SUPPORTING WORD LEARNING WITH LANGUAGE-INTERNAL DISTRIBUTIONAL STATISTICS: A PLACE FOR THE RECURRENT NEURAL NETWORK LANGUAGE MODEL?

BY

PHILIP A. HUEBNER

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Doctoral Committee:

Professor Jon Willits, Chair
Professor John Hummel
Professor Aaron Benjamin
Professor Cynthia Fisher
Professor Gary Dell

Abstract

Prior work has demonstrated that statistical dependencies between words in language input can be used to construct word clusters broadly conforming to lexical classes in adult language (Cartwright & Brent, 1997; J. L. Elman, 1990; T. H. Mintz, 2003; Redington et al., 1998), and that children use this information to guide inferences during word learning in the absence of perceptual information (Lany & Gómez, 2008; Lany & Saffran, 2011; Wojcik & Saffran, 2015). Building on these insights, this thesis examines whether the simple Recurrent Neural Network (simple RNN) could be used to model children's acquisition of form-based lexical semantic category knowledge and whether this knowledge could be used to help children infer category-associated features of novel words. In order to determine the feasibility of the RNN as a cognitive model of this procedure, I discuss several desiderata concerning how corpus-derived distributional semantic statistics should be encoded and accessed in the network, and undertake comprehensive simulations that address basic questions concerning the mechanism by which the RNN acquires lexical semantic category knowledge, and how learned representations are influenced by the statistical properties of the input.

In particular, I show that the construction of form-based lexical semantic representations by the simple RNN is extremely vulnerable to a particular kind of redundancy, which occurs when an item in the left context can be reliably used to predict an item in the right context of a target word. This co-occurrence pattern in the data allows the RNN to 'ignore' the intervening target word, yielding semantically impoverished representations that are less useful for guiding children's inferences during word learning. In order to better understand and overcome this limitation, I developed a theory that formalizes how the training data, learning dynamics, and training strategy conspire to shape lexical semantic representations in the RNN. Semantic Property Inheritance (SPIN) theory makes recommendations for how to choose training data that maximizes the acquisition of statically accessible lexical semantic category knowledge. In particular, the theory is concerned with atomicity, which requires that the (distributional) semantic properties of a target word are encoded in the representation of the target word as opposed to words that also occur in the same sentence. Further, SPIN theory predicts that training the RNN on child-directed transcribed speech that has been ordered by the age of the target child results in more atomic lexical semantic representations for nouns than training in reverse order. I test and confirm this prediction, and discuss implications of this finding for child language acquisition, the importance of studying model learning dynamics, model-data interactions, and the gradual refinement of learned representations over the course of training on non-stationary data.

To my high-school English teacher who first introduced me to cognitive science.

Acknowledgments

I am grateful to many people who have contributed to the work herein, and have supported me in my scientific endeavors, personally and academically. I am particularly grateful for my doctoral advisor Dr. Jon Willits who opened my eyes to the world of cognitive science. Our early discussions concerning long-standing theoretical issues and philosophy of science, provided me with the kind of inspiration and drive that still motivates me today. Being the first graduate student in his laboratory, there were many logistic and technological hurdles to overcome; Jon trusted me to make my own decisions, and provided me with the resources and space to develop my own perspective on controversial scientific questions. While an environment of creative freedom has often led me down one or more rabbit holes, Jon has a special knack for getting me back on my feet.

I am also grateful for the input provided by my doctoral committee, which, apart form Jon Willits, also includes Cynthia Fisher, Jon Hummel, Gary Dell, and Aaron Benjamin. Their early feedback on drafts of my thesis and presentations have challenged me to make connections between the world of machine learning and cognitive psychology — an area I believe to be full of potential for understanding how human brains make sense of the world.

I would also like to acknowledge people who have helped me develop a scientific mindset during my undergraduate education at the University of California, Davis. In particular, I want to thank Mirna Lechpammer for bringing me on board of her neuro-pathology research team at the UC Davis Medical Center, and Evan Fletcher, who first introduced me to the world of neuro-scientific research at the UC Davis Imaging of Dementia and Aging (IDEA) lab. I also extend my gratitude to other faculty who have opened their laboratories to me, and the Neurobiology, Physiology and Behavior club at the University of California, Davis, for exposing me to inspirational people and ideas.

