.

A possible answer to this is suggested by the criteria that were used to construct the Ehrenhofer stimuli—specifically, the requirement that there be a high-probability continuation for each argument order. The hypothesis that this suggests is that while each argument order in the Ehrenhofer stimuli will have a strong possible continuation, this may not be the case for the Chow15-Reversal stimuli. Because no specific guarantee was

made by Chow et al. such that the reversed argument orders would have high probability continuations, there may be a greater asymmetry in those stimuli. This hypothetical distinction is illustrated in Table 5.3. As we will show in this simulation, this state of affairs could explain why Chow et al.'s reversal context alone might select for less sophisticated cues.

Ehrenhofer et al.
"which bull the cowboy had ..." ridden = high probability
"which cowboy the bull had ..." gored = high probability
Chow et al.
"which customer the waitress had ..." served = high probability
"which waitress the customer had ..." ??? (no high probability continuation)

Table 5.3: Hypothesized situation of symmetric continuation strength for Ehrenhofer (2018), and asymmetric continuation strength for Chow15-Reversal. Note that although our particular example here of "which waitress the customer had ..." may have a reasonably high-probability continuation (e.g., *tipped*), the point here is to convey a situation in which, in the Chow15-Reversal stimuli, the reversal condition represented by that example tends overall to lack high-probability continuations.

The second blank to fill in is the specific definition of prediction strength that we will use to decide which cues (which information sources) will drive the final prediction that influences the N400. In this simulation we will define a "strong" prediction as one that is comparatively unambiguous. An obvious way to quantify this is the entropy of the distribution over candidate predictions. Another possible way of quantifying this is the ratio of the strength of the top candidate prediction to the strength of the next several candidate predictions. We will show results for both of these.

This account requires us to assume the existence of mappings from different sets of cues to probabilities of event interpretations. In this simulation we are limiting ourselves to two types of cues—asyntactic lexical cues and syntactically-constrained cues—so we

assume only two types of mappings. The lexical mappings are formulated as "given the presence of words x and y, what is the probability distribution over possible events being described?". Syntactic mappings, by contrast, are formulated as "given that we have x in subject position and y in object position, what is the probability distribution over possible events being described?"

Importantly, we formulate the latter mapping based on syntactic roles (subject, object) rather than semantic roles (agent, patient) because at the critical point in the experimental sentences, e.g., "which waitress the customer had –", the language comprehender cannot actually know with certainty that the customer is the agent, given that the sentence could continue with a passive (e..g, "... been served by"). Using mappings based on syntactic rather than semantic role is therefore more accurate, and allows us to accommodate the possibility of passives in our simulation (keeping in mind that passive constructions are less frequent).

Note also that in this formulation, in keeping with Kuperberg's approach, our asyntactic lexical mechanism is conceived of as a probabilistic mapping from a given set of words to an event prediction, rather than as a measure of simple lexical relations. We will return to our use of lexical relations in the next section.

To create mappings between cues and event predictions, we create a simple synthetic data environment along the lines of those used by the connectionist models described above—however, we skip the training of a neural network, and instead compute conditional probabilities directly based on the instances that make up the environment. In doing this, we make a basic assumption that the strength of mappings between cues and event predictions will faithfully reflect conditional probabilities in the environment.

Our synthetic dataset consists of sentence instances characterized by identification of the syntactic subject, the syntactic object, and the event being described (this is all the information needed to compute the relevant mappings for this simulation). For the purpose of this section, I will refer to a given assignment of syntactic roles to nouns (e.g., *bull*-SUBJ *cowboy*-OBJ) as a "construction".

Sentence instances are defined based on two types of argument pairs: symmetric (*bull, cowboy*) and asymmetric (*waitress, customer*). Note that because the purpose of this exercise is simply to demonstrate how hypothesized probabilistic properties would interact with predictive mechanisms to produce the observed outcomes, the probabilistic properties that define these item types are hypothetical and not based on corpus data or human responses. In Section 5.5 we will shift to quantifying properties of the actual experimental stimuli with greater precision.

For symmetric *bull/cowboy* sentence instances, each construction is strongly associated with one particular event description: *cowboy*-SUBJ *bull*-OBJ is associated with a *ride* event 50% of the time, and *bull*-SUBJ *cowboy*-OBJ is associated with a *gore* event 50% of the time. Additionally, each of these constructions has a small number of instances (5%) in which it is associated with the strong verb from the *other* construction, to reflect the influence of passives—for instance, *bull*-SUBJ *cowboy*-OBJ would in 5% of instances be associated with a *ride* event, expressed as a passive. Each of the remaining 45% of instances is associated with one of nine generic verbs, distributed uniformly such that each generic verb occurs in 5% of the instances of a given construction.

In the asymmetric *waitress/customer* stimuli, one construction—*waitress*-SUBJ *customer*-OBJ—resembles the above constructions in being associated with a single event of

served 50% of the time. However, unlike the constructions above, we do not include any portion of instances representing passives, based on the assumption (which we will move to momentarily) that the reverse construction of *customer*-SUBJ *waitress*-OBJ does not have any strong events to inject as a passive. Thus, the remaining 50% of instances for this construction are occupied by ten generic verbs, each occurring 5% of the time.

In the case of *customer*-SUBJ *waitress*-OBJ, the assumption is that this construction does not have any strong mappings to associated events, so unlike the previous constructions, it lacks any verb that dominates its distribution. It has all of the same generic verbs as the reverse construction, distributed uniformly and occurring in the same frequencies that they occur in the reverse construction. Additionally, based on the same intuition that strong events are also likely to occur as passives, we include a number of instances of *serve* as the associated event, in the same frequency as the generic verbs.[2]

Based on these collections of instances, we then compute the probability distributions for possible events, conditioned on a) the identities of the syntactic subject and object (our syntactic cue mapping), and b) the identities of the nouns in the construction, irrespective of syntax (our lexical cue mapping).

To determine which cue-type will be chosen for a given condition, we then compute our measures of strength—entropy and average candidate ratio—based on these prob-

[2]This means that this construction overall is not only characterized by a different distribution—it is also less frequent than the other constructions. It seems reasonable that this could reflect actual frequency characteristics—constructions that have no stereotypical event associations may also be less frequent (perhaps we talk about waitresses doing things to customers more frequently precisely because there is a stereotypical event of this kind to describe). But more to the point, this is actually an assumption that allows for the simulation to work best—it remains to be determined whether the assumption reflects reality, but it is worth noting that within the settings used for this simulation, this frequency assumption helps to achieve the best results.

Construction	SYNT	LEX
waitress-SUBJ customer-OBJ	**1.84**	2.14
customer-SUBJ waitress-OBJ	2.40	**2.14**
cowboy-SUBJ bull-OBJ	**1.84**	2.06
bull-SUBJ cowboy-OBJ	**1.84**	2.06

Table 5.4: Entropy values for two different cue types for each of four constructions. Lower entropy values represent stronger predictions. Values corresponding to stronger cues have been bolded.

Construction	SYNT	LEX
waitress-SUBJ customer-OBJ	**10.0**	5.5
customer-SUBJ waitress-OBJ	1.0	**5.5**
cowboy-SUBJ bull-OBJ	**10.0**	4.38
bull-SUBJ cowboy-OBJ	**10.0**	4.38

Table 5.5: Average candidate ratio values for two different cue types for each of four constructions. Higher average ratio values represent stronger predictions. Values corresponding to stronger cues have been bolded.

ability distributions. Average candidate ratio is computed as the average ratio of the probability of the top candidate to the probabilities of the next four candidates.

Tables 5.4 and 5.5 show the entropy and average candidate ratio values for each cue type within each syntactically-defined condition. Cues that yield distributions with *lower* entropy represent stronger predictors, while cues that yield distributions with *higher* average candidate ratios represent stronger predictors. Values corresponding to stronger cues have been bolded.

We see from these results that either of these measures is able to simulate the desired effect: for both orders of the symmetric *cowboy-bull* pair, the syntactic cues will be deemed the best predictors for either argument order, resulting in syntactically informed predictions, which would yield an N400 effect. By contrast, in the asymmetric *waitress-customer* pair, for only one order will the syntactic cues be deemed the best predictor—for

the construction with no strong event associated, the lexical cues are instead selected as the stronger predictor. This results in a prediction of *serve* as the event for both constructions, which would yield a lack of N400 effect on a target of *served*. This is also consistent with Kuperberg's claim that the *customer*-SUBJ *waitress*-OBJ condition alone would have a different profile of cue selection.

This is a useful exercise, in allowing us to lay out a potential implementation of the Kuperberg (2016) theory based on a likely distinction between the Ehrenhofer and Chow15-Reversal stimuli, and to confirm how this implementation of the model can account for the contrast in outcomes. However, because these properties are hypothetical, this simulation can only tell us how a system of this kind would play out given the assumption that these probabilistic properties of the stimuli hold. Similarly to the connectionist simulations described in Chapter 4, we are limited by the assumptions that the probabilistic properties that we have assigned to the stimuli in this synthetic environment actually hold of the true experimental stimuli.

For this reason, we will move in the next section to examining the real stimuli used in the Chow15-Reversal and Ehrenhofer experiments, to determine what quantifiable differences could in fact be driving the contrast in results.

5.5 Examining the stimuli

For a more concrete sense of what could be driving the different outcomes between the Chow15-Reversal and Ehrenhofer experiments, we need an analysis of the stimuli. In this section we describe the results of our conducting such an analysis.

In the above simulation we hypothesized that the Chow15-Reversal and Ehrenhofer stimuli differ in terms of the strength of predictions associated with different conditions. A good index of prediction strength for a given stimulus—and indeed the measure used by Ehrenhofer as the criterion for ensuring existence of strong continuations—is max cloze: the maximum probability of continuations provided by subjects in the cloze task. For example, a max cloze of .6 means that the most frequently-given completion for the corresponding context was given 60% of the time. A high max cloze is considered to correspond to a more constraining context, such that many people converge on the same completion, while a low max cloze suggests a less constraining context, such that subjects do not converge strongly on any common completion.

We make use of cloze values collected by Ehrenhofer for each of the Chow and Ehrenhofer stimulus sets. Examining the max cloze values computed from these cloze tasks, we find that the Chow15-Reversal stimuli have an average max cloze of .29 for high-cloze (canonical) stimuli, and .22 for low-cloze (reversal) stimuli. Comparing this with our hypothetical scenario from above, this corresponds to a .29 max cloze for stimuli in the category of *waitress*-SUBJ *customer*-OBJ, which we designed to have a single strong continuation, and a .22 max cloze for stimuli in the category of *customer*-SUBJ *waitress*-OBJ , which we designed to have no strong continuation. What this means is that the max cloze values do show a difference in the predicted direction, but it is not nearly as stark as in our idealized situation above.

How does this compare to Ehrenhofer's stimuli? Because Ehrenhofer uses the stimuli symmetrically as described above, the distinction of high-cloze and low-cloze *contexts* is not well-defined—this is because unlike Chow15-Reversal, Ehrenhofer uses each context

to construct both high- and low-cloze continuations ("which bull the cowboy had [ridden/gored]" and "which cowboy the bull had [gored/ridden]"). However, we can note that the average max cloze across contexts for Ehrenhofer is .35, which represents a stronger overall constraint than either context type in the Chow15-Reversal stimuli.

This brings us to the first major difference that we can identify between the Chow15-Reversal and Ehrenhofer stimuli: on average, the Ehrenhofer stimuli have higher max cloze, suggesting that they are on average more constraining / more predictive of a particular continuation than are the Chow et al. stimuli. The distributions of max cloze values for the stimuli from the two different experiments are shown in Figure 5.1, which shows that the max cloze values for the Ehrenhofer stimuli are in general shifted higher than are the values for the Chow stimuli. This will form the basis of two of the models tested below, as we reason about the specific mechanisms by which this difference—not quite what we envisioned in the previous section, but a difference nonetheless—could give rise to an N400 effect for Ehrenhofer and a lack of effect for Chow.

What about the lexical relations that occupied our attention Chapter 3? The role reversal stimuli are specifically constructed such that the lexical content of the preceding context is held constant between high- and low-cloze conditions—two of the nouns have simply been reversed in their syntactic position—so a lexical account alone should ostensibly not be able to account for the N400 contrast observed by Ehrenhofer. However, given that the order of the lexical items varies, there is a possibility that if we take into account decay of influence over time, as we did in modeling the Federmeier and Kutas experiment in Chapter 3, we could see a difference between conditions at that level. We may

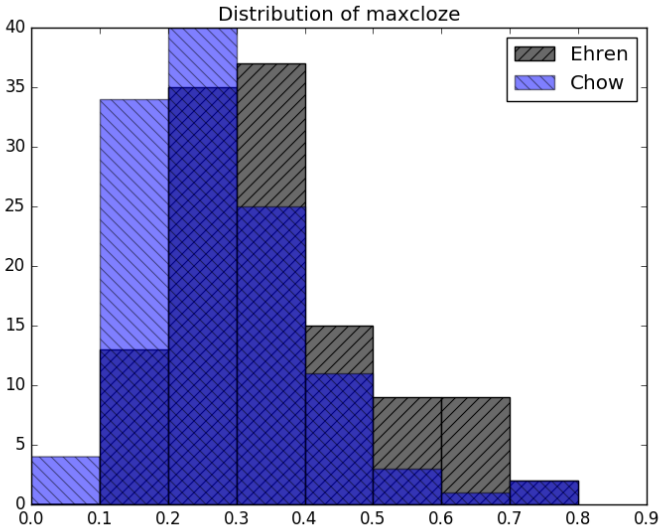

Figure 5.1: Distribution of max cloze for Chow and Ehrenhofer stimuli

also see general differences between the Chow and Ehrenhofer stimuli overall, collapsing across conditions.

To quantify lexical relations in these stimuli, we use pre-trained GloVe vectors (Pennington et al., 2014) of 50 dimensions, trained on English Wikipedia and Gigaword.[3] Using these vectors, we compute a context vector by taking the four most informative words as determined by inverse frequency (the four least frequent words, as in the agnostic models in Chapter 3), which in the case of these stimuli essentially selects out the nouns and the matrix verb. We then compute either an average or a weighted average

[3]We make the switch from word2vec (which was used for modeling the Federmeier and Kutas results) to GloVe after observing a lack of difference between these methods in modeling human behavioral results in Chapter 3, because pre-trained GloVe vectors are easily available and widely-used. Rabovsky et al. (2016) also reports "reasonable correspondence" between similarity relations using GloVe vectors and similarity relations using their own handcrafted word representations. As for the dimensionality, in preliminary testing using different dimensionalities of GloVe vectors, we found generally similar patterns of results, but we found that use of 50-dimensional vectors improved the capacity of the corresponding models to explain the human results. Whether this corresponds to an important assumption about the nature of lexical representations in humans is a question for future work.

(based on linear distance, as in Chapter 3) of these vectors, and compute the cosine of this context vector with the target vector.[4]

Finally, we compute one additional lexical relation: that between the final content word of the context (the subject, e.g., *waitress* in "which customer the waitress –") and the target.[5] Taking this measure is motivated by the fact that the subject is the most recently encountered word, and therefore it might be expected to have the strongest lexical effect—and also by the fact that when experimenters construct items of this kind, one of the things that is likely to drive the selection of constraining contexts is consideration of the stereotypical things that a given subject does in general (possibly with less regard to the nature of the object).

Tables 5.6 and 5.7 show the results. For the unweighted average, this cosine value is of course the same for both high- and low-cloze conditions, as identical words are being averaged in an unweighted manner. However, we do see from these values that the Chow contexts are in general more highly related to the targets, with an average cosine of .486, as compared to Ehrenhofer et al.'s .437.[6] When we move to the unweighted average, the high and low cosine conditions diverge slightly as expected, but the difference between conditions is comparable between Chow and Ehrenhofer.

By contrast, looking to the relation of the final content word (the subject) to the target verb, we find that the pattern of this relation in fact follows the observed pattern of results: for the Chow15-Reversal stimuli the relation is almost identical between conditions, while

[4]For multi-word targets, we remove function words and average together the vectors for remaining content words.

[5]If the subject consists of a multi-word phrase such as "restaurant owner", the vectors for the content words of this phrase are averaged.

[6]This is interesting, given that the Chow stimuli are also on average less constraining according to max cloze.

Measure	High-cloze	Low-cloze	Difference
Context - target	.437	.437	0.0
Context (weighted) - target	.438	.430	.008
Last word - target	.322	.288	.034

Table 5.6: Ehrenhofer, 50-dimensional vectors

Measure	High-cloze	Low-cloze	Difference
Context - target	.486	.486	0.0
Context (weighted) - target	.487	.477	.01
Last word - target	.322	.323	.001

Table 5.7: Chow, 50-dimensional vectors

for the Ehrenhofer stimuli we see a difference that is more than 30 times the size of the difference for Chow15-Reversal. This suggests a potentially promising direction for explaining the presence of an effect for Ehrenhofer and not for Chow.

In subsequent sections, we will incorporate these observed differences into processing models within the context of experimental simulations that will allow us to make claims about statistical significance by subjects, as is reported in the Chow15-Reversal and Ehrenhofer experiments.

5.6 Mapping to theories and models

In the previous section we identified two types of difference between the Chow15-Reversal and Ehrenhofer stimuli. First, the Ehrenhofer contexts are in general more constraining (higher max cloze) than the Chow contexts. Second, when we measure the lexical relation between the target verb and its subject (the final content word before the arrival of the verb), we see almost no difference between conditions for the Chow stimuli, but

a much more substantial difference between conditions for the Ehrenhofer stimuli.[7] In this section we will discuss how we can map the identification of these differences to the development of theories that can account for the Chow/Ehrenhofer contrast.

The pattern of difference observed in the subject-verb cosine relationship lends itself to a temptingly simple account: that the N400 amplitudes observed in these experiments are driven by the lexical relation between the target and the most recent content word (or noun phrase). Since this relation, when measured in the experimental stimuli, shows no difference between conditions for the Chow experiment and a reasonable difference for the Ehrenhofer experiment (in the observed direction, with higher cosine corresponding to lower N400 amplitude), this serves as a legitimate possible explanation.

Our other observed difference between the stimulus sets—the difference in average max cloze—does not pattern so straightforwardly with the observed N400 results. What additional components can be incorporated in order for this difference to form the basis of an account?

A promising direction for constructing such an account emerges when we consider the following fact: we have various forms of evidence suggesting that the extent to which the N400 reflects message-based expectations (which, as we discussed in Chapter 1, would be based on syntactically-constrained sentence meaning in combination with world knowledge) may be a function of the constraint of the context. This principle is implicit, for instance, in the original Federmeier and Kutas account of their result, which suggests that

[7] We also see stronger lexical relations between context and target for the Chow stimuli than for the Ehrenhofer stimuli—this difference will not drive any of the models described in this chapter, but it could motivate additional models in future work.

message-based prediction of the expected item influences facilitation of within-category targets—but only in high-constraint contexts.

This principle is also suggested by the Chow18-Delay role reversal result (remember that this is different from the Chow15-Reversal study that we have been comparing to the Ehrenhofer study): with the introduction of the pre-verb delay in that experiment, it is not *all* stimuli that show an N400 effect after a delay, but only stimuli in the "high-predictability" bin, as determined by the cloze of the expected/canonical items. To the extent that we understand this emergence of role-sensitivity in the N400 to be attributable to a more sophisticated expectation forming during the course of the delay, this suggests that such expectations may only form (or only do so strongly enough to show an effect) in more constraining contexts.

Two classes of possibility are suggested by this type of result. One obvious possibility is that message-based expectations can form given enough time, but they are only formed when the context is sufficiently constraining, as determined by some form of threshold. The Federmeier and Kutas high-constraint bin occupies a cloze range (for the expected target, thus comparable to our max cloze value) of .784-1.0, average .896. This is substantially higher than the majority of the max cloze values in the Chow and Ehrenhofer experiments, but it also comes from a very different type of stimulus, lending itself to greater constraint (two-sentence contexts are able to provide more description). By comparison, the "high-predictability" bin for the Chow18-Delay timing experiment occupies a cloze range (for canonical items, again comparable to max cloze) of .41-.97, average .64 . Since these stimuli are more similar to those used in the Chow15-Reversal and Ehrenhofer experiments, we might take this as our benchmark, and posit that contexts in a max

cloze range of .41 and higher can be expected to produce message-based expectations given enough time, while less constraining contexts may not.[8] The idea of a threshold of this nature gains further traction when we revisit the distributions of max cloze values shown in Figure 5.1, and note that it is in the .3-.4 max cloze range that we begin to see greater density in the distribution of values from the Ehrenhofer stimuli as compared to the Chow15-Reversal values.

Alternatively to this threshold-based theory, we might consider the possibility that message-based expectations are formed under all (or most) circumstances, but their influence is simply weaker for less constraining contexts, such that they may not make enough impact to be detected in N400 amplitude.

Both of these variants make the implicit suggestion that as a default, N400 amplitude may be driven (either primarily or exclusively) by lexically-based facilitation, *unless* the context is constraining enough, and there is enough intervening time, for the amplitude instead to be driven (primarily or exclusively) by message-based facilitation. This idea bears some similarity to the Kuperberg (2016) theory (and, by extension, to our hypothetical model above), in that it involves different cues driving N400 amplitude, depending on a measure of prediction strength. The accounts diverge in that Kuperberg assumes that all cues lead to an event prediction, and that cue dominance is based on event prediction strengths across all cues.[9] By contrast, our idea here does not require an assessment of prediction strength across all cue types—rather, we need only assess the extent to which message-based information gives rise to a strong expectation (perhaps equivalent to an

[8]Note that there is a range of cloze values that falls between the high and low predictability bins, which could potentially fall in either of these categories.

[9]Assessment of event prediction across all cues is not an explicit requirement in Kuperberg's description of the account, but it follows logically as a necessary component if the account is interpreted as described.

event prediction). In light of this, we additionally need not assume that lexical cues are leading to an active event prediction—we can alternatively consider lexically-based facilitation to arise in a more passive manner, as discussed in previous chapters.