At the University of Illinois, Urbana-Champaign, I am particularly indebted to Prof. H. F. Köhn, who has inspired me and many others to think deeply and rigorously about probability and statistics, and Shufan Mao for many stimulating intellectual discussions on the topic of computational semantics.

I would also like to acknowledge thinkers whose written works have inspired and influenced my scientific development and introduced me to fundamental issues in neuroscience, psychology, and artificial intelligence. Among these writers are Oliver Sacks, V.S. Ramachandran, Robert Sapolsky, Douglas Hofstadter, Michael Spivey, the Parallel Distributed Processing (PDP) Research Group (Jeffrey L. Elman, James McCelland, David E. Rumelhart, and others), Michael Tomasello, Steven Pinker, Carl Sagan, Antonio Damasio and Paul Bloom.

Finally, I would like to extend gratitude to my mother whose love and care has allowed me to flourish into the person I am today, my father, for supporting me despite great odds and across great distance, and my long-time partner, Bethany Judson, for sticking with me through 8 years of graduate school.

Table of contents

List of Abbreviations

AO-CHILDES	Corpus of Age-Ordered transcripts obtained from the CHILDES database
DECAF	Distributionally-Mediated Extension of Category-Associated Features
ED	Effective Dimensionality
DS	Divergence from Superordinate
LSTM	Long Short-Term Memory network
POS	Part-of-Speech
PSS	Perceptual Symbol Systems theory
RNN	Recurrent Neural Network
SG	Sentence Gestalt model
SGD	Stochastic Gradient Descent
SPIN Theory	Semantic Property Inheritance theory
SVD	Singular Value Decomposition

Chapter 1

Introduction

Over the course of several years, children rapidly acquire the meanings of thousands of words in their native language (P. Bloom & Markson, 1998; Golinkoff et al., 2000; Medina et al., 2011). How they do this is an enduring mystery to language acquisition researchers, computational modelers, philosophers, and linguists. One explanation for children's rapid progress is their ability to infer aspects of novel word meanings in the face of great referential uncertainty (e.g. what does the novel word label in the referential context?), and in the absence of direct perceptual experience (e.g. what does the referent look like, feel like, etc.). While many solutions have been proposed to account for children's success in the former situation (the referent labeled by the novel words is present in the extra-linguistic context), much less is known about how children deal with the latter situation (the referent is completely absent). In this thesis, I use computational modeling to examine whether the linguistic contexts in which words occur provide information that children could use to guide their inferences about word meaning, and whether this information can be acquired and used by a statistical learning system based on next-word prediction.

The development of lexical semantic knowledge (i.e. knowledge about word meanings) is an extremely complex phenomenon, that involves input from all perceptual modalities, makes use of many psychological processes (Bowerman & Choi, 2001; E. V. Clark, 2017a; Golinkoff, 1975; Murphy, 2004), and is constrained by multiple inductive biases (P. Bloom & Markson, 1998; Golinkoff et al., 2000; S. S. Jones et al., 1991a). A fundamental question in lexical semantic development is how children's inferences about novel word meanings succeed in the face of large amounts of noise and statistical uncertainty that characterize the language learning environment (Pinker, 1987; Siskind, 1996). For instance, caregivers only rarely engage in ostensive labeling, whereby objects in the referential context are directly and explicitly labeled in the target language (Callanan, 1985). It is thought that to compensate for the uncertainty that arises in the absence of ostensive labeling, children either engage in (i) 'cross-situational learning' (Fazly et al., 2010; Locke, 1847; Siskind, 1996; Yu & Smith, 2007) and/or 'fast-mapping' (Medina et al., 2011). First, cross-situational learning refers to tracking associations between words and objects across multiple referential contexts. In noisy natural environments that license a large number of potential mappings, cross-situational learning would enable children to converge on the correct word-referent mapping by accumulating word-referent co-occurrence statistics across many learning episodes. Support for this idea comes from behavioral evidence showing that children compute distributional statistics across the co-occurrences of words and referents at multiple moments (Fazly et al., 2010; Räsänen & Rasilo, 2015; Siskind, 1996; Yu & Smith, 2007). Second, fast-mapping provides an alternative account, which claims that children use sophisticated inductive biases to hypothesize only a single meaning and retain their hypotheses until disconfirmed. In contrast to cross-situational learning, alternative hypothesized meanings are not retained.

Which of these two word learning strategies ends up providing a better account of children's word learning is inconsequential to the work presented in this thesis. Neither strategy is useful in situations in which a novel word is uttered in the absence of perceptual information about its referent. Without the possibility of mapping the novel word to an object in the referential context, what can children do? On the one hand, they might not learn anything at all, or, at worst, form incorrect hypotheses; on the other hand, it is likely that children exploit *language-internal* information to induce candidate semantic properties for words presented alongside unhelpful referential contexts. For example, it is well known that children infer aspects of novel word meanings — primarily verbs — using syntactic cues such as transitivity and the number of nominal arguments (Fisher et al., 2010; L. Gleitman, 1990). But there is more: Emerging evidence in the field of language acquisition and statistical learning suggests that children may leverage finer-grained language-internal distributional information — how words pattern in fluent speech — to group words into lexical semantic categories (B. Ferguson et al., 2014; Lany & Saffran, 2010; Unger et al., 2020). These distributionally constructed lexical semantic categories could guide children's inferences about novel word meanings based entirely on the linguistic context in which novel words are presented. This thesis zooms in on this idea. Specifically, I aim to answer the following questions: Could a statistical learning system, based on tracking how words relate to other words in the input children hear, produce knowledge that is useful for inferring category-associated semantic features of novel words? For instance, the linguistic context in which a novel word is presented may provide cues about whether the novel word is a member of the distributionally constructed category ANIMAL or VEHICLE. If so, this would provide preliminary evidence that children's inferences about word meanings would benefit from tracking co-occurrence associations between the words they hear. Using computational modeling, and analyses of children's language input, this thesis aims to formalize and organize our understanding of the potential mechanism and role of language-internal distributional learning in children's inferences about novel word meanings.

1.1 Learning to Map Words onto Meaning

Lexical semantic development requires the taming of several challenging induction problems. The observation that word learning is at first slow and laborious and takes off at twenty months (an average of ten new words per day) suggests that children begin to use formulas or strategies that build on their previous successes. These formulas may be innate or learned; in either case, the learner quickly converges onto strategies that facilitate his or her word learning. The process of word learning can be broken down into smaller problems, each requiring a unique strategy: First, the infant needs to learn what perceptual signals are most important when considering which aspects of the referential context are most likely to be labeled by a novel word. This is the classic mapping problem: How does a child know that the word *rabbit* labels the whole object rather than a property, part, or action performed by the labeled object (Quine, 2013)? More broadly, how does the child know that *rabbit* refers to the rabbit and not another co-present object at all? Next, how should learned words be extended to novel objects? Is a dog without four legs still referred to as *dog*, and should the word *ball* also be used to refer to a drawing of a ball on a piece of paper? Lastly, how do children infer the meanings of novel words in the absence of a referent or perceptual information related to the referent? This question is particularly perplexing when considering that many abstract words do not have highly imageable counterparts in the natural world — at least not in a format that can be straightforwardly