Why would it be beneficial for the processor to operate in this way, forming message-based expectations only when the message is sufficiently constraining? Under the assumption that expectations or predictions are only of value when they accurately anticipate characteristics of upcoming inputs, it is sensible to suppose that in the case of an insufficiently constraining message, it simply would not be useful to form any kind of expectation based on that message. What use would there be, for instance, to form a prediction from the message-based interpretation of "A robin is not —"? The space of possibilities is simply too unconstrained to form expectations that will be of any value. In such cases, instead, the processor may simply fall back on more passive, automatic facilitation effects based on co-occurrence probabilities at the lexical level.

This also suggests an answer to the question of why the processor would benefit from a differentiation between lexical and message-based facilitation effects. Message-based cues will be more precise (and they are of course necessary for ultimate comprehension), but they may more often than not be unhelpful for the purpose of forming predictions (too slow, too unconstraining, or both). Instead, lexical co-occurrence statistics may be the most reliable overall information source for effectively facilitating incoming words.

So this leaves us with a number of possible accounts for the contrast in outcomes between the Chow and Ehrenhofer experiments.[10] It remains to be seen whether these

[10]The accounts that we discuss here are focused on the explanatory power of the stimulus characteristics themselves. Another possibility—which we will not explore in detail here, but which is worth considering—involves the difference in the stimulus distributions between the Chow15-Reversal and Ehrenhofer studies. As we discussed above, the Chow15-Reversal experiment includes high-cloze continuations

accounts will produce successful simulations of the two experiments, particularly when we consider the fine-grained dynamics of individual stimuli used in the experiments. This will be the purpose of the models and simulations described below, after we revisit and reincorporate considerations of the timing dimension.

5.7 Rethinking timing

Having established a foothold for explaining the contrast in outcomes between the Ehrenhofer (2018) and Chow15-Reversal experiments, it is now time to reincorporate the timing dimension introduced by the Chow18-Delay results, and to consider the implications of that dimension for the theories proposed in the previous section.

The Chow18-Delay timing result gives us strong reason to believe that for sufficiently predictive contexts, the N400 *can* reflect the effects of role reversals—if enough time is provided. The result relatedly suggests that there is a limitation on the speed with which these more sophisticated message-based effects can come into play (or at least a limitation on how quickly these effects can be reflected in the signal).

The Chow15-Reversal result, which shows no N400 effect for the role reversal, has accordingly been assumed by Chow et al. to fall in the too-early time window, thus reflecting the limitations observed in the Chow18-Delay study's no-delay condition. Of

for only one argument order from a given item, while Ehrenhofer includes high-cloze continuations for both argument orders. Although no individual participant will be exposed to these different argument orders or continuations for any given item (each item is presented in no more than one form for a given subject), it remains possible that the different stimulus distributions between the two experiments overall could somehow have influenced the nature of the processing of these stimuli. This is a question that we leave to future work.

course, with the sensitivity seen in the Ehrenhofer result, this assumption may be called into question.

Either way, this assumption is an important one for us to examine at this point—because if we are to propose any accounts of the Ehrenhofer (2018) and Chow15-Reversal N400 results that involve more sophisticated message-based mechanisms, then we need to reconcile this with the fact that such mechanisms are apparently only reflected in N400 amplitude if there is adequate delay between the final argument and the verb.

In order to reconcile this, let us look more carefully at the Chow18-Delay and Chow15-Reversal stimuli.

Chow18-Delay – no-delay condition
last week *police* BA *suspect arrest* ... (lit. "Last week the police arrested the suspect")
last week *suspect* BA *police arrest* ... (lit. "Last week the suspect arrested the police")
Chow18-Delay – with-delay condition
police BA *suspect* last week *arrest* ... (lit. "Last week the police arrested the suspect")
suspect BA *police* last week *arrest* ... (lit. "Last week the suspect arrested the police")
Chow15-Reversal
the restaurant owner forgot which *customer* the *waitress* had *served* ...
the restaurant owner forgot which *waitress* the *customer* had *served* ...

Table 5.8: Comparison of Chow18-Delay stimuli (no-delay and with-delay conditions) with Chow15-Reversal stimuli.

In the Chow18-Delay Mandarin stimuli, the no-delay contexts involve the target word falling immediately after the second argument, for an SOA of 600ms between the second argument and the verb. The with-delay stimuli introduce an additional intervening 1200ms, for a total SOA of 1800ms between onset of the second argument and onset of the verb.

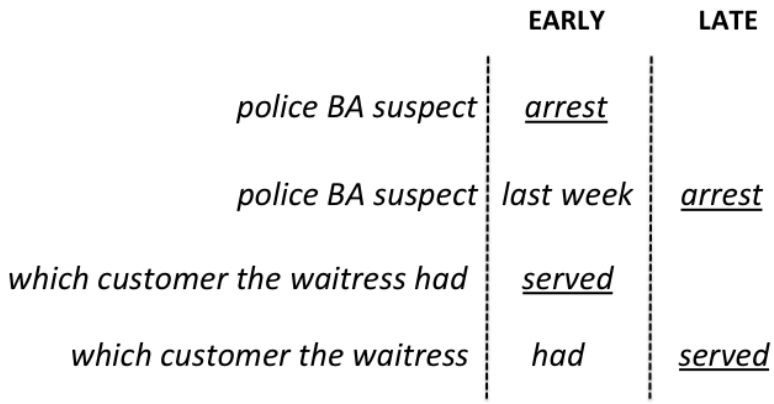

Figure 5.2: An illustration of the two possible construals of the timing in the Chow15-Reversal and Ehrenhofer experiments. The verb *served* may be construed to fall in the earlier time window associated with the no-delay condition of Chow18-Delay, or it may be construed to fall in the later time window associated with the with-delay condition.

The Chow15-Reversal stimuli, by contrast, do not place the verb immediately after the second argument, instead including the auxiliary "had" between the second argument and the verb—for an SOA of 1060ms between the onset of the second argument and the onset of the verb (530ms per word). The same applies to the Ehrenhofer stimuli.

The processing time for the Chow15-Reversal and Ehrenhofer experiments thus falls in a range intermediate between the Chow18-Delay no-delay and with-delay conditions. While we cannot guarantee that 1060ms is enough time for the relevant mechanisms to generate message-based facilitation, we hold that this longer SOA (in combination with the observation of a role-sensitive N400 at exactly this delay in the Ehrenhofer experiment) is adequate reason to believe that the intervening word may potentially introduce enough time for those mechanisms to come into play.

This leaves us free to entertain two different classes of possibility: 1) that the Chow15-Reversal and Ehrenhofer stimuli fall in the no-delay category, and as such only less sophisticated (e.g., lexical) mechanisms are at play in driving the results, as assumed by the

Chow15-Reversal account—or 2) that the Chow15-Reversal and Ehrenhofer stimuli fall in the with-delay category, in which case more message-based influences can be considered as a possible component in the resulting N400 amplitudes. These two possibilities are illustrated in Figure 5.2.

5.8 Modeling framework

The models that will be described in the following sections differ from the connectionist models of role reversals presented in previous sections, so in this section I will discuss the implications and motivations of the present modeling choices.

The question of primary interest to us here concerns the relationship—with respect to the generation of the N400—between the asyntactic lexical processes that we have been discussing in previous chapters, and the syntactically-constrained message-based processes that appear to be reflected in role-sensitive N400 results described in this chapter. In the modeling experiments reported here, rather than focusing on the specific mechanisms that constitute those processes, we will simply assume the existence of two distinct lexical and message-based processes (as we have done throughout this dissertation) and focus our modeling effort on understanding the dynamics of the interplay between them.

More specifically, in the modeling experiments presented below, we will assume the existence of these two types of processes, characterized by the types of information that they have access to and the types of effects that they can have as a result, and then we will quantify the facilitation generated by each of these processes, in order to model the tradeoff in contribution between them.

For the lexical processes, as in Chapter 3, we will be using cosine between vector representations to quantify facilitation arising from asyntactic lexical processes. As we have discussed in previous chapters, our mechanistic commitments are non-zero in using these vectors, given that the vectors are learned in a particular manner (even if their final content is fairly opaque), and more importantly that they are being combined in a way that has certain clear limitations—specifically, a way that is necessarily asyntactic. (This is true even if the vectors pick up on some statistical regularities of syntactic usage when they are learned. As we will discuss in greater detail in Chapter 6, regardless of the information contained in word vectors, the syntax of the source sentence itself cannot possibly be captured by an averaging procedure.) However, to a reasonable extent we are still not making very specific commitments about the mechanisms that output the facilitation levels that these cosine values represent—for instance, as we discuss in Chapter 2, we could imagine these averaging-based models to correspond either to a more passive priming-like mechanism, or alternatively to an active, lexically-based mechanism of plausibility assessment. In using cosines, we do imply certain assumptions about the way that facilitation is generated: as discussed in Chapter 2, the use of cosine is most straightforwardly compatible with facilitation based on feature overlap. However, it is likely that the cosine measure could correlate with the outcomes of other types of facilitation mechanism as well.

For our syntactically-informed message-based process, we will use the cloze probability of the target to quantify the relevant facilitation. This too involves non-zero assumptions: for one thing, we assume cloze (as an offline task performed by competent language comprehenders) to reflect access to both syntax and world knowledge, which character-

ize our message-based processes. Additionally, in using cloze to estimate facilitation, we make an assumption that the message-based process produces facilitation proportional to how frequently the target word is given as a response in the cloze task. (This is, of course, in line with the massive body of evidence showing the N400 to correlate with cloze probability.) Our use of this metric is suggestive of a probability-based mechanism of facilitation, as described in Chapter 2, by contrast to the overlap-based mechanism suggested by our use of cosine for the lexical facilitation. However, we are not in fact making strong assumptions about the use of probability distributions *per se*—cloze probability, like cosine, could plausibly correlate with the outcomes of other facilitation mechanisms.

Apart from the above assumptions, our use of cloze values allows us to remain particularly uncommitted about the specific mechanisms that give rise to this message-based facilitation—given that cloze values are produced in black-box fashion based on the responses of human participants. A key motivation for this approach is that at the present time, there are no existing computational models able to approximate the use of syntactic composition and world knowledge underlying the sophisticated and fine-grained assessments that humans make in a cloze task (and that we expect to be reflected in our message-based facilitation). We saw an example of this type of limitation in the Brouwer model in Chapter 4, and we will explore this problem in greater detail with more state-of-the-art models in Chapter 6—but for now, suffice to say that the values that would come from existing models would fall dramatically short of human cloze values in terms of nuance and accuracy. For this reason, we derive a significant advantage in using cloze values: though we sidestep commitments about the specifics of the mechanisms that produce these values, we gain significant item-level accuracy.

Note that this combination of corpus-based and human-based quantification of stimulus properties represents a hybrid modeling approach, which to our knowledge is a novel method for addressing these kinds of processing questions.

In what specific ways does our modeling approach differ from those of the connectionist models above? Obviously, since we are focusing on the interactions of the facilitations output by these processes rather than on the internal details of the processes themselves, we differ from the connectionist models in avoiding commitments to mechanistic specifics. In the remainder of this section, I will highlight three advantages of our use of this approach.

Zeroing in on relative contributions Our priority in these simulations is to zero in on the interactive dynamics and relative contributions of two classes of process that appear to be reflected in the N400. Because the types of connectionist models used by Rabovsky et al. (2016) and Brouwer et al. (2017) are largely opaque with respect to the relative contributions of different factors in producing the final outcome (much like the brain itself), these models are not well-suited to asking this kind of question. It is possible (and reasonably likely) that these connectionist models do pick up on and make use of a combination of lexical, syntactic, and "world knowledge" (training data probability) information—but we cannot straightforwardly determine in what specific ways these different types of information are being used and combined to yield the final result. By breaking things down into two independent sources of facilitation, quantifying the facilitation from each of these sources, and testing different hypotheses about how these might interactively

contribute to N400 amplitude, we allow ourselves to ask a more targeted question about this interaction.

Use of real data While there is certainly value in using synthetically-generated data to demonstrate how mechanisms embodied by a model interact with a well-defined and controlled probabilistic environment, the fine-grained complexities of real stimuli mean that even if the probabilistic properties built into the synthetic environment accurately reflect the corresponding properties in the real stimuli (which is unlikely, given that synthetic data tends to have greater uniformity than we can expect from real stimuli), we still have the problem that real stimuli have a host of other properties that may influence results, but that are not part of the intended experimental manipulations that the synthetic data reflect. (This is clear from our analysis in Section 5.5.)

So while synthetic data are useful for understanding how mechanisms respond to particular circumstances, this kind of method must necessarily be complemented by methods that make use of the real experimental stimuli, and that draw on accurate quantifications of relevant characteristics that these stimuli have.

This is the tradeoff that we make in using the present modeling method. By using corpus-derived vectors to quantify lexical relations, we are able to draw on real-world statistics to estimate the lexical relations at play for each individual stimulus used in the Chow15-Reversal and Ehrenhofer experiments. This allows us to pick up on unintended or unintuitive lexical idiosyncrasies of the datasets—a perfect example of this being the difference that emerges when we quantify the lexical relation of the subject to the verb in Section 5.5. Similarly, by using item-by-item cloze values, we can quantify strength

of message-based expectations at a much finer-grained level than the binary distinction of high- versus low-cloze, drawing on the power of real human expectations to derive the estimates.

One can of course question just *how* accurate these real-world quantifications of the item characteristics are—certainly, any estimates are bound to be imperfect. However, given that we draw on the statistics of large corpora (in the case of the lexical relation quantification) and the predictive behaviors of human respondents (in the case of quantifying message-based facilitation) to estimate these characteristics, we have good reason to believe that these methods should capture the relevant characteristics at least to a respectable extent.

Testing of different models with transparent connection to hypotheses In the case of Rabovsky et al. (2016) and Brouwer et al. (2017), the authors present a single model as proof-of-concept, showing the relevant model to be successful in simulating a set of results. There are two limitations here, related to the limitations discussed in Chapter 4. First, presentation of a single working model does not allow us to impose much in the way of partition on the hypothesis space—there are bound to be a great many models that can account for a given set of results, so our goal should be to test a variety of models such that we can begin drawing boundaries between hypotheses that can and cannot be viable accounts. Here we will take a step in this direction by contrasting a number of different models and corresponding hypotheses.

Second, although there is, of course, no principled limitation that prevents connectionist models from presenting contrasting variations and describing which ones work

and which do not (Brouwer et al. (2017) do in fact include a list of failed models in the appendix), the challenge comes in the straightforwardness of the mapping between a given neural network and a corresponding cognitive hypothesis. As we discuss and illustrate in Chapter 4, the opacity of internal computations and representations within neural network models mean that it is often challenging to map clearly between distinctions relevant to the network's performance, and corresponding distinctions in cognitive hypotheses.

By comparison, although the models here certainly involve unknowns (e.g., the specific content of the vector representations), the variations that we employ to differentiate between models map in a straightforward manner to the contrasts between hypotheses that we wish to test. In this way, the use of these simpler models affords us greater ability to partition the cognitive hypothesis space—and in doing so to extract cognitive insights—based on the models' performance.

5.9 Hypotheses/models

Having established the framework within which we are working, we will now lay out the different models that we test in the present simulations. We divide our models into two hypothesis categories: those that operate solely on the basis of lexical relations, and those that have access to message-based (syntactic and world knowledge) information. While each of these models differs in non-trivial ways from the formulations of previously existing accounts, this division aligns somewhat with the distinction between theories like that of Chow et al. (2015), who posit that these role reversal results reflect processing

based on limited information (in our case, lexical relations only), and theories like that of Kuperberg (2016), which posit that all levels of information can be in play in influencing these results.

5.9.1 Lexical-only hypotheses

For the lexical category of models, we consider two hypotheses, based directly on the corresponding lexical measures discussed above. Here we reiterate the hypotheses that these measures suggest, and define the models that will be used in the full experimental simulations below.

Decaying-context relation This hypothesis holds that N400 facilitation is based on the aggregate lexical relation of the content words in the context with the target word. In its simplest form, where all content words have equal weight, this hypothesis is a strawman with respect to the Ehrenhofer (and Chow et al. (2018)) results—the preceding content words are identical between the high- and low-cloze conditions. However, with a linear-distance-based decay of the kind used in Chapter 3, this is a reasonable account to test (if unlikely to succeed based on our analyses in Section 5.5). The decay-based version of this hypothesis is therefore the one that we simulate here: the hypothesis that N400 facilitation reflects aggregate lexical effects of the most informative (least frequent) words in the context, weighted by recency. To model this, a weighted average vector is constructed for the context as in Chapter 3, and the cosine of this vector with the target word vector is computed.

Final-word relation This hypothesis holds that N400 facilitation reflects lexical relation effects between the most recent content word (or phrase, as in the case of e.g., "restaurant owner") and the target. We implement this hypothesis simply by taking the cosine of the final content word/phrase (the subject, in the case of these stimuli) with the target word.

The decaying-context relation and final-word relation hypotheses could be thought of as endpoints on a continuum: if we implement a sharp enough decay in the effects of the context words, then eventually we will arrive at a scenario in which the final content word is fully dominating the signal. So to the extent that the final-word hypothesis is successful, it could support either a categorical limitation of the lexical effect to the most recent content word, or it could support a sharper decay of lexical effects.

This class of hypotheses is consistent with an account of the Chow/Ehrenhofer results which disallows access to message-based processes, rather reflecting simpler and more limited lexical processes.

5.9.2 Combination hypotheses

For the second category, which incorporates message-based facilitation, we also consider two hypotheses. In both of these categories, we consider N400 facilitation to reflect an interplay of lexical- and message-based facilitation. Both of these hypotheses, as discussed above, build on the intuition that message-based expectations might vary in influence based on how constraining the context is.

Graded combination This hypothesis holds that N400 facilitation reflects a weighted average of message-based and lexically-based facilitation, with weighting determined by the level of constraint of the context. Specifically, the greater the max cloze (constraint) of the context, the greater the contribution of the message-based facilitation as quantified by target cloze. The lower the max cloze of the context, the greater the contribution of the lexical facilitation. This is formalized in the following way:

$$\text{facilitation} = (\text{max cloze})(\text{target cloze}) + (1 - \text{max cloze})(\text{context cosine})$$

where "context cosine" refers to the cosine between the target and the unweighted average vector of the informative context words.

To illustrate with an example: consider the context "the superintendent overheard which tenant the landlord had —", which has a max cloze value of .63 (corresponding to the word *evicted*).

If the target word is *evicted*, for which the target cloze in this context is .63 (same as the max cloze) and the cosine of context and target is .38, then by the formula above, we will calculate our facilitation value as (.63)(.63) + (.37)(.38) = .54.

If the target is *complained about*, for which the target cloze in this context is 0 and the cosine of context and target is .25, then the facilitation for this target is accordingly (.63)(0) + (.37)(.25) = .09.

In this particular example, because the max cloze of the relevant context is high, the relative contribution of the message-based facilitation component is also high.

The intuition behind this hypothesis is that message-based facilitation is contributing at all times, but it contributes more weakly when the context is less constraining.

Threshold-based selection This hypothesis holds that N400 facilitation reflects *either* lexical facilitation *or* message-based facilitation at a given time, with the contributing mechanism decided based on a context constraint threshold. The intuition behind this hypothesis is that below a certain constraint threshold, no message-driven expectations will be generated, and the facilitation will be driven solely by the lexically-based activations. By contrast, if the constraint threshold is exceeded, a message-based expectation will be generated, at which point the facilitation will be driven solely by the fit of the target to that expectation. We model this with a max cloze threshold of .4, in accordance with the discussion in Section 5.6—if max cloze is less than .4, then the facilitation value is taken to be the cosine between the average (unweighted) context vector and the target vector. If max cloze is greater than or equal to .4, then the facilitation value is taken to be the target cloze value.

This final hypothesis bears the strongest resemblance to our above interpretation of the Kuperberg (2016) account, in which facilitation is based on a categorical selection of cues according to a measure of cue strength.

5.9.3 Mixing cloze and cosine

Both of the latter two models involve mixing of cloze and cosine values for the final average estimates of facilitation. These are very different measures, so this is an issue that deserves some attention.

For the purpose of these simulations I have opted to do minimal transformation on these values prior to combining them—specifically, I simply clip cosine values at zero (a very small number of the original cosine values fall just below zero), such that both of the measures are distributed in a range between 0 and 1.

We can compare the actual distributions of these values for the stimuli—Figures 5.3 and 5.4 show the scatter plots of max cloze (which is equivalent to target cloze for any targets that will receive non-zero cloze-based facilitation) against context-to-target cosine, for the stimuli from the Ehrenhofer (2018) and Chow15-Reversal experiments respectively.[11]

In the case of the Ehrenhofer stimuli, we see a reasonably comparable distribution of values, though the cloze values trend lower than cosine. In the case of the Chow et al. stimuli, consistent with the observation of lower overall cloze values in that experiment, the cloze values cluster more substantially lower relative to the cosine values.

Though the distributions of the cloze and cosine values for these stimuli are not identical, the fact of the matter is that we lack any well-motivated *a priori* hypothesis about how one would shift the distributions of these measures, or otherwise transform the values, such that they would better reflect how the corresponding sources of facilitation can be expected to differentially contribute to overall facilitation in reality. This is a useful direction for future work, but for the current purposes, I opt to mix the values with only the minimal transformation such that they vary in the same range.

[11]Note that Chow et al. have half as many datapoints, due to Ehrenhofer's inclusion of symmetric continuations. This means that Ehrenhofer has twice as many distinct stimuli—however, the experiments consist of the same overall numbers of trials, as do our simulations below.

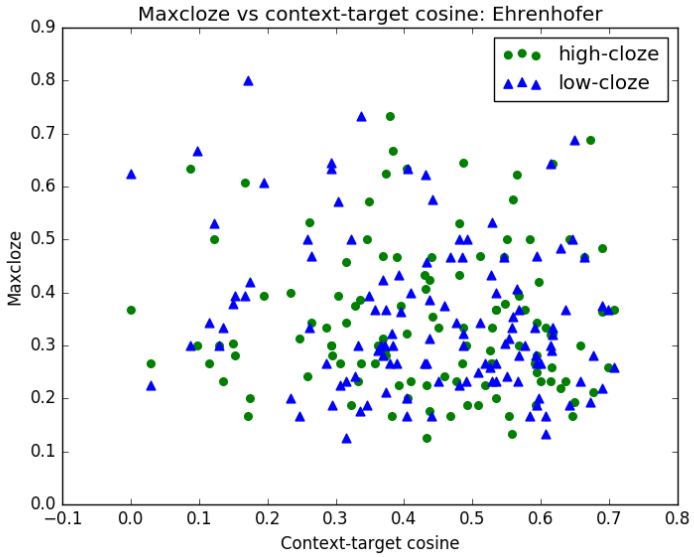

Figure 5.3: Scatter plot of max cloze against context-to-target cosine for all stimuli: Ehrenhofer (2018)

5.10 Simulations

To simulate the Chow15-Reversal and Ehrenhofer (2018) experiments based on these hypotheses, we run full simulated experiments, each with 24 simulated subjects (the sample size of the Chow and Ehrenhofer experiments), dividing the stimuli into lists as done by Chow et al. and Ehrenhofer, such that a given subject sees 30 items from each of the Chow et al. conditions, or 15 items from each of the Ehrenhofer conditions. (This results in the same number of trials across the two experiments, since Ehrenhofer has twice as many stimuli due to use of symmetric pairings.) The goal of running the simulation in this way is to allow for assessment of significance by subjects, as is reported by Chow et al. and Ehrenhofer.

To simulate variation between subjects, we introduce a small amount of noise to the average simulated N400 value for each subject. Not having a strong *a priori* assumption

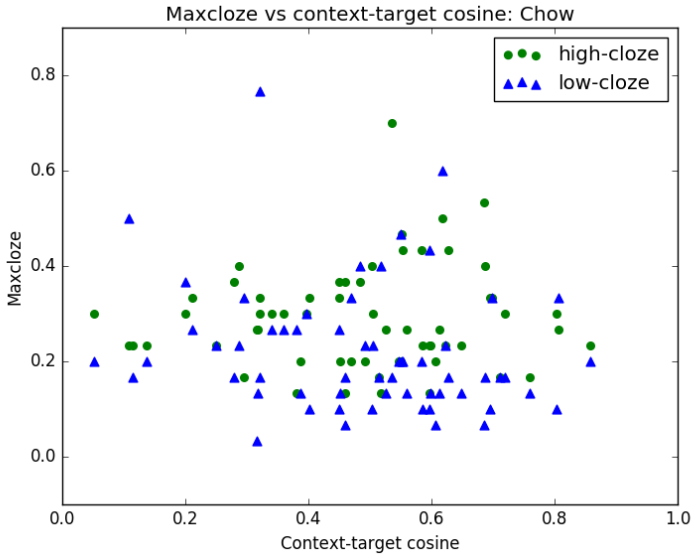

Figure 5.4: Scatter plot of max cloze against context-to-target cosine for all stimuli: Chow, Smith, Lau, and Phillips (2015)

with respect to how subjects will vary, and for the practical purpose of limiting the extent of variation between runs of the experiments, we opt to restrict noise such that it should only reduce the size of the N400 effect, rather than spuriously increasing it. To this end, for each subject and each cloze condition (high- or low-cloze) a noise term is drawn from a normal distribution, but is subjected to a number of constraints to restrict variability.[12]

The practical value of this particular noise model is that it generates a small amount of variation between subjects, mitigating inflation of significance values, but it also restricts noise sufficiently to ensure consistency between runs, and to ensure that variation between conditions is driven primarily by the stimulus characteristics identified above. However, it is important to note that our use of this noise model—and our interpretation of the resulting significance values—makes some fairly strong assumptions about the nature of

[12]The noise terms were clipped at zero such that they could only increase low-cloze facilitation or decrease high-cloze facilitation, and they were also capped so as not to exceed half of the total difference between the averages for a given subject.

variation between subjects, which may not align with actual subject variation. In future work, with access to trial-level data from these experiments, an ideal alternative will be to compute significance by items in order to avoid the need for modeling subject variation in this way.

There are few free parameters in these models. The decaying-context model includes a parameter for the number of context words to include in the weighted context average, for which we choose 4 (this is consistent with the Federmeier and Kutas simulations from Chapter 3, and essentially selects out all of the content words for the Ehrenhofer and Chow15-Reversal stimuli). The threshold model includes the threshold parameter, which we set to .4 based on the Chow18-Delay result, as discussed above.

In all of the simulations below, N400 amplitude is represented (inversely) by the facilitation values produced by the different models, as described in Section 5.9.

5.10.1 Chow/Ehrenhofer simulation results

Results are shown in Figures 5.5 and 5.6. Figure 5.5 shows the raw simulated N400 amplitudes for high- and low-cloze conditions across the two different experiments (in this figure the y-axis is inverted to show correspondence of higher facilitation values to lower N400 amplitudes), while Figure 5.6 shows the differences between cloze conditions, with 95% confidence intervals. In the latter plots, zero is indicated by a dotted horizontal line to facilitate the assessment of whether the confidence intervals for the difference overlap with zero. Statistical significance is determined by paired-sample t-tests.

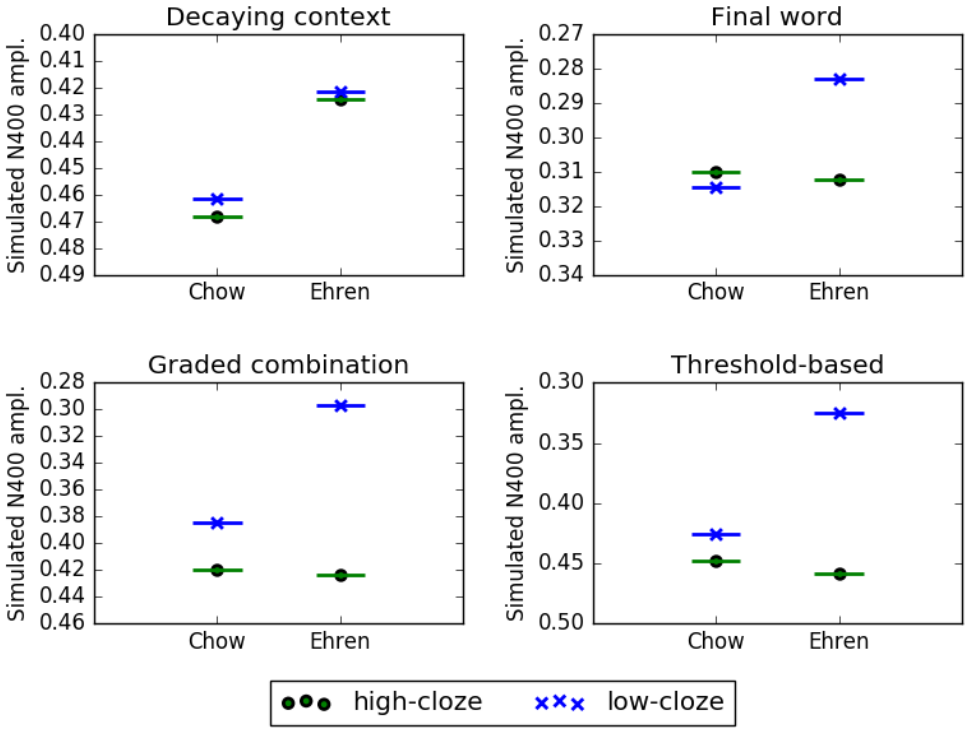

Figure 5.5: Simulated N400 amplitudes for Chow et al. (2015) and Ehrenhofer (2018) experiments under four hypotheses. y-axis is inverted to show correspondence of higher facilitation values to lower N400 amplitudes.

The context relation model, despite the decay with linear distance, fails to simulate the observed results, with both simulated experiments instead showing no significant difference (Chow: t(23) = .416; p = .681, Ehrenhofer: t(23) = .243; p =.810), corresponding to no N400 effect for either experiment.

By contrast, the last-word relation model does successfully simulate the observed results, with no significant difference for Chow (t(23) = -.596; p =.557) but a just-significant difference for Ehrenhofer (t(23) = 2.26; p = .034), corresponding to a significant N400 effect for Ehrenhofer but not for Chow.

Moving to our combination hypotheses: the gradient prediction model fails to simulate the observed results, with both simulated experiments now showing a significant

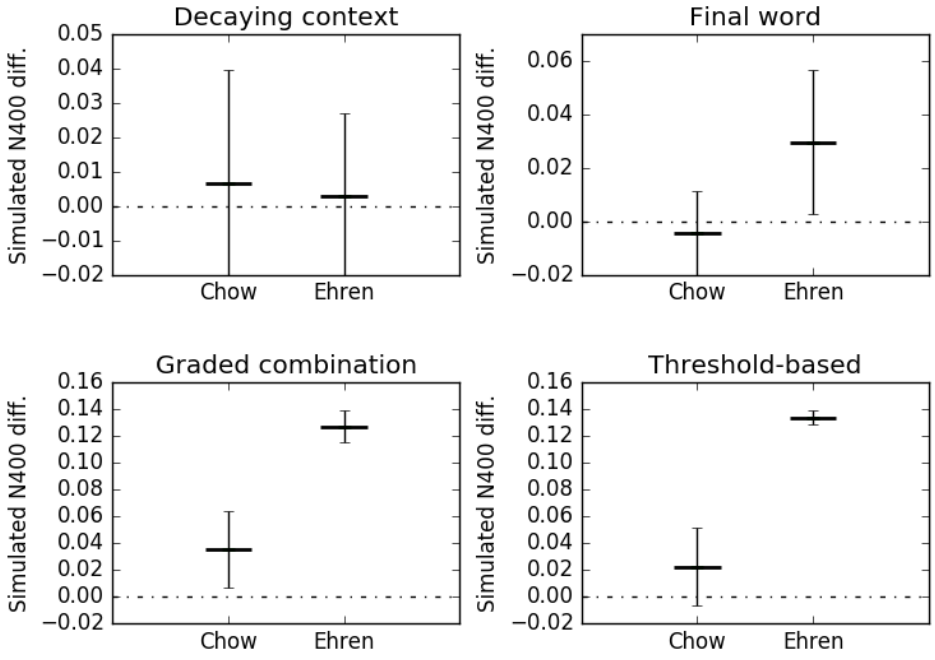

Figure 5.6: Simulated N400 amplitude differences for Chow et al. (2015), and Ehrenhofer (2018) experiments under four hypotheses, with 95% confidence intervals

difference (Chow: t(23) = 2.53; p = .019, Ehrenhofer: t(23) = 21.87; p < .0001), corresponding to an N400 effect for both experiments.

Finally, our thresholded prediction hypothesis does successfully simulate the observed results, with no significant difference for Chow et al. (t(23) = 1.58; p =.127) but a highly significant difference for Ehrenhofer (t(23) = 53.92; p < .0001) corresponding to an N400 effect for Ehrenhofer but not for Chow.

5.10.2 Interim discussion

The above simulations indicate that two of our hypotheses are able to account for the contrast observed between the Chow and Ehrenhofer experiments: an account in which

N400 amplitude reflects only the lexical relation between the most recent content word and the target, and an account in which N400 amplitude reflects either lexical relations or message-based expectations, depending on whether the constraint of the context meets a certain threshold (.4 max cloze in the current model). This means that we have a successful representative from each category of hypothesis: one purely lexical, and one that incorporates message-based expectations.

The results of these experiments are, as expected, largely consistent with the differences observed in Section 5.5 that motivated these models. However, these simulations allow us now a) to make distinctions in terms of statistical significance, based on the sample sizes and experimental settings used by Chow et al. and Ehrenhofer, and b) to identify a failure of the gradient prediction model to distinguish between the studies, by contrast to the thresholded prediction model. (The failure of the context relation model, though worth confirming, is less surprising.) Importantly, again, this partitioning of models based on statistical significance relies on the assumptions made by our current model for variation between subjects, and the partitioning may change with a different model of noise. However, given our current assumptions with respect to subject variation, we can conclude that two of our models are able to account for the observed results, while the other two do not.

As we have discussed above, there are bound to be a host of models that can account for a given result—our job is to partition the hypothesis space to the best of our ability, and then to test how well the successful hypotheses generalize to further results. To this end, in the next section we will add to our simulations the stimuli from one additional experiment, in order to take a further step in testing the generalizability of the hypotheses.

5.10.3 Nouns as targets (Ehrenhofer (2018))

The third experiment that we simulate is an experiment also run by Ehrenhofer (2018), in tandem with the role reversal experiment described above. This experiment aimed to test N400 sensitivity to anomaly under circumstances in which the verb is located in the preceding context, with the object noun serving as the target word. For this purpose, the experiment (which we label as "NVN") uses stimuli with subject-relative clauses, such as "The queen wondered which maid had dressed the princess ...".

Ehrenhofer reports a significant N400 effect between high- and low-cloze conditions for this experiment as well. We obtained the stimulus set and cloze values for this experiment, allowing us to compute the relevant item-level facilitation values and test the generalizability of our hypotheses to this result.[13]

5.10.4 NVN simulation results

Figures 5.7 and 5.8 show the results of the simulations, with the noun-prediction "NVN" experiment now added alongside the role reversal results.

Interestingly, both the final-word hypothesis and the threshold-based hypothesis successfully predict a significant N400 effect for this experiment as well (FW: t(23) = 2.72; $p = .012$, TP: t(23) = 7.79; $p < .0001$). The gradient prediction hypothesis accurately predicts a significant effect for this third experiment (but, as we saw above, fails to account for the lack of effect in Chow et al.) (t(23) = 17.46; p <.0001), while the context

[13] We thank Lara Ehrenhofer for provision of stimuli and cloze values, and for extensive discussion of the relevant experiments.

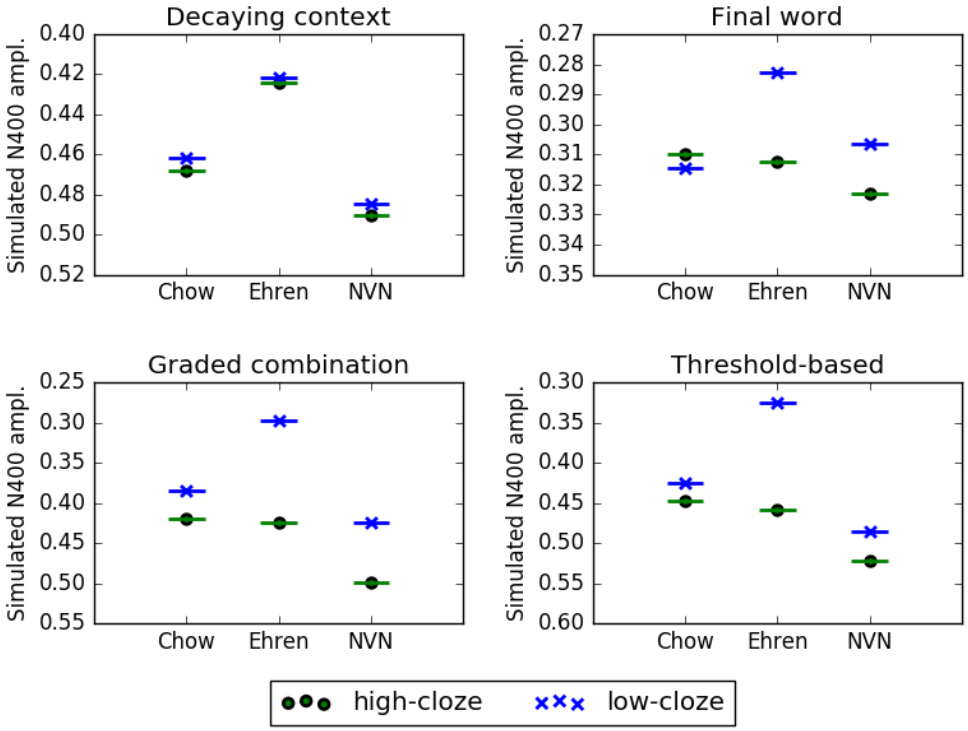

Figure 5.7: Simulated N400 amplitudes for Chow et al. (2015), and Ehrenhofer (2018) role reversal and NVN experiments, under four hypotheses. y-axis is inverted to show correspondence of higher facilitation values to lower N400 amplitudes.

relation hypothesis incorrectly predicts a lack of effect for this experiment ($t(23) = .829$; $p = .416$).

5.11 Discussion

Two hypotheses prove able to account for the distinction between the Chow and Ehrenhofer role reversal experiments, while also successfully predicting a significant N400 effect for a third experiment (which was not involved in the analyses that inspired the hypotheses in the first place). The conclusions for these simulations assume a particular model of variation between subjects which influences the significance values for our

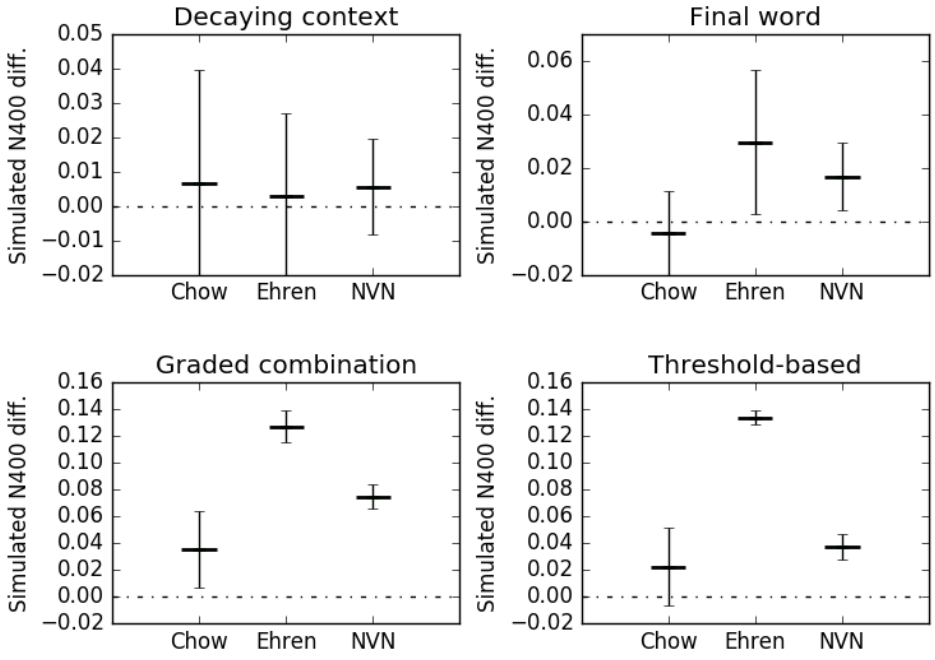

Figure 5.8: Simulated N400 amplitude differences for Chow et al. (2015), and Ehrenhofer (2018) role reversal and NVN experiments, under four hypotheses, with 95% confidence intervals

simulated effects—while this model of variation may change with more sophisticated understanding of variation between subjects, it allows us for now to draw conclusions about the adequacy of our hypotheses given the assumptions of that variation model.

These simulations support the conclusion that the contrast in role reversal results between Chow15-Reversal and Ehrenhofer (2018) can be accounted for by two different classes of model: one in which the N400 is driven by simple lexical relations between the most recent content word and the target word, and one in which the N400 reflects alternate influences of lexical relations and message-based expectations, depending on whether the strength of the latter expectations exceeds a given threshold.

How can we further adjudicate between these models? The simplest answer is that we can further expand to simulation of a broader range of results. Attempting to extend these models to account for additional results based on other stimulus sets is sure to force further elimination, or at very least adjustment, among the models in order to increase generalizability. Particularly of note are some of our less principled modeling choices (like the mixing of cloze and cosine, or the model of subject variance) for which we may identify better methods as we proceed—with adjustments to these modeling decisions and simulation settings, we may see changes in the partitioning of the models. In particular, it is clear that the gradient prediction model is only significant by a small amount for the Chow et al. experiment, which suggests that certain adjustments might shift it so as to show the correct pattern. In preliminary exploration of different settings, we found no condition under which that model produces a successful simulation of the results—but we acknowledge that we may not want to consider the corresponding hypothesis to be off the table.

In the meantime, we will subjectively compare our currently successful accounts, by considering their general plausibility, and likely generalizability, relative to other results reviewed above.

Considering first the last-word hypothesis, one might quickly object that although this account proves able to simulate N400 results for the three studies modeled here, it seems a tenuous proposition that the relation between the most recent content word and the target word should be able to account for N400 amplitude across the full range of reported N400 results in the literature.

Critically, it is not *all* N400 results that this relation would need to account for, as this hypothesis operates within the larger context of the Chow18-Delay timing result: it assumes that the Chow15-Reversal and Ehrenhofer (2018) stimuli fall within the no-delay category from Chow18-Delay, and that as a result the N400 to the verb in those experiments reflects the more limited processes of the "early" processing time window. Consequently, this hypothesis does predict that the N400 will be driven by the lexical relation between the most recent content word and the target word, but *only* in cases where those items fall within one intervening word of one another.

For some of the studies cited above, such as Hoeks et al.(2004), this property would apply: as we see in Table 4.1, the target verb falls immediately after the second argument in these stimuli. In this particular case, it appears that this simple pairwise relation may indeed hold—with the N400 effect being absent for stimuli that have adjacent pairings such as "athletes thrown", and present for stimuli that have adjacent pairings like "athletes summarized". As another example, the Kim and Osterhout (2005) study has a single word intervening between the argument and the verb, as in the Chow and Ehrenhofer studies, placing it in the gray zone discussed above—and if the last-word hypothesis is to account for the Chow and Ehrenhofer results, then it must be expected to extend to other studies in which the argument and verb are separated by a single intervening word. By this logic, then, the last-word hypothesis would predict that the lexical relation between the last argument and the verb in Kim and Osterhout's attraction condition ("meal ... devouring") and passive condition ("meal ... devoured"), should be stronger than the lexical relation between those items in the non-attraction condition ("tabletops ... devouring"). Given the

The superintendent overheard which tenant the landlord had evicted ...
The superintendent overheard which realtor the landlord had evicted ...

Table 5.9: Argument substitution contrast, Chow et al. (2018)

fact that the attraction condition is explicitly constructed to generate semantic attraction between the argument and the verb, this does seem plausible.

By contrast, in experiments such as Kuperberg et al. (2003), where the argument and verb are separated by multiple words, we might reasonably assume that the verb in this circumstance falls in the Chow18-Delay "late" processing window, in which case the N400 would be subject to different influences, which this specific hypothesis does not address.

Although the final-word hypothesis shows promise for generalizing to the classic role reversal results, this hypothesis encounters problems when we expand beyond the role reversal literature. In particular, if we consider the argument substitution contrast tested by Chow et al. (2015), reviewed in Chapter 4 and reiterated in Table 5.9, we see that this is a distinction that involves identical targets and final nouns across high- and low-cloze conditions—but it produces a significant difference in N400 amplitudes between these conditions. Given that the final-word hypothesis would predict no N400 difference for this contrast, this result suggests trouble for this hypothesis.

One potential adjustment to the last-word hypothesis, which could potentially afford it greater flexibility to account for such results, would be the alternative discussed above, in which the dominating influence of the last word is not categorical, but continuous—the rate of decay would be sharp enough such that the dominating influence of the final word

would persist, but that influence could be modulated somewhat by weak influences of more distant context words.[14]

Moving now to our other successful model, the thresholded prediction model, we can ask the same question: does this model seem likely to generalize? By contrast to the last-word hypothesis, the thresholded prediction hypothesis assumes that a single intervening word is indeed enough time for more sophisticated message-based effects to emerge, and that the Chow15-Reversal and Ehrenhofer (2018) results thus reflect an interplay between more and less sophisticated facilitation. It furthermore assumes that more sophisticated facilitation patterns will come into play only when the strength of message-based expectation exceeds a certain threshold—otherwise the signal will reflect lexical facilitation.

This hypothesis predicts that for experiments like Kuperberg et al. (2003) and Kim and Osterhout (2005), where intervening words give the opportunity for message-based expectations to form, the lack of N400 effect would be due to too few of the experimental items exceeding the constraint (max cloze) threshold, resulting in lexical facilitation dominating the signal instead. This seems in principle a plausible scenario—but in order to test this prediction, we need access to the cloze numbers for the stimuli in those studies. This will be a direction for future work.

[14] An additional possibility that should be considered with respect to this hypothesis is that, to the extent that influence of the final content word is indeed privileged in the Chow and Ehrenhofer role reversal results, this is not because it is the most recent content word, but because it is the subject. This would be consistent with the claim of Chow et al. (2015) that the role-insensitive effects are not fully asyntactic, in that they have access to the parse and the identification of the verb's arguments. Importantly, in our simulations above, we are in fact testing a hypothesis of the last content word *per se*, due to the fact that in the NVN study we are measuring the relation between the verb and the target object. But this does not rule out the possibility that a better hypothesis is one in which effects in the role reversal experiments are driven by the last word's status as subject.

Finally, let us now sketch out a rough picture of the mechanisms that these hypotheses envision. Both hypotheses operate under the assumption that there is an evolution over time in the influences that are at play with the N400. The last-word hypothesis makes no claims with respect to what drives facilitation in the later processing time window, while the thresholded prediction hypothesis claims that these results *do* reflect the later time window, and that this later time window involves alternating influences of lexical and message-based processes.

This picture is consistent with the existence of parallel processes set in motion by the processing of the sentence: one fast, automatic, and lexical, and the other slower, syntactically-constrained, and having access to world knowledge. A visualization of such a parallel architecture is shown in Figure 5.9. This conceptualization envisions both lexical and syntactic information becoming available quickly, with the delay in message-based effects being accounted for by the additional computation step involved in consulting world knowledge. In this example, because the max cloze item falls below the .4 threshold, the facilitation would come from the lexical pathway—which we must therefore assume would carry forward in time, and not be limited to the early time window.

Are these hypotheses also consistent with an alternative single-stream architecture? The final-word hypothesis surely is—it accounts for the N400 effects while making no claims about the later part of the processing evolution, so we can imagine that this later part could evolve directly as a later stage of the lexical process. As for the thresholded prediction hypothesis, this is perhaps less consistent with a single-stream hypothesis, given that it assumes lexical effects to remain in play even after syntax has been computed, and world knowledge consulted, to determine message-based expectation levels.

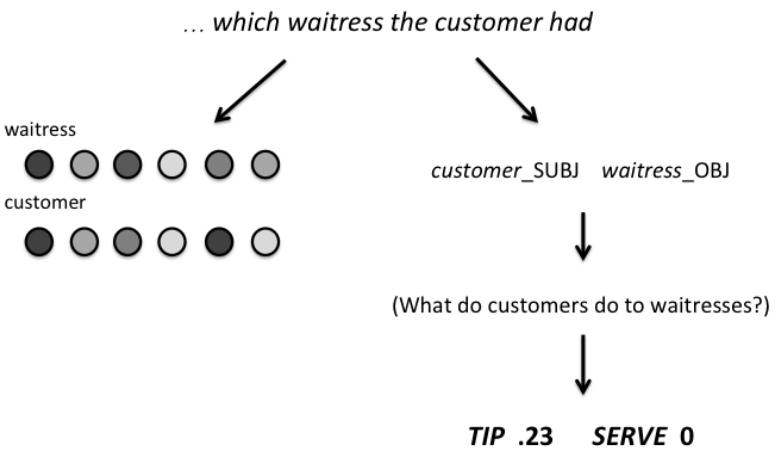

Figure 5.9: A hypothetical parallel architecture in which lexical facilitation is available quickly and automatically, while syntax- and world-knowledge-based facilitation happens more slowly and strategically.

In sum, these simulations provide us with evidence in favor of two promising models that can allow us to account for the complex pattern of results in studies from Chow15-Reversal (among other role reversal results), Ehrenhofer (2018), and Chow18-Delay, by contrast to two related models which under our current assumptions do not. It is of course critical that we acknowledge these models to represent only a small number among the innumerable possible hypotheses in our hypothesis space—but these simulations allow us a foothold in partitioning that space based on the results of interest.

5.12 Future directions

The next steps for this work will involve two critical components. First, we will expand the testing of models to additional results, in order to allow for further refinement of the models and adjustment of the hypotheses (including, as needed, formulation of new hypotheses that allow for broader generalization). As an immediate next step, we will

test the generalization of the models to the Chow et al. (2015) argument substitution contrast, which may allow us to distinguish between our final-word and threshold-based hypotheses—while the final-word hypothesis is anticipated to fail on this contrast, as discussed in the previous section, the threshold-based hypothesis could very plausibly succeed. If both models fail, then this result will provide a useful testing ground for forming additional hypotheses for testing.

The second immediate direction for this work is to shift to testing significance effects by items rather than by subjects. Because we do not have a strong *a priori* theory of subject variation, the necessity of a subject variation model is a disadvantage of assessing significance by subjects. In immediate future work, we will obtain trial-level data from these experiments, in order to test our hypotheses without the need for a subject variation model.

5.13 Conclusion

In this chapter we have discussed new results that complicate the picture of the N400 relative to the role reversal paradigm, along with a computational simulation demonstrating how one existing account might potentially account for these new results. We then moved on to examining the actual stimuli used in the relevant experiments, allowing us to identify characteristics that could underlie the contrasting results. Based on these analyses, we proposed several new hypotheses, and corresponding models, to account for these results while taking into account the complex fine-grained dynamics of the individual stimuli. With the modeling framework that we use to address these questions, we introduce a

novel hybrid approach which incorporates both corpus-based and human-based quantification of stimulus characteristics. The results of the simulations testing these hypotheses suggest two viable accounts able to explain the observed pattern of results: one in which the N400 reflects lexical relations between the target and the most recent content word, and one in which the N400 reflects a trade-off between lexical facilitation and message-based expectations, depending on contextual constraint. We discussed the potential for further generalization of these hypotheses, as well as the alignment of these hypotheses with the notion of parallel versus single-stream processing architectures.

Chapter 6

Modeling composition of complex lexical content

[Section 6.4 in this chapter is adapted from Ettinger, A., Elgohary, A., & Resnik, P. (2016). Probing for semantic evidence of composition by means of simple classification tasks. *Proceedings of the First Workshop on Evaluating Vector Space Representations for NLP, ACL 2016*.]

[Section 6.5 in this chapter is adapted from Ettinger, A., Elgohary, A., Phillips, C., & Resnik, P. (2018). Assessing Composition in Sentence Vector Representations. *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*.]

6.1 Introduction

In the previous chapters, our discussion of the relationship between lexical and syntactic processes has focused on a distinction between asyntactic lexical effects that we see reflected in real-time language comprehension in humans, and the syntactically-constrained processes that give rise to interpretation of sentence meaning.

We have set aside until now the question of how lexical and syntactic mechanisms relate *within* the context of the composition process that gives rise to sentence meaning. In particular, our use of complex vector-based word representations lends itself to the following question: how does complex lexical content interface with the composition system to produce complex sentence meaning?

This will be the focus of the present chapter. In particular, we will focus on the problem of sentence composition in natural language processing (NLP) systems, for which the problem of composing complex lexical representations is particularly salient. This will represent a shift from our previous focus on modeling language comprehension in humans, to a focus on engineering of language processing in machines—however, the underlying question of how lexical content should be represented, and how compositional operations should interface with that content, is one that is also of fundamental importance to our understanding of language in humans.

It is important at this point that we clarify our use of the term "composition". On one hand, composition can mean any function that takes the component words of a phrase or sentence and produces an output—it is in this sense that the various models that perform this kind of operation can all be called "composition" models. However, we of course

cannot be satisfied with just any composition function: we want composition functions to produce outputs that reflect the *meaning* of the sentence. So although we will refer to a variety of NLP models that combine words to produce sentence representations as "composition" models, we will also operate with the understanding that there is a goal of achieving *correct* composition, which we define roughly as composition that accurately derives a representation of the meaning of the composed phrase or sentence. What we will need to address in the upcoming sections is what it means to do composition "correctly", how to evaluate how well composition is being done—and ultimately, how to approach doing composition optimally.

We will discuss first the nature of existing approaches to composition in NLP, reviewing existing composition models and digging into a key prerequisite of improving these models: determining how to evaluate them. This is a particularly difficult problem for the types of models that we will be discussing, which produce composed representations in the form of opaque vectors. We will then introduce a new method, inspired by analysis methods in neuroscience, for analyzing and assessing such representations. We will describe the details of this method and demonstrate the use of it on several influential composition models in NLP. Finally, we will discuss implications and directions for using this method to improve the capacity of these models to do composition.

6.2 Composition of complex lexical representations

The problem of composing sentence meaning has of course been addressed in substantial detail within the domain of compositional semantics. However, as we discussed in

Chapter 2, standard approaches in compositional semantics tend to set aside the specifics of lexical content, such that a sentence such as "Mary likes John" might be represented with a placeholder for the verb meaning, resulting in a composed representation such as LIKE(Mary, John).

The importance of specifying lexical content within the compositional system is thrown into stark relief when we try to implement sentence composition in NLP systems. If we give an NLP system a representation such as LIKE(Mary, John), the system has no way of discerning, for instance, that this meaning is comparable to LOVE(Mary, John), while the meaning represented by DISOWN(Mary, John) is quite a different thing. Without a complex representation of the relevant lexical content, the system has little information about meaning at all.[1]

Lexical content is not only critical for the system overall—there is a strong argument to be made for lexical content interacting to a substantial extent with the composition process itself. This is argued for by Pustejovsky (1991), and is perhaps most clearly illustrated by considering the problem of sense selection. Consider the verb *bake* in the sentence "John baked the potato" versus "John baked the cake". As Pustejovsky points out,[2] the verb has different meanings in these two sentences, describing a change of state in one, and an act of creation in the other. Although one can suppose that these are simply different lexical entries, that solution quickly becomes unwieldy when we consider the very many subtle shades of meaning involved in polysemy—instead, it seems that a much more efficient system would be one in which word meanings and composition functions

[1]This is of course a recognized need within NLP and artificial intelligence more generally. Examples of relevant work include Resnik and Diab (2000) and Brachman and Schmolze (1988).

[2]Referencing Atkins, Kegl, and Levin (1988).

interact systematically to give rise to the proper senses. This is the type of system argued for by Pustejovsky, and as we will see in Section 6.3, proper sense selection is precisely the goal pursued by some of the composition work in NLP.

This need to specify and manipulate lexical content — including in the context of composition — has increasingly been addressed in NLP by the use of vector space models. As we discussed in Chapter 2, vector space models may or may not be an adequate approach to representing word meanings for participation in phrase and sentence meaning composition, but they stand as an automatically-derivable (and therefore scalable) approach to deriving representations of lexical content. If nothing else, representing words by distribution-based vectors allows for computation of graded similarity relations, such that there is at least non-zero information about the different relationship between LIKE(Mary, John) and LOVE(Mary, John) as opposed to that between LIKE(Mary, John) and DISOWN(Mary, John). For this reason, these models have become quite popular for purposes of NLP models.

Are these vectors a reasonable representation of the lexical meaning? Up to this point we have avoided making any commitments with respect to the relationship between distribution-based word vectors and the type of lexical meaning content that would participate in the sentence composition process. In the previous chapters we have taken advantage of these vectors because they capture co-occurrence-based lexical relations in a way that resembles asyntactic lexical effects on the N400—but those relations may or may not have any direct connection to the composition of sentence meaning. The encoding of association information, in particular, seems extraneous to the meaning content that we would expect to interface with composition.

This is not to say, however, that vector space representations could not conceivably capture the lexical meaning that is needed for composition. As we discussed in Chapter 2, many theories of lexical meaning and lexical concepts consist of distributed feature-based structures—a class that we may reasonably consider VSMs to belong to. This suggests that although distribution-based VSMs in their current form may not capture word meaning accurately, the vector space approach may be reasonably suited to capturing meaning information, if we can identify an effective way of doing so.

Of course, it is not only representation of lexical meaning that NLP models need to master—it is also the composition of those meanings to form representations of phrase and sentence meanings. In the next section, we will review some of the approaches that have been used in NLP to implement sentence composition with vector space representations. Then, in the following sections we will address the question of how we can assess whether these models are doing composition effectively or not.

6.2.1 NLP vector composition models

Composition models in NLP take a variety of forms. We will focus on approaches that operate on word vector representations, or that otherwise use the word content of a phrase or sentence to produce a vector representation for the phrase or sentence. Note that we will sometimes use the term "word embedding" or "sentence embedding" to refer to these vector representations.

Fixed composition functions

Some approaches to vector composition simply select fixed mathematical operations for combining vectors.

One example of this type of model is the averaging approach that we have taken in previous chapters: taking pre-trained vectors and simply averaging them together. Despite its simplicity, this approach has seen widespread use in NLP, and has often been found to be surprisingly effective (Adi, Kermany, Belinkov, Lavi, & Goldberg, 2016; Arora, Liang, & Ma, 2016; Wieting, Bansal, Gimpel, & Livescu, 2015).

A variant of this approach is introduced by Kintsch (2001), who rather than simply averaging vectors together, instead computes the sum of a given predicate and argument, along with additional neighbors in the semantic space that are most closely related to both the predicate and the argument. The goal of this approach is to use the semantic neighbors to better select for features of the predicate that relate to the argument.

Mitchell and Lapata (2008) similarly use pre-trained word vectors with fixed composition operations, but they test a range of mathematical operations as potential composition functions: additive operations, multiplicative operations, and combinations thereof. Some of their functions incorporate weighting parameters, such that one element of the composed pair will contribute more substantially to the final representation—this is intended to allow for some syntactic awareness. They find that based on their evaluation metric (which we will discuss below) the multiplicative model and a combined additive/multiplicative model perform best. Mitchell and Lapata (2010) expand the set of tested composition functions to a number of other operations, including a function which

dilates the representation of the phrase head in the direction of the modifier, also allowing for some measure of syntactically-informed behavior.

By contrast to the above approaches, which pre-train word vectors and then seek an appropriate composition operation for those vectors, Fyshe et al. (2015) take the opposite approach: pre-selecting a weighted addition operation as the composition function, and then learning word representations to be compatible with that composition function. To do this, Fyshe et al. (2015) derive distribution-based representations for adjective-noun *phrases*, by computing the distributions of those phrases in a corpus. They then optimize their word vectors such that across phrases p consisting of words i and j, the weighted addition of the vectors for i and j is maximally close to the vector for p. For example, given a distribution-based vector representation for the phrase *military aid*, this method would optimize the individual vectors for *military* and *aid* such that when those vectors are combined by weighted addition, the result is maximally close to the distribution-based vector for the phrase *military aid*.

Note that apart from the generic averaging model, the above models do not attempt composition of full sentences, but instead focus on composition of two-word phrases.

Neural network composition models

Many of the more recent composition models have consisted of neural networks, which learn more complex functions to map from the words of a sentence to a sentence representation.

In models of this kind, a key consideration for encouraging the network to learn how to compose sentences is the nature of the training objective used to guide the model's

learning. The relevant question in choosing a training objective is the following: given that the model is going to make predictions based on sentences, what type of predictions should it be asked to make, so as to force it to learn how to compose the sentences well?

Along these lines, Kiros et al. (2015) develop a model that is trained to predict the preceding and following sentences based on the representation of the current sentence. Hill, Cho, Korhonen, and Bengio (2015) leverage connections between words and dictionary definitions, training a neural network to map between composed dictionary definitions and the words that they are defining. Hill, Cho, and Korhonen (2016) try a similar method, but instead of using dictionary definitions, they use composed caption representations to predict vectors for images. Hill et al. (2016) also introduce a sequential denoising autoencoder model, which is trained to predict sentences based on distorted versions of themselves (with words sometimes dropped or reversed in order). Conneau, Kiela, Schwenk, Barrault, and Bordes (2017) train a neural network model to predict entailment relations between sentences, based on those sentences' representations.

Syntactically-guided neural network models

Although some of the above models attempt to incorporate syntactic sensitivities in indirect ways, none incorporate syntactic structure explicitly. By comparison, some neural network approaches to composition integrate syntactic structure directly into the functioning of the network, such that composition proceeds according to that structure.

An example of such an approach is Socher, Huval, Manning, and Ng (2012) who introduce a neural network method that involves computing both a matrix and a vector for every node in a syntactic tree (starting by assigning a matrix and vector to every word of

the vocabulary). The idea of this approach is that the vector will capture meaning content, while the matrix captures whatever transformation the node performs when it combines with another—such that in every composition of nodes, each node both influences and is influenced by the other. This model is trained so as to optimally predict sentiment labels, at each node of the syntactic tree.

The models of Bowman et al. (2016) and Dyer, Kuncoro, Ballesteros, and Smith (2016) also use syntactically-guided neural network composition methods, which either use or infer the parse of the sentence, and combine constituents based on that structure. For guiding the neural network's learning, Bowman et al. (2016) use an objective based on a) parsing accuracy of the model, and b) entailment relations between sentences. Dyer et al. (2016) train their model specifically as a parser, and optimize based on maximizing the likelihood of parses in a corpus.

6.3 Evaluating composition

Having reviewed a variety of models that attempt to do composition, we arrive naturally at the question: are these models any good?

This brings us to the problem that will drive the work presented in the following sections: that of how to evaluate the representations produced by these NLP models, so as to assess how effectively they have executed composition of the sentence meaning.

Why is this a difficult problem? The most salient difficulty for the models reviewed above is that all of these models operate within a vector space representation framework, which means that the sentence representations that they produce take the form of dense

vectors, which are highly opaque to interpretation. Unlike sentence representations in linguistics, which take human-readable forms, we cannot assess the information contained in these vectors simply by examining them.

6.3.1 Existing evaluation approaches

Existing evaluation approaches can be divided into two classes: methods that evaluate based on performance on a downstream task, and methods that evaluate more directly based on the properties of the composed representations. This corresponds roughly to the distinction between "extrinsic" and "intrinsic" evaluation measures.

Extrinsic (downstream tasks)

Most of the neural network methods described above are evaluated using extrinsic metrics, assessing composition systems based on the performance that their sentence representations produce on downstream tasks. Kiros et al. (2015) evaluate on paraphrase detection. Hill et al. (2016) evaluate on a number of tasks including paraphrase detection, sentiment analysis, and question classification. Conneau et al. (2017) evaluate on a variety of tasks as well, including sentiment analysis, question classification, and entailment. Dyer et al. (2016) evaluate on parsing and language modeling tasks. Socher et al. (2012) evaluate primarily on sentiment analysis, the task on which the model is trained. Similarly, Bowman et al. (2016) evaluate on entailment, the basis of their training objective.

The advantage of evaluating based on downstream tasks is that this allows us to assesses directly how useful a composition system, and its representations, are for the types

of tasks that NLP systems are designed to be able to perform. From a practical standpoint, this is a necessary type of evaluation to have.

However, from the perspective of assessing sentence composition *per se*, evaluating on downstream tasks is prohibitively indirect—for any given task, there are many factors that might contribute to strong or poor performance, with accurate composition of sentence meaning being only one (consider for instance the sentiment analysis task, and the relevance of factors like pragmatic reasoning and world knowledge for performing this task). Even if we were able to confidently attribute the strength of performance to the composition process *per se*, it would be very difficult to zero in on the particular aspects of composition that these models can or cannot do.

Intrinsic (similarity based)

By contrast to the extrinsic metrics, intrinsic metrics focus on properties of the composed representations themselves. Most of these metrics rely on assessing similarity relations between representations produced by the composition models.

The most common intrinsic task for evaluating sentence representations is a sentence-similarity task: pairs of sentences are assigned similarity ratings based on human judgments, and models are assessed based on how well the similarities between their representations of the sentences correlate with the similarity ratings from the humans (e.g., Marelli et al., 2014). Kiros et al. (2015), Hill et al. (2016) and Conneau et al. (2017) evaluate on the sentence similarity task as well as on their extrinsic tasks.

The phrase-composition models above make use of different intrinsic measures, geared toward phrases but still based on similarity. Kintsch (2001) uses a small number of hand-

constructed examples that involve comparing composed phrase vectors to "landmarks" selected to emphasize target senses of the verb. For instance, he composes "horse ran" and "color ran" and computes the similarity of the resulting vectors to "gallop" and "dissolve", which are synonyms of each of the correct verb senses, respectively. Mitchell and Lapata (2008) take the same basic evaluation approach as Kintsch, but they expand from Kintsch's small set of hand-selected examples to a larger test set of 120 items, and they collect similarity ratings to be correlated with, rather than assessing subjectively.

Fyshe et al. (2015) evaluate by comparing their composed representations with the corresponding distribution-based phrase vectors (recall that their approach involves computing phrase vectors based on distributions of phrases in text).

Similarity-based intrinsic measures have the advantage of assessing composed representations more directly, isolating away from other task-related factors that may influence performance on downstream tasks. They also make use of the reasonable and intuitive notion that if composed representations of sentences or phrases capture meanings accurately, then they will be similar to representations of other sentences or phrases with similar meanings.

However, similarity scores (particularly at the level of sentences) are notoriously subjective—and more to the point, they are far too coarse-grained to capture all of the nuances of meaning involved in sentence composition (consider sentences such as "the man and the son slept", "the man or the son slept", "the man with the son slept", all of which would surely receive relatively high similarity scores, but which differ in distinct and important ways). Like the extrinsic metrics, similarity metrics do not allow us to zero in on specific aspects of composition that systems may or may not be capturing.

How our approach will differ

In the following sections, we will introduce an alternative method that aims to probe more directly for the extractability of specific linguistic information from sentence representations produced by these NLP models. By contrast to the extrinsic measures, our method targets properties of the composed representations *per se*—and by contrast to the similarity-based intrinsic approaches, our method is designed to probe for different types of linguistic information at a finer-grained level than is afforded by similarity rating alone.

By contrast to extrinsic measures, our approach operates on the assumption, argued for in Chapter 1, that the goal of composition models in NLP should be to produce task-general representations based on core sentence *meaning*.

We prioritize evaluation as a first step in addressing sentence composition because without clear assessments of how well systems are executing composition, and which aspects of composition they can and cannot do, we have no way to implement guided improvements of the models. This work thus stands as a necessary prerequisite to identifying the key driving components for allowing systems to do composition with complex lexical content.

6.3.2 Related work

The work described in the following sections relates closely to a growing effort to increase interpretability of neural network models in NLP—including use of visualization to analyze what neural networks learn (Kádár, Chrupała, & Alishahi, 2016; Li, Chen, Hovy, & Jurafsky, 2015), efforts to increase interpretability by generating explanations

of model predictions (Lei, Barzilay, & Jaakkola, 2016; Li, Monroe, & Jurafsky, 2016; Ribeiro, Singh, & Guestrin, 2016), and work submitting adversarial examples to systems in order to identify weaknesses (Ettinger et al., 2017; Jia & Liang, 2017; Zhao, Dua, & Singh, 2017).

Methodologically the most closely related work is that of Adi et al. (2016), which like our method uses classification tasks to probe for information in sentence vectors. We will discuss that work and its relationship to ours in greater detail below.

Our focus on assessing linguistically-motivated information relates to work on evaluations that aim for fine-grained analysis of systems' linguistic capacities (Emily M. Bender, Flickinger, Oepen, & Zhang, 2011; Marelli et al., 2014; Rimell, Clark, & Steedman, 2009a). The present work contributes to this effort with new tasks that assess composition *per se*, and that do so in a highly targeted manner via careful controls. Our use of synthetically-generated data to achieve this level of control relates to work like that of Weston et al. (2015), which introduces synthetic question-answering tasks for evaluating the capacity of systems to reason with natural language input.

Our examination of the capacity of neural sequence models to identify abstract relations in sentence representations also relates to work by Linzen et al. (2016), who explore whether LSTMs can learn syntactic dependencies, as well as Williams, Drozdov, and Bowman (2017), who investigate the extent to which parsers that are learned based on a semantic objective produce conventional syntax.

Finally, importantly related work is that concerned specifically with testing systematic composition. Lake and Baroni (2017) investigate the capacity of RNNs to perform zero-shot generalization using composition, and Dasgupta, Guo, Stuhlmüller, Gershman, and

Goodman (2018) construct an entailment dataset with balanced lexical content in order to target composition more effectively. We contribute to this line of inquiry by establishing an analysis method that can take output embeddings from sentence composition models and query them directly for specific types of information to be expected in properly compositional sentence representations.

6.4 Applying neuroscientific analysis to sentence vectors

In this section we introduce a different approach to evaluating composition in sentence vector representations. The approach is based on the multivariate pattern analysis (MVPA) method used for analysis of recorded brain data. Our proposal of this approach builds upon one key observation: MVPA tests for encoding of specific target information types in vectors of brain data—if we can adapt the approach to target specific information of interest for composition, then we can apply it to probe for information in vectors produced by NLP systems for representing sentences.

In this section I will introduce the original proposal and preliminary experiments, as described in Ettinger, Elgohary, and Resnik (2016). In the following section I will move on to describing the subsequent system that developed out of this proposal.

6.4.1 MVPA

Multivariate pattern analysis (Haxby, Connolly, & Guntupalli, 2014) makes use of machine learning to test whether certain types of information can be extracted from vectors

of recorded brain data, in order to draw conclusions about the types of information being encoded in the activity of a given brain region.

Consider the following example. Suppose that we want to test whether brain region X is encoding information about the animacy of a visual object. To use MVPA to test this, we might record the brain activity from many instances of a human viewing both animate and inanimate objects. Having obtained recordings of activity from region X corresponding to processing of animate and inanimate visual objects, MVPA now asks a single critical question: is there a consistent underlying difference between the activations that correspond to "animate" and the activations that correspond to "inanimate", such that we can distinguish the activations based on animacy?

For instance, a very simple way for this criterion to be satisfied would be if a single voxel (corresponding to a single dimension of the recorded vector) or a cluster of voxels is "on" when the image depicts something animate, and is "off" when the image depicts something inanimate. In this case, we could use this voxel or cluster of voxels to distinguish between the animacy categories—and this would be cause to conclude that brain region X (or more specifically, that cluster of voxels) is encoding information about animacy. Figure 6.1 shows an illustration of this scenario.

Of course, the keen observer will point out that it may not be animacy *per se* that the region encodes. What other things might be confounded with animacy? What if the region is encoding something like theory of mind? Capacity for motion? The problem of controlling for confounds will be covered in detail below.

The tool used in MVPA for testing this distinguishability of activation vectors is a classifier. Classifiers are machine learning models that are trained to make predictions

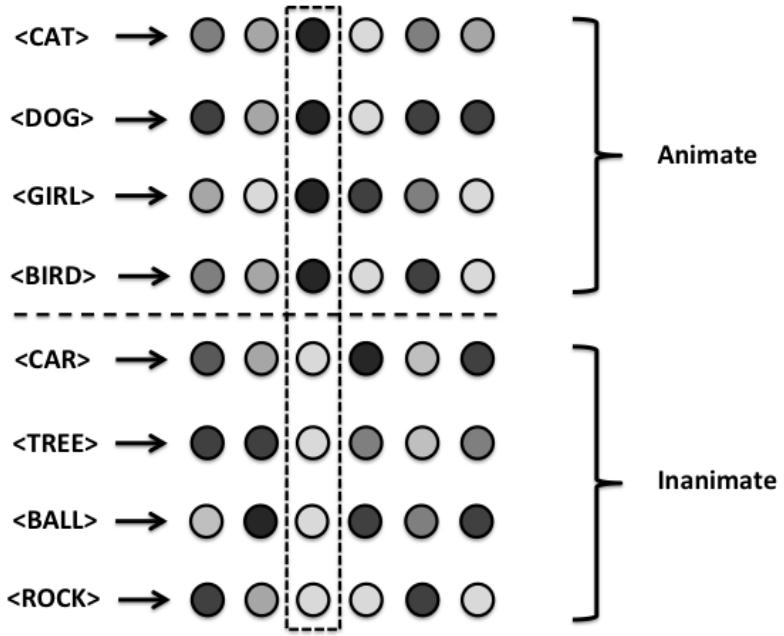

Figure 6.1: Illustration of an idealized scenario of animacy encoding in brain activations. Visual objects (shown here with descriptions such as <DOG>) produce vectors of recorded brain activations, and in this scenario we see that there is one dimension of these vectors that has high values (dark color) in the case of animate objects, and low values (light color) in the case of inanimate objects, which would enable classification of these vectors based on animacy.

based on input features. In the case of our MVPA example, the input features given to the classifier are the dimensions of the vector of activations from region X, and the prediction that the classifier is trained to make is whether the activation vector corresponds to an animate object or an inanimate one. If the classifier is able to classify the animacy of the vectors, we conclude that the vectors contain information about animacy.

An important concept that needs to be employed for this method to work is the standard machine learning notion of a train/test split. This is necessary to ensure that the classifier's correct predictions are generalizable to new data. Consider the following possibility: vectors a, b, c, d, e, f contain no information about animacy, but we train the classifier to recognize that vectors a, c and e correspond to "animate", and vectors b, d and f correspond to "inanimate". Now we present vector b to the classifier, and it correctly produces a prediction of "inanimate", despite there being no animacy information in the vectors.

In order to ensure that the classifier is picking up on generalizable information rather than simply memorizing the trained vectors, we split the vectors into separate training and test items. So if, for instance, we train the classifier on vectors a, b, c, d, and then test it on vectors e and f, it should only be able to classify these correctly if it has picked up on generalizable characteristics of the training items, rather than simply memorizing them individually. In MVPA this typically corresponds to training on some recorded trials and testing on a held-out set of other trials.

Frankland and Greene (2015)

Serving as particular inspiration for our method here is a study by Frankland and Greene (2015), who use MVPA to investigate the encoding of semantic role information. In order to do this, they record the brain activity of subjects reading simple sentences consisting of events described by transitive verbs in active or passive form, as in Table 6.1.

the dog chased the cat
the man was bitten by the dog
the cat saw the girl
the girl was bumped by the car

Table 6.1: Frankland and Greene (2015) example sentences.

The authors then test whether classifiers can be trained to identify the agents and patients of the described events, based on the recorded brain activations corresponding to these sentences. For instance, a classifier trained to identify agents would be tested on whether it could classify the sentence "the dog chased the cat" with a label of "dog". The authors are able to use this method to identify brain regions that appear to encode agent and patient information, respectively.

This study serves as an example of using MVPA to probe vectors (of brain activity) for fundamental linguistic information. We use this model as inspiration for probing vector sentence representations for linguistic information relevant to assessing composition.

6.4.2 Probing for semantic information with targeted classification tasks

The reasoning of our method is as follows: if we take a set of sentences—each represented by a composed vector—and introduce a classification scheme requiring identification of a particular type of semantic information, then by testing classification accuracy on this task, we can assess whether the composed vector representations give access to the information in question.

Our approach assumes there to be identifiable components of *meaning* that we can expect in well-formed sentence representations. For instance, the sentence "the dog chased the girl" contains the information that there was a chasing event, and a dog was the chaser (*agent* of chasing) and a girl the chasee (*patient* of chasing). The sentence "the dog did not bark" conveys that a barking event did not happen.

In order to have maximum confidence in our interpretation of performance in these tasks, our sentences must have sufficient diversity to ensure that there are no consistently correlating cues—confounds—that would allow for strong performance without capturing the relevant compositional information. Relatedly, we want to ensure that the classification tasks cannot be solved by memorization (rather than actual composition) of phrases.

Linguistic properties

There are many types of linguistic information that we might probe for with this method. For our purposes here, we are going to start with two basic types, which are understood

in linguistics to be fundamental components of meaning, and which also have clear ties to tasks for NLP systems: semantic role and negation.

The importance of semantic role information is well-recognized both in linguistics and in NLP—obviously it has occupied a substantial amount of our attention in reasoning about human language comprehension in Chapters 4 and 5, and in NLP it figures critically in tasks such as semantic role labeling and textual entailment. Similarly, the significance of negation is clear from linguistics, and is also widely acknowledged in NLP, in particular for applications such as sentiment analysis (Blunsom, Grefenstette, & Hermann, 2013; Iyyer, Manjunatha, Boyd-Graber, & III, 2015).

Example classification tasks

Once we have identified our information of interest, we can design classification tasks to target this information.

Semantic role If a sentence representation has captured semantic roles, a reasonable expectation would be extractability of the entity-event relations contained in the sentence meaning. So, for instance, we might choose *professor* as our entity, *recommend* as our event, and AGENT as our relation—and label sentences as positive if they contain *professor* in the AGENT relation with the verb *recommend*. Negative sentences for this task could in theory be any sentence lacking this relation—but it will be most informative to use negative examples containing the relevant lexical items (*professor*, *recommend*) without the relation of interest, so that purely lexical cues cannot provide an alternative classification heuristic.

Positive label	Negative label
the professor recommended the student	**the student recommended** the professor
the administrator was **recommended** by **the professor**	the professor was **recommended** by **the administrator**
the school hired the researcher that **the professor recommended**	the school hired the professor that **the researcher recommended**
the school hired **the professor** that **recommended** the researcher	the school hired the professor that was **recommended** by **the researcher**
the professor that liked the school **recommended** the researcher	**the school** that hired the professor **recommended** the researcher

Table 6.2: Labeled data for professor-as-agent-of-recommend task (*recommend* verb and its actual agent have been bolded).

Examples illustrating such a setup can be seen in Table 6.2. In this table we have included a sample of possible sentences, varying only by active/passive alternation and placement of relative clauses, and holding lexical content fairly constant. The verb *recommend* and its agent have been bolded for the sake of clarity.

An important characteristic of the sentences in Table 6.2 is their use of long-distance dependencies, which cause cues based on linear order and word adjacency to be potentially misleading. Notice, for instance, that sentence 5 of the positive label column contains the string *the school recommended*, though *school* is not the agent of *recommended*—rather, the agent of *recommended* is located at the beginning of the sentence. We believe that incorporation of such long-distance dependencies is critical for assessing whether systems are accurately capturing semantic roles across a range of naturally-occurring sentence structures (Emily M Bender, Flickinger, Oepen, & Zhang, 2011; Rimell, Clark, & Steedman, 2009b).

Negation Negation presents somewhat of a challenge for evaluation. How can we assess whether a representation captures negation properly, without making the task as sim-

ple as detecting that negation is present in the sentence (a substantial confound observed by Bentivogli et al. (2016) and Lai and Hockenmaier (2014) for the SICK entailment dataset (Marelli et al., 2014))?

One solution that we propose is to incorporate negation at various levels of syntactic structure (corresponding to different negation scopes), which allows us to change sentence meaning while holding lexical content relatively constant. One way that we might then assess the negation information accessible from the representation would be to adapt our above classification task to focus not on a semantic role relation *per se*, but rather on the polarity of an event described by the sentence meaning. For instance, we might design a task in which sentences are labeled as positive if they describe an event in which a professor actually performs an act of recommending, and negative otherwise.

The labeling for several sentences under this as well as the previous classification scheme are given in Table 6.3. In the first sentence, when negation falls in the relative clause—and therefore has scope only over *like the school*—the professor entity does perform an act of recommending. In the second sentence, however, negation has scope over *recommend*, resulting in a meaning in which the professor, despite being agent of *recommend*, is not involved in performing a recommendation. By incorporating negation in this way, we allow for a task that assesses whether the effect of negation is being applied to the correct component of the sentence meaning.

sentence	prof-ag-of-rec	prof-recommends
the professor that *did not* like the school **recommended** the researcher	TRUE	TRUE
the professor that liked the school *did not* **recommend** the researcher	TRUE	FALSE
the school that liked the professor **recommended** the researcher	FALSE	FALSE

Table 6.3: Sentence labeling for two classification tasks: "contains *professor* as AGENT of *recommend*" (column 2), and "sentence meaning involves professor performing act of recommending" (column 3).

6.4.3 Preliminary experiments

As part of this original proposal, we conducted preliminary experiments to test that this method could yield results patterning in the expected direction on tasks for which we have clear predictions about whether a type of information could be captured. Here we report the results of those preliminary experiments, which will set us up to describe the full method and subsequent experiments.

Preliminary experiment settings

For these preliminary experiments we compared three methods for producing sentence vectors: 1) Simple averaging of GloVe vectors, 2) Paraphrastic word averaging embeddings (Paragram) trained with a compositional objective (Wieting et al., 2015), and 3) Skip-Thought (ST), the neural network model from Kiros et al. (2015) described in Section 6.2.1.[3] For each task, we used a logistic regression classifier with train/test sizes of 1000/500.[4] The classification accuracies are summarized in Table 6.4.

[3]We used the pre-trained models provided by the authors. GloVe and Paragram embeddings are of size 300 while Skip-Thought embeddings are of size 2400.

[4]We tuned each classifier with 5-fold cross validation.

We used three classification tasks for the preliminary testing. First, before testing any actual indicator of composition, as a sanity check we tested whether classifiers could be trained to recognize the simple presence of a given lexical item, specifically *school*. As expected, we see that all three models are able to perform this task with 100% accuracy, suggesting that this information is well-captured and easily accessible. As an extension of this sanity check, we also trained classifiers to recognize sentences containing a token in the category of "human". To test for generalization across the category, we ensured that no human nouns appearing in the test set were included in training sentences. All three models reach a high classification performance on this task, though Paragram lags behind slightly.

Finally, we did a preliminary experiment pertaining to one of our information types of interest: semantic role. We constructed a simple dataset with structural variation stemming only from active/passive alternation, and tested whether models could differentiate sentences with *school* appearing in an agent role from sentences with *school* appearing as a patient. All training and test sentences contained the lexical item "school", with both active and passive sentences selected randomly from the full dataset for inclusion in training or test sets. Note that with sentences of this level of simplicity, models can plausibly use fairly simple order heuristics to solve the classification task, so a model that retains order information (in this case, only ST) should have a good chance of performing well. Indeed, we see that ST reaches a high level of performance, while the two averaging-based models never exceed chance-level performance.

Task	GloVe	Paragram	ST
Has-school	100.0	100.0	100.0
Has-human	99.9	90.5	99.0
School-as-agent	47.98	48.57	91.15

Table 6.4: Percentage correct on has-school, has-human, and has-school-as-agent tasks.

6.4.4 Interim discussion

This section has presented the original proposed form of our method for analyzing compositional information in sentence vectors. This proposal draws inspiration from MVPA to allow for probing of opaque vectors, and makes use of classification tasks designed in a targeted manner in order to assess the extractability of certain types of linguistic information of interest. Preliminary experiments demonstrate that the results of this method are consistent with our expectations on tasks for which we have clear predictions.

It should be noted that although we focus on neural composition models and sentence embeddings in the discussions and experiments here, this analysis method can also be applied more broadly. Since the method simply operates by classification of sentence representations, it can be applied to any format of sentence representation that can be input as features to a classifier.

The next section will describe the more recently developed version of the method, in which we flesh out and strengthen the original proposal with a number of more rigorous controls aimed at better isolating the information of interest, and we substantially expand the scope of the tests through the use of a more sophisticated sentence generation system.

6.5 An analysis system for assessing sentence composition

In the previous section we described the original foundation of our method for assessing compositional information in sentence vector representations. In the present section we will describe the fully formed version of the method that has subsequently developed based on that proposal.

To ensure validity of our tests we introduce three mechanisms of control. First, to create controlled datasets at the necessary scale, we develop a generation system that allows us to produce large sentence sets meeting specified semantic, syntactic and lexical constraints, with gold-standard meaning annotation for each sentence. Second, we control the train-test split so as to require more robust generalization in order to perform the tasks successfully. Third, we employ a sanity check leveraging known limitations of bag-of-words (BOW) averaging models: for any tasks requiring order information from the source sentence (which BOW models cannot logically retain), we check to ensure that BOW models are at chance performance.

Our use of these controls stands in contrast to the prevailing approach in NLP, which typically does not employ careful control over evaluation datasets. This issue has gained increasing attention in recent work: many existing evaluation datasets have been shown to contain biases that allow for high performance based on superficial cues, thus inflating the perceived success of systems on a broad range of downstream tasks (Bentivogli et al., 2016; Gururangan et al., 2018). In the present work, our first priority is careful control of our tasks such that biases are eliminated to the greatest extent possible, allow-

ing more confident conclusions about systems' compositional capacities than are possible with existing metrics.

A highly relevant illustration of the effects of lack of control is the result reported in Adi et al. (2016). In a study released at roughly the same time as our original proposal described in the previous section, Adi et al. (2016) use classification tasks to test sentence vector representations for surface variables of sentence length, word content, and word order. Rather than developing controlled datasets, the authors draw their data automatically from naturally-occurring corpora.

One somewhat striking outcome of the Adi et al. (2016) study is the finding that their BOW composition model, which is created by simply averaging word vectors together, attains 70% accuracy on a binary word order classification task (well above chance level of 50%). This result is surprising, given that BOW averaging models necessarily sacrifice any order information from the source sentence. This suggests that the above-chance performance relies on statistical regularities of word ordering in general, independent of the source sentence.

Although there is certainly much use for the kinds of statistical regularities that likely contribute to this result, it is critical that we recognize that *this result cannot possibly reflect a capacity of the vectors to provide information about the order of words in the* ***source*** *sentence*. A correct classification can *only* reflect the extent to which the source sentence accords with order regularities in general.

For our purposes, we are critically concerned with *systematic composition of the source sentence itself*, abstracting away from general statistical regularities. For this reason, our use of controls is essential to the validity of our conclusions. The necessity of

distinguishing source sentence encoding from general regularities is the reasoning behind our BOW sanity check, discussed in Section 6.5.1. The inherent biases in naturally-occurring data, further highlighted by this Adi et al. (2016) result, also motivate our use of generated data, for the sake of maintaining the necessary level of control.

It is worth noting that we also differ from Adi et al. (2016) in specifically targeting more abstract, dependency-based sentence characteristics relevant to composition, by contrast to surface variables of word content, word order, and sentence length that Adi et al. test for. Our targeting of this more abstract information is made possible by our use of generated sentences for which we have detailed (automatically generated) syntactic and semantic annotations.

6.5.1 The analysis method

Classification tasks

The formulation of our classification tasks for the fully-developed method differs somewhat from the formulation in the original proposal. The original proposal attempted to define fixed lexical items for the classification task: for instance, fixing the target relation as <*professor* as AGENT of *recommend*>, and testing whether vectors can be distinguished based on the presence or absence of this lexically-specific relation. This formulation is ideal in that it lends itself well to use of a linear classifier, which is the standard approach in MVPA.

Unfortunately, after extensive testing with this formulation of the classification tasks, we found it prohibitively difficult to control the sentences adequately so as to isolate the

semantic information of interest. An example of a problem that arises with this method is the existence of sequential strings that can be used as relatively superficial cues—e.g., "recommended by the professor".

To circumvent this issue, we transitioned to the variable-target framework used by Adi et al. (2016). This allows us to define the identities of the target components in a variable fashion, and gives us increased flexibility to control the task and filter out superficial cues.

As in the original proposal, we target two meaning components as our starting point: semantic role and negation. We stick to these components because they are fundamental to the meaning of a sentence, and because they represent information types that can be heavily distorted with respect to surface variables like word content and order: to know semantic role and negation information, it is not enough to know which words are in the sentence or which words come earlier in the sentence.

We formulate the semantic role classification task ("**SemRole**") as follows: "Given representation $\mathbf{n}$ of probe noun n, representation $\mathbf{v}$ of probe verb v, and embedding $\mathbf{s}$ of sentence s (with s containing both n and v), does n stand in the AGENT relation to v in s?" For example, an input of $\{n$: "professor", v: "help", s: "the professor helped the student"$\}$ would receive a positive label because *professor* is AGENT of *help* in the given sentence.

We formulate the negation classification task ("**Negation**") as follows: "Given a representation $\mathbf{v}$ of a probe verb v, and an embedding $\mathbf{s}$ of sentence s (with s containing v, one negation, and one other verb), is v positive or negated in s?" For example, an input of $\{v$: "sleep", s: "the professor is not actually helping the student who is totally sleeping"$\}$ receives a positive label because *sleep* is not negated in that sentence. A concern that

arises with these sentences is the confound of adjacency between negation and a verb, so to remove this confound we insert variable-length adverb sequences (e.g., "not *actually, totally* helping") before the verbs in the dataset (both negated and non-negated), to ensure that the negation is not always adjacent to the verb that it affects.

Means of control

The most critical consideration in this work is ensuring that we can draw valid conclusions about composition from performance on our classification tasks. To this end, we take a number of measures to control our data, to avoid biasing cues that would make the tasks solvable independent of the information of interest—a problem observed in many existing datasets, as mentioned above (Gururangan et al., 2018).

Generation system A critical component of isolating abstract meaning information is employing syntactic variation, such that the meaning information of interest is the single underlying variable distinguishing label categories. For instance, we might use sentences like "the professor helped the student", "the student was helped by the professor", and "the student that the professor helped was sleeping"—which vary in structure, but which share an underlying event of a professor helping a student.

In order to produce sentence sets that exhibit this level of variation—and that reach the necessary scale for training and testing classifiers—without allowing the biases inherent in naturally-occurring data, we developed a generation system that takes as input lexical, semantic and syntactic constraints, and that produces large sentence sets meeting those constraints. In addition to allowing us to produce controlled datasets, this system also

ensures that the generated datasets are annotated with detailed semantic and syntactic information. This generation system is described in greater detail in Section 6.5.2.

Train/test splits To be confident that the classifier is picking up on underlying meaning information and not simply a union of different superficial cues across syntactic structures, we make careful provisions in our train/test split to ensure generalization (beyond the obvious split such that sentences in test do not appear in training). For our semantic role task, certain (n,v) probe combinations are held out for test, such that no combinations seen at test time have been seen during training. This is done to ensure that the classifier cannot rely on memorized sequences of words. For our negation task, which uses only one probe, we hold out certain adverbs from training (as described above, adverbs are used as material to separate the negation and the verb), such that at test time, the material separating the negation and the verb (or preceding the non-negated verb) has never been seen in training.

BOW as control As described above, it is logically impossible for BOW models to encode information that requires access to word order from the source sentence itself. We leverage this knowledge to create a sanity check baseline for use in monitoring for lexical biases: if, for any task requiring access to word order information, the BOW baseline performs above chance, we know that the datasets contain lexical biases affecting the classification results, and we can modify them accordingly.

6.5.2 Generation system

In this section we describe the generation system that we use to create large, controlled datasets for our classification tasks. As described above, this system takes input constraints targeting semantic, syntactic, and lexical components, and produces diverse, meaning-annotated sentences meeting those constraints.

Event/sentence representations

As a framework for specifying semantic and syntactic constraints, we use a class of event representations that contain both lexicalized semantic information and necessary syntactic information, such that there is a deterministic mapping from a fully-populated event representation to a corresponding surface sentence form. These representations fall roughly within the category of "lexicalized case frame" outlined by Reiter, Dale, and Feng (2000) for natural language generation. Figure 6.2 shows an example representation, in fully-specified textual form, and in simplified graphical form.

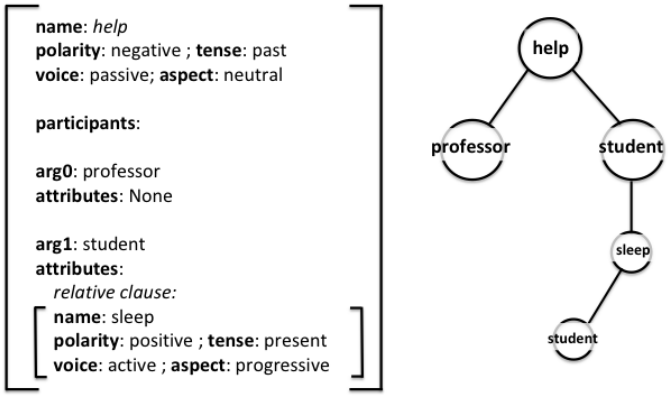

Figure 6.2: Event representation for *"The student who is sleeping was not helped by the professor"*

Our representations are currently restricted to events denoted by transitive and intransitive verbs, with the arguments of those verbs and optional transitive or intransitive relative clauses on those arguments.

These representations are comparable in many ways to abstract meaning representation (AMR) (Banarescu et al., 2013), but rather than abstracting entirely away from syntactic structure as in AMR, our event representations encode syntactic information directly, along with the more abstract meaning information, in order to maintain a deterministic mapping to surface forms. Relatedly, while AMR uses PropBank frames (Palmer, Gildea, & Kingsbury, 2005) to encode meaning information, we encode information via English lemmas, to maintain control over lexical selection during generation.

These representations can be partially specified to reflect a desired constraint, and can then be passed in this partial form as input to the generation system—either as a required component, or as a prohibited component. This allows us to constrain the semantic and syntactic characteristics of the output sentences. In addition to partial events, the system can also take lists of required or prohibited lexical items.

Event population

The system uses a number of structural templates into which partial events can be inserted. Structural templates vary based on the transitivity of verbs and the presence or absence of relative clauses on arguments—for instance, if the nodes in the right side of Figure 6.2 were unpopulated, it would depict an empty structural template consisting of a transitive main verb with an intransitive relative clause on arg1. Once we have inserted a partial event into a subsection of an empty structural template (events can be inserted into either

the main clause or a relative clause), the system populates the remainder of the event components by iterating through available verbs and nouns of the vocabulary, and through available values for unfilled syntactic characteristics (such as polarity, tense, voice, etc.).

For simplicity, we control plausibility of argument/predicate combinations by setting the system vocabulary such that it contains only animate human nouns, and only verbs that can take any of those nouns in the relevant argument slots. This is a reasonable task due to the capacity of the system to generate thousands of sentences from only a handful of nouns and verbs. We leave incorporation of more sophisticated selectional preference methods (Resnik, 1996; Van de Cruys, 2014) for future work.

Our goal is to find the optimal balance between the critical need of this method for structurally variable, carefully controlled sentences, and the practical need to avoid substantial deviation from sentence types to which systems will have been exposed during training. To this end, we draw our vocabulary from comparatively frequent words, and we impose structural constraints to limit the complexity of sentences—specifically, in the current experiments we restrict to sentences with no more than one relative clause, by omitting templates that include relative clauses on both arguments of a main verb.

Syntactic realization

Once an event representation is fully populated, it is submitted to a surface realization module that maps from the event to a surface sentence via a simple rule-based mapping. Since the representations specify syntactic information and use lexicalized meaning information, there is no significant process of lexical selection required during surface realization—only morphological inflection derivable from syntactic characteristics. As

a result, the event representations map deterministically to their corresponding surface forms. We use a grammar specified using the NLTK feature grammar framework (Bird, Klein, & Loper, 2009). Morphological inflections are drawn from the XTAG morphological database (Doran, Egedi, Hockey, Srinivas, & Zaidel, 1994)

Sentence quality

To ensure the quality of the generation system output, we manually inspected large samples of the generated sentences throughout development and after generation of the final sets, to confirm that sentences were grammatical and of the expected form. Table 6.5 shows a sample of the generated sentences.

the men were sleeping
the woman followed the lawyer that the student is meeting
the women were being helped by the lawyers
the student called the man
the scientist that the professors met is dancing
the doctors that helped the lawyers are being recommended by the student

Table 6.5: Example generated sentences

6.5.3 Implementation of lexical variability

As discussed above, we adopt the variable probe formulation used by Adi et al. (2016). This adds a dimension to the learning task that is not present in the original task formulation of Ettinger et al. (2016) described above: the classifier needs not only to identify meaning information in the input sentence—it needs to identify meaning information contingent on the identities of the particular probe words.

To identify the probe word(s) in the input features, Adi et al. (2016) use the source word embeddings, but this is problematic for our purposes, given that we want to test a wide variety of models, which use word embeddings of different types and sizes. To avoid this variability, it would be preferable to use one-hot vectors to identify word probes. To this end, we performed a series of experiments testing whether classification accuracy was affected by use of one-hot probe representations by comparison to embedding probes, in a replication of the word content task of Adi et al. (2016). Finding almost equivalent accuracies between the two input types, we use one-hot probe representations in all subsequent experiments.

Note that as a result, by contrast to Adi et al. (2016) we are not assuming the classifier to identify words in the sentence representation based on resemblance to their original word embeddings—this may not in fact be a safe assumption, given that the word's representation may distort during composition of the sentence. Instead, the classifier must learn a mapping from each one-hot representation to its manifestation in sentences. This means that all words must appear as probes in training. To facilitate the learning of this mapping, we restrict to a small (14-word) vocabulary of probes in these experiments. Because the generation system is able to produce thousands of sentences from even such a restricted vocabulary as this, this limitation does not prevent generation of adequately large datasets.

A note about the size and selection of the vocabulary: some composition tests will surely be sensitive to specific idiosyncrasies of individual words, in which case the choice of vocabulary will be of great importance. In the case of the semantic role and negation tasks described here, however, the focus is on identification of structural dependencies

between words, which are not in this case sensitive to the specific nouns/verbs used. Consequently, for these tasks—as long as vocabulary words are not out-of-vocabulary for the models (which we confirm below)—the important thing should be not what the words themselves are, but whether dependencies between them have been captured in the sentence embeddings.

6.5.4 Surface tasks: word content and order

Though our ultimate interest is in abstract meaning information, part of the goal of these experiments is to get a clear picture of the information currently captured by existing systems. For this reason, we include the content and order experiments as performed by Adi et al. (2016), to see how encoding of these surface variables compares to encoding of meaning information—and to compare with the results of Adi et al. (2016) after the more rigorous controls used in our datasets.

We structure these tasks to be maximally parallel with our meaning tasks. To this end, we have two content tasks: one-probe ("**Content1Probe**") and two-probe ("**Content2Probe**"), with the one-probe task using verb probes as in the negation task, and two-probe using noun-verb probe pairs, as in the semantic role task. Similarly, for the order task ("**Order**") we use only noun-verb pairs. The order task is thus formulated as "Given representation $\mathbf{n}$ of probe noun n, representation $\mathbf{v}$ of probe verb v, and embedding $\mathbf{s}$ of sentence s (with s containing both n and v), does n occur before v in s?". The two-word content task is formulated as "Given representation $\mathbf{n}$ of probe noun n, representation $\mathbf{v}$ of probe verb v, and embedding $\mathbf{s}$ of sentence s, do both n and v occur in s?", and the one-word con-

tent task is formulated as "Given representation $\mathbf{v}$ of probe verb v, and embedding $\mathbf{s}$ of sentence s, does v occur in s?"

6.5.5 Classification experiments

To demonstrate the utility of our analysis, we use it to test a number of existing sentence composition models. Following Adi et al. (2016), for our classifier we use a multi-layer perceptron with a single hidden layer of the same size as the input. For each of the above tasks we construct train/test sets consisting of 4000 training items and 1000 test items. For each embedding method, we run the corresponding model to produce embeddings for the train and test sentences, and we use the resulting embeddings as input features to the classifier. No tuning is necessary, as the hyperparameters of hidden layer number and size are fixed in accordance with the architecture used by Adi et al.

It is important to note that the training of the classifier, which uses the 4000 items mentioned above, is to be distinguished from the training of the sentence embedding methods. The sentence embedding models are pre-trained on separate corpora, as described below, such that they map sentence inputs to embeddings. Once these models are trained, they are used to produce the 4000 sentence embeddings that will serve as training input to the multi-layer perceptron classifier (and the 1000 sentence embeddings used for testing).

Our use of a relatively simple classifier with a single hidden layer builds on the precedent not only of Adi et al., but also of MVPA, which in fact typically uses linear classifiers (an option that we could not employ due to our use of the variable probes). An important reason for use of simpler classifiers is to test for *straightforward* extractability of

information from embeddings—if a complex classifier is necessary in order to extract the information of interest, then this calls into question the extent to which we might consider this information to be "captured" in the embeddings, as opposed to the information being somehow reconstructable from the embeddings' encoding of other information. That said, the question of how the complexity of the classifier relates to the encoding of the target information in these sentence embeddings is an interesting issue, which we will explore further below.

For each experiment, we also run two corresponding experiments, in which random vectors are used in place of the sentence vectors and the probes, respectively. This serves as an additional check for biases in the datasets, to ensure that neither the sentence vectors nor the probe vectors alone are sufficient to perform above chance on the tasks. For all tasks, these random vectors produce chance performance.

Sentence encoding models

We test a number of composition models on these classification tasks (all of which are included in our review of models in Section 6.2.1). These models represent a range of influential current models designed to produce task-general sentence embeddings. They employ a number of different architectures and objectives, and have shown reasonable success on existing metrics (Conneau et al., 2017; Hill et al., 2016).

All sentence embeddings used are of 2400 dimensions, in accordance with the dimensionality of the pre-trained models that we use. Because our pre-trained models (SDAE, Skip-Thought) are trained on the Toronto Books Corpus (Y. Zhu et al., 2015), we use this

as our default training corpus, except when other supervised training data is required (as in the case of InferSent).[5]

BOW averaging Our first sentence embedding model (**"BOW"**) is a simple BOW averaging model, for which we use the skip-gram architecture of the word2vec model (Mikolov et al., 2013) to learn word embeddings.[6] As discussed above, the BOW model serves primarily as a sanity check for our purposes, but it is important to note that this model has had competitive results on various tasks, and is taken seriously as a sentence representation method for many purposes (Adi et al., 2016; Arora et al., 2016; Wieting et al., 2015).

Sequential Denoising Autoencoder Our second model (**"SDAE"**) is an autoencoder variant from Hill et al. (2016) for unsupervised learning of sentence embeddings. The model uses an LSTM-based encoder-decoder framework, and is trained to reconstruct input sentences from their vector representations (last hidden state of the encoding LSTM) despite noise applied to the input sentence. We use a pre-trained model provided by the authors. This model has the advantage of an unsupervised objective and no need for sequential sentence data, and it shows competitive performance on a number of evaluations.

Skip-Thought Embeddings Our next two models are variants of the Skip-Thought model by Kiros et al. (2015), in which sentences are encoded with gated recurrent units (GRUs), with an objective of using the current sentence representation to predict the im-

[5]Before sentence generation, the chosen vocabulary was checked against training corpora to ensure that no words were out-of-vocabulary (or below a count of 50).

[6]We switch from GloVe to word2vec because the GloVe model proved problematic for training word embeddings of this size.

mediately preceding and following sentences. Following the model's authors, we use both the uni-skip (**"ST-UNI"**) and bi-skip (**"ST-BI"**) variants: uni-skip consists of an encoding based on a forward pass of the sentence, while bi-skip consists of a concatenation of encodings of the forward and backward passes of the sentence (each of 1200 dimensions, for 2400 total). We use the publicly available pre-trained Skip-Thought model for both of these variants.

Skip-Thought sentence embeddings have been used as pre-trained embeddings for a variety of tasks. They have proven to be generally effective for supervised tasks and passable for unsupervised tasks (Hill et al., 2016; Triantafillou, Kiros, Urtasun, & Zemel, 2016; Wieting et al., 2015). Like the SDAE model, the Skip-Thought model is able to use unsupervised learning, though it requires sequential sentence data. However, more so than the SDAE model, the Skip-Thought model uses an objective that is intended to capture semantic and syntactic properties, under the authors' assumption that prediction of adjacent sentences will encourage more syntactically and semantically similar sentences to map to similar embeddings.

InferSent Our final model is the InferSent model (Conneau et al., 2017), which uses multi-layer BiLSTM encoders with max pooling on the hidden states of the last layer to produce vector representations of the sentences. This model is trained with an entailment (natural language inference) objective, and for this reason we train it on the Stanford Natural Language Inference dataset (Bowman, Angeli, Potts, & Manning, 2015).

The InferSent model is intended to produce "universal" sentence representations, and has been shown to outperform unsupervised methods like Skip-Thought on a number of

	Accuracy				
	Content1Probe	Content2Probe	Order	SemRole	Negation
BOW	100.0	97.1	55.0	51.3	50.9
SDAE	100.0	79.8	92.9	63.7	99.0
ST-UNI	100.0	88.1	93.2	62.3	96.6
ST-BI	96.6	79.4	88.7	63.2	74.7
InferSent	100.0	70.1	86.4	50.1	97.2

Table 6.6: Classification results

tasks as reported by Conneau et al. (2017). More generally, the NLI objective is believed to encourage learning of compositional meaning information, given that inference of entailment relations should require access to meaning information.

Results and Analysis

Table 6.6 shows the accuracy of the different models' sentence embeddings on our classification tasks. Figures 6.3-6.7 show the results with bootstrapped 95% confidence intervals.

The first thing to note is that our BOW control allows us to confirm nearly complete lexical balance in the sentence sets: the averaged word embeddings perform roughly at chance on all but the content tasks. By contrast, BOW performs with near-perfect accuracy on the content tasks, lending support to the intuitive conclusion: the one thing that BOW *does* encode is word content. The quality of performance of the BOW model on this task exceeds that reported by Adi et al. (2016)—we speculate that this may be due to our use of a smaller vocabulary to facilitate the learning of the mapping from one-hot probes.

While BOW has very high performance on two-probe word content, SDAE, ST-UNI, ST-BI and InferSent have much lower accuracy (albeit still far above chance), suggest-

ing that some detail with respect to word content is sacrificed from these representations in favor of other information types. This is exemplified by the order task, on which all non-BOW models show significantly higher accuracy than on the word content tasks, supporting the intuitive conclusion that such sequence-based models retain information about relative word position. This result is generally consistent with the Adi et al. (2016) result, but due to the additional control that brings BOW roughly to chance, we can conclude with greater confidence that the performance on this task pertains to order information in the source sentence itself.

Turning to our meaning information tasks, we see that with the exception of ST-BI, the sequence models perform surprisingly well on the negation task, despite the fact that this task cannot be solved simply by detecting adjacency between negation and the verb (due to our insertion of adverbs). Instead, we speculate that these sequence models may be picking up on the utility of establishing a dependency between negation and the *next* verb, even in the face of intervening words. This is not a complete solution to the problem of representing the meaning and dependencies of negation, but it is a useful step in that direction, and suggests that models may be sensitive to some of the behaviors of negation.

Interestingly, ST-BI shows markedly weaker performance on the negation task. We see two potential reasons for this. First, it may be due to the reduced dimensionality of each of the two concatenated encodings (recall that ST-BI involves concatenating 1200 dimensional encodings of the forward and backward passes). Second, the reduced performance could be influenced by the inclusion of the backward pass: while the forward pass can leverage the strategy of linking negation to the next verb, the backward pass

cannot use this strategy because it will encounter the relevant verb before encountering the negation.

Turning to the semantic role task, we see a stark contrast with the high performance for the negation task. InferSent performs squarely at chance, suggesting that it retains as little compositional semantic role information as does BOW. SDAE, ST-UNI and ST-BI perform modestly above chance on the semantic role task at 62-63% accuracy, suggesting that they may provide some amount of abstract role information—but no model shows any substantial ability to capture semantic role systematically.

These results accomplish two things. First, they lend credence to this method as a means of gaining insight into the information captured by current models. Second, they give us a sense of the current capacity of sequence-based models to capture compositional meaning information. The picture that emerges is that sequence models are able to make non-trivial headway in handling negation, presumably based on a sequential strategy of linking negation to the next verb—but that these sequence models fall significantly short when it comes to capturing semantic role compositionally. Another point that emerges from these results is that despite the fairly substantial differences in architecture, objective, and training of these models, capacity to capture the compositional information is quite similar across models, suggesting that these distinct design decisions are not having a significant impact on compositional meaning extraction. We leave the testing of more substantially distinct models, like those with explicit incorporation of syntactic structure (Bowman et al., 2016; Dyer et al., 2016; Socher et al., 2013) to future work.

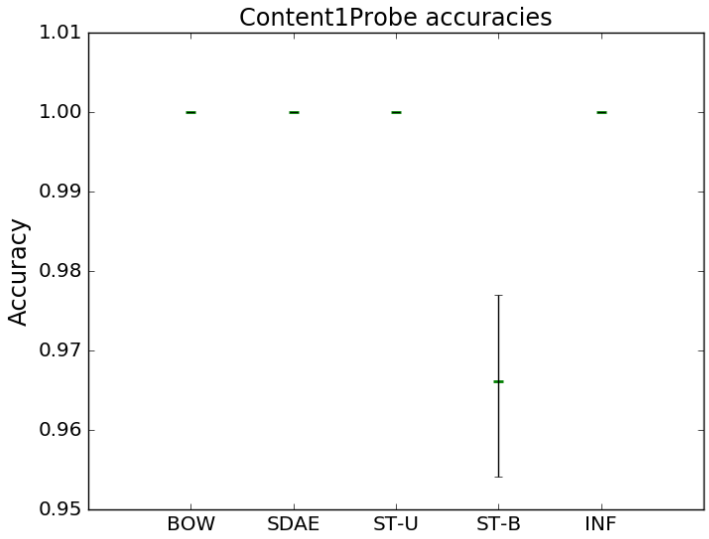

Figure 6.3: Content1Probe accuracies

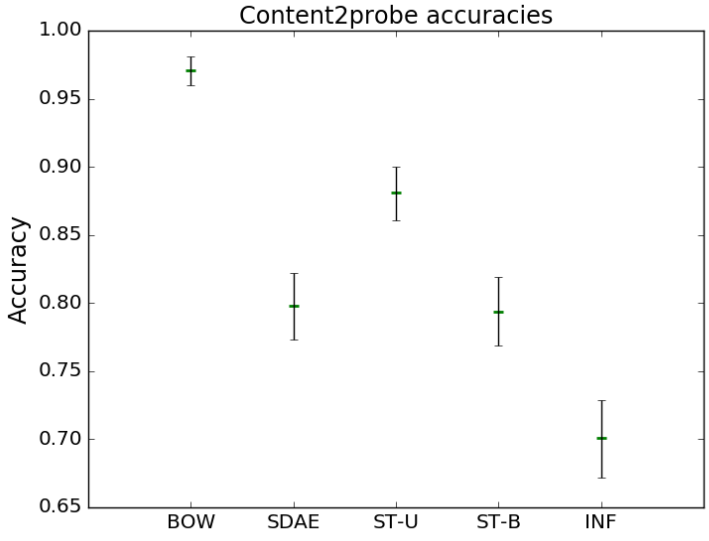

Figure 6.4: Content2Probe accuracies

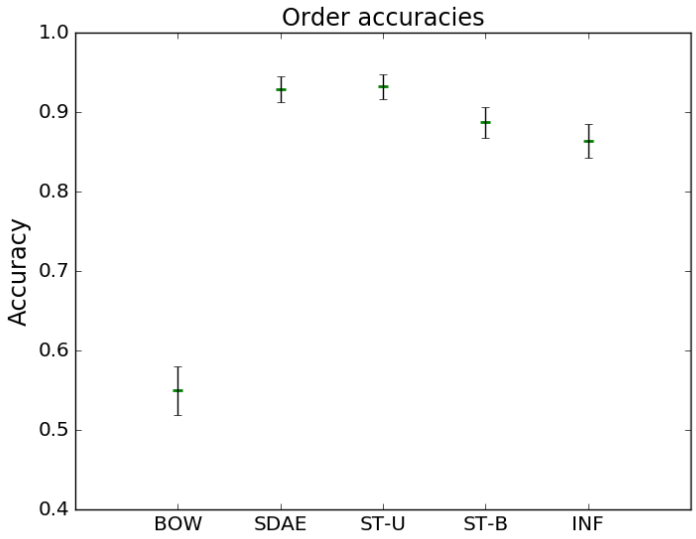

Figure 6.5: Order accuracies

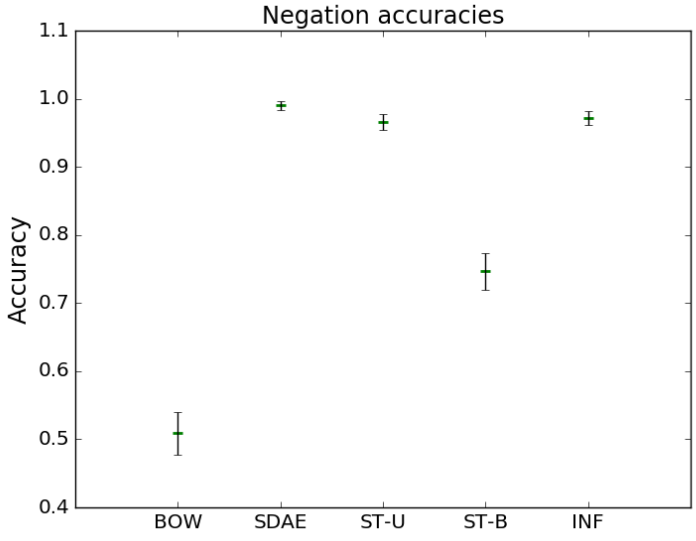

Figure 6.6: Negation accuracies

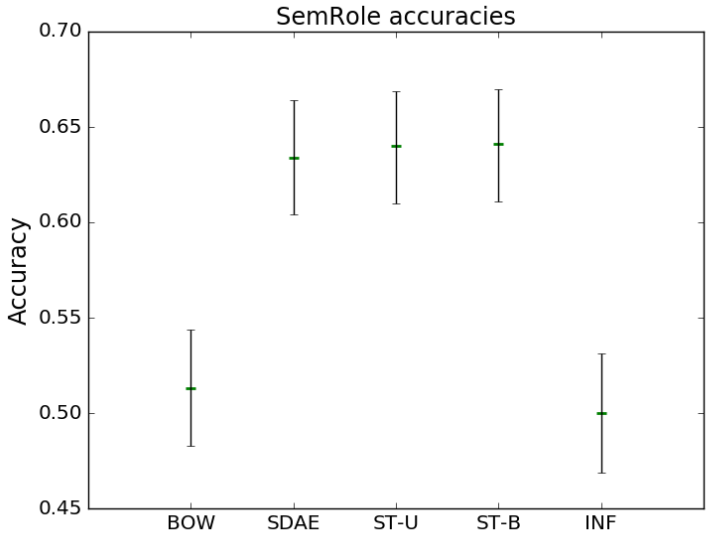

Figure 6.7: SemRole accuracies

Deeper classifiers

As discussed above, for the purposes of these experiments we have strong motivations for restricting to a single hidden layer in our multi-layer perceptron classifier. Ideally, for consistency with MVPA, we would use a linear classifier, but because our use of the variable probe formulation renders a linear classifier not feasible, we satisfy ourselves with a relatively simple non-linear classifier (which is also consistent with the precedent of Adi et al.)

However, this leaves us with the question of what *would* happen if we increased the complexity of our classifiers. Would the accuracies improve if, for instance, we simply add more hidden layers to our classifier? In order to investigate this question, we run the SemRole and Negation experiments with multi-layer perceptron classifiers of two and three hidden layers, to test whether this will allow for extraction of additional information.

Table 6.7 shows the results. For the BOW vectors, as is to be expected, there is no increase from chance-level performance—increasing the complexity of the classifier cannot extract information that simply cannot be there.

For the Negation task, most embedding methods remain at roughly the same accuracy, with slight fluctuations—the exception is InferSent, which drops to chance performance with three hidden layers. Why this sudden drop with three layers occurs only for the InferSent vectors is not entirely clear, though we speculate that it could involve some form of overfitting to spurious properties of the training data.

By contrast, for the SemRole task, we see that increasing the number of layers does in fact non-trivially improve performance, especially in the case of SDAE, which reaches accuracy of 73%. What this suggests is that there is information encoded in these vectors which has a more non-linear relationship to the desired output (that is, to the target classification involving semantic role) such that when a more complex classifier is employed, that relationship is able to be established. Whether this amounts to an encoding of semantic role information *per se* is questionable (as we have argued, in favor of our use of the simpler classifier)—but it does leave open the possibility that with adequately sophisticated classifiers, these vectors may allow for reasonably accurate extraction of semantic role information, while the InferSent vectors, seemingly, will not.

6.6 Discussion and looking ahead

We have presented an analysis method and accompanying generation system designed to address the problem of assessing compositional meaning content in sentence vector

	Accuracy					
	SR	SR2lyr	SR3lyr	Neg	Neg2lyr	Neg3lyr
BOW	51.3	50.1	50.0	50.9	50.3	49.9
SDAE	63.4	69.9	73.7	99.0	97.3	98.5
ST-UNI	64.0	69.9	68.5	96.6	96.4	97.3
ST-BI	64.1	64.8	68.5	74.7	72.0	74.0
InferSent	50.0	51.9	50.0	97.2	94.1	50.0

Table 6.7: Accuracies with more MLP layers

representations. We have also presented the results of applying this method for analysis of a number of current sentence composition models, demonstrating the capacity of the method to derive meaningful information about what is captured in these models' outputs, and showing that even state-of-the-art sequence-based neural network models, when subjected to a controlled test, show little sign of having mastered the dependencies necessary for systematic composition—particularly for semantic role information. The SDAE and Skip-Thought models do show above-chance performance on semantic roles, with the SDAE model allowing for 73% accuracy with a deeper classifier, which suggests that there is information retained in these models that is relevant to semantic role and that is absent from the BOW and InferSent models. However, it is clear that this information is either imperfect or incomplete, given that these accuracies are still only marginally above chance.

Looking forward, we have three things to consider.

6.6.1 Expansion of tested information types

We have developed a method, based in analysis methodology from neuroscience, that allows us to select and target information that we believe should be present in properly

composed sentence vectors. In the experiments reported above, we design tasks to target semantic role dependencies, and dependencies of negation.

It is clear that our methods as implemented here have limitations. In particular, our test of negation is not so much a test of the properties of negation *per se*, but of dependency between negation and a subsequent verb. While we take measures to decouple our task from a task of adjacency, the task remains confounded, as we see in the discussion above, with the task of linking negation to the next verb in the sentence.

Ideally, we will be able to identify variations on this task that will get more directly to the heart of negation and its scoping properties (efforts up to this point have unfortunately been unsuccessful, due to the variety of constraints involved and the persistent problem of new confounds arising as we eliminate others).

However, as we look ahead to expanding our tests of composition, we want to improve our tests not only of syntactic dependencies that underlie composition, but also of more semantically-oriented properties. For this type of direction, we might be able to take advantage of the effect of negation on adjectives, and the synonymy of such phrases with existing adjectives (e.g., "not closed", "open"). As another example, we may be able to take advantage of veridicality effects of embedding verbs, such as factives like "He *knew* that the car was red" as opposed to "He thought that the car was red".

A challenge with implementing more semantic tests of this kind lies in identifying how to generate corresponding tests sets on the scale of those described above. However, if we are to establish a comprehensive and rigorous battery of tests for composition, it will be important to incorporate not just syntactic dependencies, but also semantic properties *per se*.

6.6.2 Testing of additional models

The models tested here represent only a subset of approaches to composition. Most notably from the perspective of our theme in this dissertation, none of the models tested here incorporate explicit syntactic information, as the Socher et al. (2012), Bowman et al. (2016) and Dyer et al. (2016) models do (though some of these models may pick up on and utilize information from syntactic structure implicitly).[7]

Obviously we expect it to be the case that good composition models will require syntactic information—whether explicitly incorporated or implicitly inferred—in order to execute composition successfully. From this perspective, it is certainly critical that we test models that incorporate syntax on top of the vector and neural network frameworks, in order to establish whether, and to what extent, this factor improves upon the performance that we see above.

However, it is important to note that presence or absence of syntax cannot be the entire story—in order to do good composition, we need not only to compose things in the proper order—we need to start with units that capture the proper meaning, and that interface correctly with the compositional operations to produce target phrase meanings. This is a much more difficult problem, and incorporating syntactic information is only the beginning of solving it.

It is also important to note that the models tested above are highly-respected sentence encoders considered to be among the state of the art in NLP. This serves to highlight the

[7]We did in fact run the Bowman et al. model on these tests, but finding chance-level accuracies even on the simpler tasks, we were forced to conclude there to be a bug in the available implementation, which we have been in contact with the authors about, but which has not yet been resolved.

point argued for in Chapter 5: even the most sophisticated NLP models currently fall short of the human capacity for composing and interpreting actual sentence meaning.

6.6.3 How to improve models

This work serves as a necessary precursor to solving the problem of greater interest: helping to determine how to get NLP systems to accomplish representation and composition of meanings. In this chapter we began tackling this problem by designing an analysis system to better assess how well sentence encoding models are doing at composition in the first place. Moving forward, the goal is to apply insights gained from this analysis method in order to identify more precisely which design decisions lead to effective capturing of meaning information, in order to guide system improvement. The tests and results described in this chapter may not yet be sufficiently developed to be able to provide conclusive answers to this question—but in this section we will discuss general considerations for improving composition models.

What are the most promising directions for improving the capacity of models to do composition? One component that seems almost certainly necessary is joint learning of word representations and composition function—or design of lexical representations with an explicit theory of how they interface with composition, as done by Pustejovsky (1991). Given the complexity of the manipulation of lexical content during composition, it seems certain that any lexical representations that do not take into account composition—or compositional operations that do not take into account lexical content—will fall short.

If we are to use a learning-based approach, we also need to identify the most effective assumptions to guide how models will learn and generalize from their environments (inductive biases), and the most effective selection of data to comprise those learning environments. Is it enough to train a recurrent neural network to make sentiment predictions on movie reviews and expect that it will learn how to do composition? Almost certainly not. Is it enough to train a network to predict entailment relations between pairs of sentences and expect it to learn composition? This seems admittedly more plausible by comparison to sentiment, given the fundamental nature of entailment as a measure of meaning—but even so, it seems highly unlikely that the full scope and complexity of the compositional system could be learned based on the signal provided by entailment relations between sentences, without building in additional targeted assumptions to guide learning. Relatedly, the models that we test here are primarily based on neural networks, and all operate within a vector space representation framework—but it is important to consider whether either neural networks or vector-based representations are adequately suited to achieving the nuance of human sentence meaning composition.

By comparison to learning-based approaches, Pustejovsky (1991) takes a rule-based approach to this problem. Pustejovsky argues that to accomplish these goals computationally, we will need structured lexical representations in human-annotated databases—and pre-defined composition operations—by contrast to approaches that attempt to learn word representations and composition functions automatically. Given the level of detail and complexity that Pustejovsky observes to be relevant for capturing the necessary components of meaning, it is easy to see why he takes this stance—the discussion in that paper

is a useful jumping-off point for understanding the extent of what is likely to be needed in lexical representations in order for them to interface with the compositional system.

Does this mean that learning-based models simply cannot accomplish what we need? In considering this, it is of course worth keeping in mind that humans do manage to learn word meanings and how to compose them. What is the signal that these human learners use? What is the framework of innate knowledge that scaffolds this learning process? These are fundamental questions in linguistics and language acquisition—and while NLP need not emulate humans precisely, this particular problem is one that human brains have solved brilliantly, and that remains unsolved both in NLP and in cognitive science. This suggests that in bridging these domains and drawing on the insights to be shared from each, we stand in the end both to gain valuable insights into the human language capacity, and to significantly improve NLP systems.

6.7 Future directions

In line with the considerations discussed above, our directions for future work have three basic components. First, we will continue to develop the analysis method, improving the existing tests to target the desired linguistic information with greater precision, and expanding the range of information types that we are able to probe for. As we improve and expand the analysis method, the goal will be to increase our capacity to use these tests to draw specific conclusions about which model characteristics do and do not lead to satisfactory capturing of the relevant aspects of composition.

Also critical to being able to draw such conclusions is our second component of future work: broader and more systematic testing of composition models, in order to narrow down and identify classes of model characteristics that allow for capturing of desired properties.

Finally, the third component of future work is to develop new composition models to improve the state of the art in NLP. For this purpose, we intend to use not only insights gained by testing existing models using this analysis method—but also insights from linguistic theory and language acquisition—to identify and implement the most effective approaches to composing sentence meaning for NLP.

6.8 Conclusion

In this chapter we have shifted to discussing the relationship of lexical and sentence-level processes in the context of sentence composition, and in particular, the problem of composing complex lexical content in NLP systems. In this work we take a first step in approaching this problem by addressing a necessary prerequisite: establishing adequate assessments of composition in these types of models. To this end, we present an analysis method that is inspired by neuroscience analysis, allowing us to increase the interpretability of sentence vector representations produced by NLP systems, and to target and probe for specific types of information relevant to composition. We use this method to probe for two types of dependencies—semantic role and negation—in sentence vectors produced by sequence-based neural network models. Finally, we discuss limitations of this ap-

proach and directions for future work, as we move toward using this type of analysis in order to understand how best to accomplish composition of complex lexical content.

Chapter 7

Conclusion

In this dissertation I have explored aspects of the relationship between "lexical" and "syntactic" processes and representations, bridging approaches and questions from cognitive neuroscience, linguistics, and NLP. My use and definition of this lexical/syntactic distinction is motivated partly by the standard delineation between access of lexical content and syntactic composition of that content—but also by the observation of lexical processes in real-time comprehension which appear to be asyntactic in nature, and which therefore contrast with the syntactically-constrained interpretation processes (which I also refer to as "message-based" processes) that give rise to understanding of sentence meaning. One of the questions underlying the discussions in the previous chapters is the extent to which these versions of the distinction are tapping into the same neural mechanisms, and the extent to which they diverge.

7.1 Overview

In Chapter 1, I began the discussion by laying out the nature of the distinction of interest between "lexical" and "syntactic" processes, with a particular focus on illustrating the influences of asyntactic lexical effects that occupy much of our attention in subsequent chapters. Because of their apparent reflection of lexical content but not syntactic information, these effects lend themselves to interesting questions about what processes these effects reflect, what role they have in processing, and how they relate to the more structured interpretation processes that result in compositional sentence meaning. These effects are also significant in that it is important to understand the influences of such effects on our measured signal, relative to the influences of other types of mechanisms of interest.

In that chapter I also discussed the manifestation of this distinction within NLP, and the considerations that come with drawing parallels of this kind between human language comprehension processes and NLP systems. Among other things, we find that although a key difference in NLP is that it need not emulate coarser processes that arise from real-time comprehension, the reality is that representations and processes that exist in current NLP systems are often in fact more reflective of those real-time effects—like the influence of associative relations—than of the more structured meaning-based processes that we believe to be involved with the syntactically-driven mechanisms.

In Chapter 2, I delved deeper into the relevant considerations involved in modeling and reasoning about these issues in humans, and I explored the parallels that can be drawn between the human and NLP domains in this respect. I discussed three relevant dimensions

of the hypothesis space: word representation, combinatorial processes, and facilitation mechanisms, all of which are important to consider as we design models and hypotheses to account for human comprehension effects. I reviewed some classes of existing theories with respect to these dimensions on the human side, and introduced relevant parallel approaches from the NLP side. Finally, I discussed in detail the implications of using vector space models (VSMs), which are commonly used in NLP, for modeling aspects of human language processing. I discussed the fundamental properties of this representation framework, as well as the assumptions that I make in using distribution-based VSMs for cognitive modeling. In particular, I do not assume these vectors necessarily to represent word meaning content (as is often assumed in NLP), but only that they capture useful co-occurrence-based lexical relations which may align well with passive lexical priming-like processes. While the relations modeled by these vectors may also align with effects of a more active asyntactic lexical process, they lend themselves naturally to a more passive notion of these lexical processes.

In Chapter 3, I proceeded to demonstrate the use of these distribution-based VSMs as a means of modeling and testing hypotheses pertaining to lexical processes. First, I showed how these vectors could be used to capture graded lexical relations that are not subjectively apparent, but that may provide alternative explanations for observed patterns of result—using as a case study the N400 results observed by Federmeier and Kutas (1999). We see that these VSM-based models are able to account for key aspects of this result, thus presenting a viable alternative account, not in terms of the message-based prediction mechanism appealed to by those authors, but in terms of low-level lexical relations. This highlights the need to be able to model the potential influences of such

effects, in such a way that we can differentiate between them and the more structured processes that are often of interest.

In that chapter I then explored the capacity of these VSMs to model lexical processes outside of the sentence context, in the form of single-word semantic priming (priming by single-word contexts). This represents a different version of lexical processes than those we focus on elsewhere, not only because it represents effects from a non-sentential context, but also because it makes use of behavioral measures, by contrast to our focus on ERP components in all other cases. Nonetheless, it is useful in allowing us to check how well these VSMs align with lexical relations as manifested in semantic priming measures. We find that the vectors capture non-trivial variance in the reaction times, specifically in the case of the lexical decision task with a short SOA. This suggests that these vectors align best with processes that are fast and automatic, and that involve lexical access, which is consistent with our notion that they are well-suited for modeling fast, passive lexical processes that may be understood as automatic byproducts of lexical access.

In Chapter 4, I shifted to consideration not just of lexical processes, but of the interaction between these observed asyntactic lexical effects and the syntactically-constrained message-level processes that drive accurate interpretation. For exploring this interaction, we focus on the phenomenon of role reversals, in which the N400 component fails to show sensitivity to anomaly created by reversing the canonical roles filled by arguments, with the N400 instead appearing to be sensitive only to the lexical content of those arguments. I reviewed the core canon of role reversal results, as well as two connectionist computational models that have successfully simulated results from this literature.

I then reported the results of a replication of one of these models, from Brouwer et al. (2017), which claims to account for both the N400 and P600 components in response to role reversal anomalies as well as standard anomalies. We find that although the model's simulation of the N400 effect is reasonably robust—though as we point out, this particular pattern of N400 results can be captured by much simpler models—its simulation of the P600 is dependent on the specific distributions of test sentences in the training data, suggesting that it is reflecting the probabilities of those sentences themselves, rather than generalizing its effect based on meaning anomaly. This suggests that this model is able to capture lexical effects, but it falls short in capturing syntax-sensitive message-level effects.

In Chapter 5, I moved on to introduce two newer results that complicate the picture, in that they show the N400 to be in fact sensitive to role reversal anomalies under certain conditions. This presents us with a valuable opportunity to explore how lexical and message-based factors might interact, holding constant the relevant ERP component and developing theories based on the effects of variation in stimulus properties and timing. After reviewing these new ERP results, I presented a brief computational simulation demonstrating how one existing theory could potentially account for the variation. I then transitioned to reporting the results of an analysis of the relevant stimuli, aimed at identifying possible explanations for the contrast in outcomes between studies. This analysis gives rise to several hypotheses about the interactions of effects responsible for the observed patterns of results across studies, which we implement in the form of several corresponding models. These models represent the introduction of a novel hybrid modeling approach, combining corpus-based and human-based quantification of stimulus

characteristics. We then test these models in a series of simulations of the two studies of interest, followed by addition of a third study to test generalization. We find that two classes of hypothesis are able to account for the variations in outcomes: one class in which the results are attributable to lexical effects alone, with those effects being driven by the most recent content word in the context, and one class in which message-based expectations drive the N400 signal under select circumstances—when given enough time and when the context is sufficiently constraining—and lexical relations drive the N400 signal otherwise.

In Chapter 6, I shifted in two respects: first, I shifted to thinking about the interaction of lexical and syntactic processes within the context of semantic composition *per se*, which in previous chapters has fallen under the umbrella of message-based processes. Second, in service of discussing that topic, I shifted to focusing on composition as an engineering problem in NLP, by contrast to the scientific questions emphasized in the previous chapters. In particular, I zeroed in on the problem of composing complex lexical content (such as the vectors that we used for modeling lexical effects in previous chapters)—which is a problem that is often set aside in compositional semantics, but that is a critical issue to address when designing NLP systems.

In that chapter, I approached this problem by first tackling a necessary pre-requisite of improving composition in NLP systems: being able to evaluate how well those systems do composition in the first place. This is a challenge for many NLP systems, especially current systems that operate within a vector space representation framework, because they output representations in the form of opaque vectors, the content of which is not straightforward to interpret. I described a method the we have proposed and developed—inspired

by methods used in neuroscience to interpret brain activity—that allows us to target and probe for specific linguistic information relevant to composition. I presented results of experiments testing this method on existing NLP systems, and discussed implications and directions for future work.

7.2 Future directions

Exploring lexical representation In the modeling experiments reported here, we have held fixed the basic type of word representation used, accepting as an assumption of our simulations that these representations will capture the relevant lexical properties with reasonable accuracy. Apart from our comparison of slightly different models in Chapter 3—which yields the conclusion that those models perform very similarly—we do not attempt to vary our lexical representations systematically to test the most accurate way of thinking about the lexical representations that give rise to these effects.

As we have discussed, our use of distribution-based VSMs makes it doubtful that we are simulating lexical meaning *per se*. We have instead made the weaker assumption that these vectors could suitably approximate automatic co-occurrence-based lexical processes, which may be the correct characterization of the observed asyntactic lexical effects that motivated much of our work here.

In future work, it will be useful to explore this dimension of the problem in greater depth. In particular, we have not yet resolved the question of whether the representations or encodings that give rise to the asyntactic lexical effects are in fact the same lexical representations that participate in meaning composition. Do associative lexical effects

arise as a simple byproduct of accessing lexical representations for composition? Or do these effects reflect a separate, parallel processing mechanism of some kind? To what extent do VSMs capture information from only one or the other of these processes?

Along this line, it will also be important to consider other approaches to lexical representation, beyond VSMs—and particularly beyond distribution-based VSMs. This direction will be particularly relevant when we want to test hypotheses with respect to lexical meaning *per se*.

Expanding modeling range In Chapter 5 we saw a number of hypotheses that showed promise in accounting for a complex set of N400 results. The natural next steps for this work are to expand the range of studies that we attempt to simulate with the corresponding models, as this is a process that will inevitably force us to modify the hypotheses and their implementations, allowing us to further partition the hypothesis space. Relatedly, as we discussed in Chapter 5, there are aspects of these models for which we may be able to identify more principled approaches—such as the combination of cloze and cosine values and the use of noise to simulate subject variance—and with the corresponding adjustments, the performances of the models may shift.

We focused in that chapter on modeling the interacting contributions of facilitatory effects of these processes on the N400. Understanding these dynamics is critical to interpreting the amplitude of the N400 as an index of the functioning of the relevant processes, so this is an important component of improving our understanding of human sentence comprehension mechanisms. However an additional goal in moving forward will be to devote greater attention to the specific nature of the underlying mechanisms that comprise

these lexical and message-based processes. This will involve, for instance, differentiating between the active and passive notions of the asyntactic lexical processes, as well as pursuing a more sophisticated capacity to model the message-based processes computationally.

Improving sentence composition models Chapter 6 introduced and tested a method for analyzing linguistic properties in sentence vector representations produced by NLP models, as a prerequisite to improving those models in a targeted manner. In moving forward, one of the primary goals of this research program is to use insights from analyses of this kind in order to identify and implement more effective models of sentence meaning composition—and relatedly, to identify promising directions for theories of the human capacity for composition of complex lexical content. This brings us to our final point.

7.3 Bridging domains

In exploring these questions, I have drawn on tools and perspectives from across cognitive neuroscience, linguistics, and NLP, with the goal of identifying points of productive contact such that these domains can be enhanced by the mutual sharing of insights and methods.

For the most part, this bridging has taken the form of applying tools or methods from one domain to the other. In modeling of asyntactic lexical processes, I drew on VSM representations commonly used in NLP, in order to quantify lexical facilitation produced by these processes in the brain. In assessing composition in NLP models, I drew on methods

from cognitive neuroscience as well as insights from linguistics, in order to design tests to probe for evidence of semantic composition in opaque vector representations.

There is another way in which I envision these fields being bridged, which would take the form of common effort toward a shared goal. Because of fundamental differences in the goals of scientific and engineering domains, it is in many cases not reasonable to expect such an alignment of efforts. However, I believe that the problem of composing complex lexical content, discussed in Chapter 6, is an area in which such a state of affairs could plausibly be reached. This is a problem with fundamental questions that span the human and NLP domains: "What might lexical representations, and compositional operations, look like in order for these components to interact to produce systematic composition of phrase and sentence meaning? How can this be learned, and what pre-existing biases are necessary to make that possible?"

It is certainly in principle possible to arrive at an answer to how these things *can* be done that does not align with how they *are* done in humans. However, in the absence of a comprehensive solution or theory in either NLP or cognitive science, it seems that any such comprehensive solution to these questions would represent a substantial step forward both for NLP and for cognitive science. For this reason, I believe this to be a promising area for real collaborative connections to be drawn between these domains.

References

Abney, S. P., & Johnson, M. (1991). Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*, *20*(3), 233–250.

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2016). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19–27). Association for Computational Linguistics.

Arora, S., Liang, Y., & Ma, T. (2016). A simple but tough-to-beat baseline for sentence embeddings.

Atkins, B. T., Kegl, J., & Levin, B. (1988). Anatomy of a verb entry: From linguistic theory to lexicographic practice. *International Journal of Lexicography*, *1*(2), 84–126.

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*(2), 283.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., ... Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse* (pp. 178–186).

Bender, E. M. [Emily M.], Flickinger, D., Oepen, S., & Zhang, Y. (2011). Parser evaluation over local and non-local deep dependencies in a large corpus. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 397–408). Edinburgh, Scotland, UK.: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/D11-1037

Bender, E. M. [Emily M], Flickinger, D., Oepen, S., & Zhang, Y. (2011). Parser evaluation over local and non-local deep dependencies in a large corpus. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 397–408).

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*(Feb), 1137–1155.

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology*, *60*(4), 343–355.

Bentivogli, L., Bernardi, R., Marelli, M., Menini, S., Baroni, M., & Zamparelli, R. (2016). SICK through the SemEval glasses. lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Language Resources and Evaluation*, 1–30.

Berant, J., Chou, A., Frostig, R., & Liang, P. (2013). Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1533–1544).

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."

Blei, D. M., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *JMLR*, *3*, 993–1022.

Blunsom, P., Grefenstette, E., & Hermann, K. M. (2013). ''Not not bad" is not "bad": A distributional account of negation. In *Proceedings of the 2013 workshop on continuous vector space models and their compositionality*.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on semantic P600 effects in language comprehension. *Brain research reviews*, *59*(1), 55–73.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *EMNLP*.

Bowman, S. R., Gauthier, J., Rastogi, A., Gupta, R., Manning, C. D., & Potts, C. (2016). A fast unified model for parsing and sentence understanding. *arXiv preprint arXiv:1603.06021*.

Brachman, R. J., & Schmolze, J. G. (1988). An overview of the KL-ONE knowledge representation system. In *Readings in artificial intelligence and databases* (pp. 207–230). Elsevier.

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science*, *41*, 1318–1352. doi:10.1111/cogs.12461

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143.

Brown, C., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. *Journal of Cognitive Neuroscience*, *5*(1), 34–44.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, *18*(4), 467–479.

Chow, W.-Y., Lau, E., Wang, S., & Phillips, C. (2018). Wait a second! Delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience*, 1–26.

Chow, W.-Y., Momma, S., Smith, C., Lau, E., & Phillips, C. (2016). Prediction as memory retrieval: Timing and mechanisms. *Language, Cognition and Neuroscience*, *31*(5), 617–627.

Chow, W.-Y., Smith, C., Lau, E., & Phillips, C. (2015). A 'bag-of-arguments' mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, *31*(5), 577–596.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, *8*(2), 240–247.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*.

Costello, F. J., & Keane, M. T. (1997). Polysemy in conceptual combination: Testing the constraint theory of combination. In *Proceedings of the nineteenth annual conference of the cognitive science society* (pp. 137–142). Hillsdale, NJ: Erlbaum.

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

Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S. J., & Goodman, N. D. (2018). Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

Deacon, D., Hewitt, S., Yang, C.-M., & Nagata, M. (2000). Event-related potential indices of semantic priming using masked and unmasked words: Evidence that the N400 does not reflect a post-lexical process. *Cognitive Brain Research*, *9*(2), 137–146.

Demberg, V., Keller, F., & Koller, A. (2013). Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, *39*(4), 1025–1066.

Doran, C., Egedi, D., Hockey, B. A., Srinivas, B., & Zaidel, M. (1994). XTAG system: A wide coverage grammar for english. In *Proceedings of the 15th conference on computational linguistics-volume 2* (pp. 922–928). Association for Computational Linguistics.

dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 69–78).

Duffy, S. A., Henderson, J. M., & Morris, R. K. (1989). Semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(5), 791.

Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. A. (2016). Recurrent neural network grammars. *NAACL*.

Ehrenhofer, L. (2018). *Argument roles in adult and child comprehension* (Doctoral dissertation, University of Maryland).

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

Estes, Z., & Glucksberg, S. (2000). Interactive property attribution in concept combination. *Memory & Cognition*, *28*(1), 28–34.

Ettinger, A., Elgohary, A., & Resnik, P. (2016). Probing for semantic evidence of composition by means of simple classification tasks. *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, 134.

Ettinger, A., Rao, S., Daumé III, H., & Bender, E. M. (2017). Towards linguistically generalizable NLP systems: A workshop and shared task. *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1606–1615). Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/N15-1184

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469–495.

Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (wac-4) Can we beat Google* (pp. 47–54).

Fischler, I., Childers, D. G., Achariyapaopan, T., & Perry, N. W. (1985). Brain potentials during sentence verification: Automatic aspects of comprehension. *Biological Psychology*, *21*(2), 83–105.

Fodor, J. A. (1990). Information and representation. In P. P. Hanson (Ed.), *Information, language and cognition*.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and language*, *140*, 1–11.

Frankland, S. M., & Greene, J. D. (2015). An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences*, *112*(37), 11732–11737.

Franks, B. (1995). Sense generation: A 'quasi-classical' approach to concepts and concept combination. *Cognitive science*, *19*(4), 441–505.

Fyshe, A., Wehbe, L., Talukdar, P. P., Murphy, B., & Mitchell, T. M. (2015). A compositional and interpretable semantic space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Gagné, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier–noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 71.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211.

Guo, J., Che, W., Wang, H., & Liu, T. (2014). Learning sense-specific word embeddings by exploiting bilingual resources. In *Proceedings of COLING* (pp. 497–507).

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. *NAACL*.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1–8). Association for Computational Linguistics.

Hale, J. T. (2014). *Automaton theories of human sentence comprehension*. CSLI Publications.

Harris, Z. S. (1968). Mathematical structures of language.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual review of neuroscience*, *37*, 435–456.

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar*. Blackwell Oxford.

Herdağdelen, A., Erk, K., & Baroni, M. (2009). Measuring semantic relatedness with vector space models and random walks. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing* (pp. 50–53). Association for Computational Linguistics.

Hill, F., Cho, K., & Korhonen, A. (2016). Learning distributed representations of sentences from unlabelled data. In *NAACL*.

Hill, F., Cho, K., Korhonen, A., & Bengio, Y. (2015). Learning to understand phrases by embedding the dictionary. *arXiv preprint arXiv:1504.00548*.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Hoeks, J. C., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, *19*(1), 59–73.

Hofmann, T. (1999). Probilistic latent semantic analysis. In *UAI*.

Holcomb, P. J. (1988). Automatic and attentional processing: An event-related brain potential analysis of semantic priming. *Brain and Language*, *35*(1), 66–85.

Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the*

50th annual meeting of the association for computational linguistics: Long papers-volume 1 (pp. 873–882). Association for Computational Linguistics.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., ... Buchanan, E. (2013). The semantic priming project. *Behavior research methods*, *45*(4), 1099–1114.

Iyyer, M., Manjunatha, V., Boyd-Graber, J., & III, H. D. (2015). Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (pp. 1681–1691).

Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Jones, K. S. (1965). Experiments in semantic classification. *Mech. Translat. & Comp. Linguistics*, *8*(3-4), 97–112.

Jones, M. N., Kintsch, W., & Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552.

Kádár, Á., Chrupała, G., & Alishahi, A. (2016). Representation of linguistic form and function in recurrent neural networks. *arXiv preprint arXiv:1602.08952*.

Keller, F. (2010). Cognitively plausible models of human language processing. In *Proceedings of the ACL 2010 Conference Short Papers* (pp. 60–67). Association for Computational Linguistics.

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of memory and Language*, *52*(2), 205–225.

Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2016). Character-aware neural language models. In *AAAI* (pp. 2741–2749).

Kintsch, W. (2001). Predication. *Cognitive science*, *25*(2), 173–202.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems* (pp. 3276–3284).

Kolk, H. H., Chwilla, D. J., Van Herten, M., & Oor, P. J. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and language*, *85*(1), 1–36.

Kos, M., Vosse, T., Van Den Brink, D., & Hagoort, P. (2010). About edible restaurants: Conflicts between syntax and semantics as revealed by ERPs. *Frontiers in psychology*, *1*.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain research*, *1146*, 23–49.

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, 1–15.

Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, *17*(1), 117–129.

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and cognitive processes*, *8*(4), 533–572.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, *62*, 621–647.

Kutas, M., & Hillyard, S. (1989). An electrophysiological probe of incidental semantic association. *Cognitive Neuroscience, Journal of*, *1*(1), 38–49.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*.

Kutas, M., Lindamood, T. E., & Hillyard, S. A. (1984). Word expectancy and event-related brain potentials during sentence processing. In S. Kornblum & J. Requin (Eds.), *Preparatory states and processes* (pp. 217–237).

Lai, A., & Hockenmaier, J. (2014). Illinois-lh: A denotational and distributional approach to semantics. *Proc. SemEval*.

Lake, B. M., & Baroni, M. (2017). Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *arXiv preprint arXiv:1711.00350*.

Lakoff, G. (1969). On generative semantics.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Lapesa, G., & Evert, S. (2013). Evaluating neighbor rank and distance measures as predictors of semantic priming. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2013)* (pp. 66–74).

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of cognitive neuroscience*, *25*(3), 484–502.

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. *Concepts: core readings*, 3–81.

Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2015). Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*.

Li, J., Monroe, W., & Jurafsky, D. (2016). Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

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

Lynott, D., Tagalakis, G., & Keane, M. (2004). Conceptual combination with PUNC. *Artificial Intelligence Review*, *22*(3), 247–267.

Mandera, P., Keuleers, E., & Brysbaert, M. (2016). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., & Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *Language resources and evaluation* (pp. 216–223).

Masson, M. E. (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 3.

McClelland, J. L. [Jay L], St. John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and cognitive processes*, *4*(3-4), SI287–SI335.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115–133.

McDonald, S., & Brew, C. (2004). A distributional model of semantic context effects in lexical processing. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (p. 17). Association for Computational Linguistics.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. Psychology Press.

Meyer, D., & Schvaneveldt, R. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of experimental psychology*, *90*(2), 227.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Mitchell, J., & Lapata, M. (2008). Vector-based models of semantic composition. In *ACL* (pp. 236–244).

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, *34*(8), 1388–1429.

Mooney, R. (2004). Learning semantic parsers: An important but under-studied problem. In *Working notes of the aaai spring symposium on language learning* (pp. 39–44).

Mooney, R. J. (2014). Semantic parsing: Past, present, and future. In *ACL workshop on semantic parsing. Presentation slides*.

Morris, R. K. (1994). Lexical and message-level sentence context effects on fixation times in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 92.

Murphy, G. L. (1988). Comprehending complex concepts. *Cognitive science*, *12*(4), 529–562.

Neelakantan, A., Shankar, J., Passos, A., & McCallum, A. (2015). Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories.

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., ... Pallier, C., et al. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 201701590.

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, *19*(12), 1213–1218.

Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2), 161–199.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, *31*(1), 712105.

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011). Using language models and latent semantic analysis to characterise the N400m neural response. In *Proceedings of the Australasian Language Technology Association Workshop 2011* (pp. 38–46).

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Retrieved from http://www.aclweb.org/anthology/D14-1162

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological review*, *107*(4), 786.

Pustejovsky, J. (1987). On the acquisition of lexical entries: The perceptual origin of thematic relations. In *Proceedings of the 25th annual meeting on association for computational linguistics* (pp. 172–178). Association for Computational Linguistics.

Pustejovsky, J. (1991). The generative lexicon. *Computational linguistics*, *17*(4), 409–441.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral science*, *12*(5), 410–430.

Quillian, M. R. et al. (1962). A revised design for an understanding machine. *Mechanical translation*, *7*(1), 17–29.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2016). N400 amplitudes reflect change in a probabilistic representation of meaning: Evidence from a connectionist model.

In *Proceedings of the 38th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society*.

Reisinger, J., & Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Human language technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 109–117). Association for Computational Linguistics.

Reiter, E., Dale, R., & Feng, Z. (2000). *Building natural language generation systems*. MIT Press.

Resnik, P. (1992). Left-corner parsing and psychological plausibility. In *Proceedings of the 14th conference on computational linguistics-volume 1* (pp. 191–197). Association for Computational Linguistics.

Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, *61*(1), 127–159.

Resnik, P., & Diab, M. (2000). Measuring verb similarity. In *Proceedings of the 22nd annual meeting of the cognitive science society* (pp. 399–404). Philadelphia, PA.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). ACM.

Rimell, L., Clark, S., & Steedman, M. (2009a). Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 813–821). Singapore: Association for Computational Linguistics.

Rimell, L., Clark, S., & Steedman, M. (2009b). Unbounded dependency recovery for parser evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 813–821).

Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, *8*, 627–633.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of verbal learning and verbal behavior*, *9*(5), 487–494.

Rumelhart, D. E., Hinton, G. E., McClelland, J. L. [James L], et al. (1986). A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, *1*, 45–76.

Rumelhart, D. E., McClelland, J. L. [James L], Group, P. R., et al. (1986). Parallel distributed processing, vol. 1. MIT Press, Cambridge, MA.

Salveter, S. C. (1979). Inferring conceptual graphs. *Cognitive Science*, *3*(2), 141–166.

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of EMNLP*.

Schuler, K. K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon.

Siskind, J. M. (1990). Acquiring core meanings of words, represented as Jackendoff-style conceptual structures, from correlated streams of linguistic and non-linguistic input. In *Proceedings of the 28th annual meeting on association for computational linguistics* (pp. 143–156). Association for Computational Linguistics.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1-2), 39–91.

Smith, E. E., & Osherson, D. N. (1984). Conceptual combination with prototype concepts. *Cognitive science*, *8*(4), 337–361.

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1201–1211). Association for Computational Linguistics.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1631–1642).

Tian, F., Dai, H., Bian, J., Gao, B., Zhang, R., Chen, E., & Liu, T.-Y. (2014). A probabilistic model for learning multi-prototype word embeddings. In *Proceedings of COLING* (pp. 151–160).

Triantafillou, E., Kiros, J. R., Urtasun, R., & Zemel, R. (2016). Towards generalizable sentence embeddings. In *Acl workshop on representation learning for nlp* (p. 239).

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*(1), 141–188.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.

Van de Cruys, T. (2014). A neural network approach to selectional preference acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 26–35).

Van Herten, M., Kolk, H. H., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research*, *22*(2), 241–255.

Van Petten, C. (1993). A comparison of lexical and sentence-level context effects in event-related potentials. *Language and Cognitive Processes*, *8*(4), 485–531.

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A., & Mikolov, T. (2015). Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Wieting, J., Bansal, M., Gimpel, K., & Livescu, K. (2015). Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.

Wilks, Y. (1975). A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, *6*(1), 53–74.

Williams, A., Drozdov, A., & Bowman, S. R. (2017). Learning to parse from a semantic objective: It works. Is it syntax? *arXiv preprint arXiv:1709.01121*.

Wisniewski, E. J. (1996). Construal and similarity in conceptual combination. *Journal of Memory and Language*, *35*(3), 434–453.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems* (pp. 649–657).

Zhao, Z., Dua, D., & Singh, S. (2017). Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*.

Zhu, X., & Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision* (pp. 19–27).