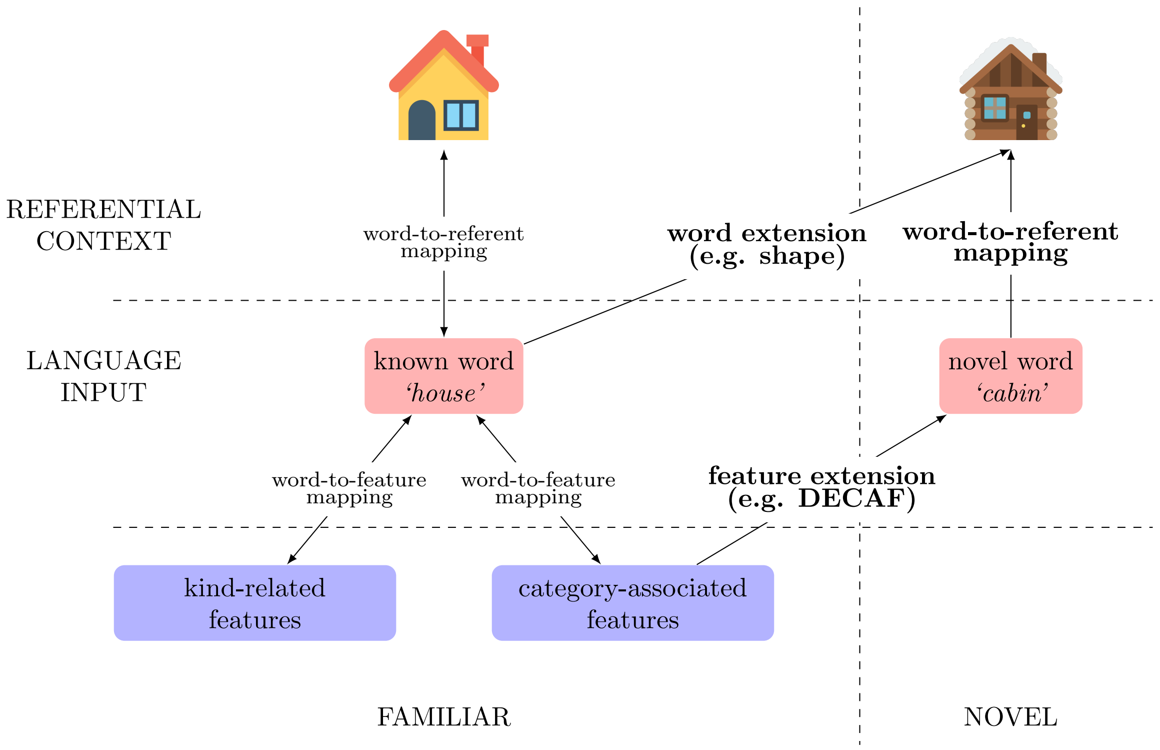


Figure 1.1: Three broad classes of inductive problems related to children's lexical semantic development. **Word-to-referent mapping:** Which object does a novel word label? **Word extension**: What additional objects does a known word label? **Feature extension**: Which known category-associated features are most relevant to a novel word? In this situation, the only cue available to the learner is the sentential context in which a novel word occurs. There are two mutually compatible strategies applicable here: First, children may perform, what I have termed, the distributionally-mediated extension of category-associated features (DECAF) — leveraging distributional similarity between a novel word and known words to extend familiar semantic features of known words to the novel word. By 'features', I refer to aspects of meaning, broadly construed. A second strategy is syntactic bootstrapping — leveraging syntactic cues such as part-of-speech and number of verb arguments to infer the ontological category of a novel word. This requires the availability of universal links between universal and syntactic categories (e.g. verb ↔ action, adjective ↔ property). Icons are made by Vector Market, and Smashicons, and obtained from www.flaticon.com.

extracted from the referential contexts in which language learning takes place. I review several strategies that children use to reduce uncertainty in each of these situations. Much ink has been spilled on the first two problems, word-referent mapping, and novel word extension. Instead, my focus is on the third problem, namely induction of word meanings for words in the absence of perceptual information. In particular, I focus on the supporting role of language-internal distributional statistics, beyond coarse-grained syntactic cues. A primary aim of this thesis is to test the feasibility of using distributional statistics to guide and constrain the extension of previously acquired semantic category-associated features to novel words.

To provide an overview of the aforementioned induction problems, and to situate learning from language-internal distributional cues relative to these problems, consider the schematic shown in Figure 1.1. I distinguish between familiar and novel information (right or left of the vertical divider, respectively), and between information available in the referential context and language input (top or bottom of upper horizontal divider, respectively). The bidirectional arrows indicate learned mappings, while the unidirectional arrows labeled with bold text illustrate the three induction problems. This thesis is most concerned with the induction problem labeled by 'DECAF', which stands for distributional extension of semantic category-associated features — the use of language-internal distributional evidence for extending known word features to novel words. DECAF is in some ways the inverse of the word extension problem: Whereas word extension is concerned with mapping known labels onto novel objects, DECAF is concerned with mapping known semantic features onto novel words.

1.1.1 Perceptual Cues

An obvious place to look for the meanings of words is in the world around us. However, this is easier said than done. The natural world is a complicated, dynamic place that does not lend itself to straightforward segmentation into discrete meaningful units. In fact, there is extensive cross-cultural diversity in the kinds of objects and events that are picked out by languages and linguistic conventions. Not knowing these conventions, however, a child is faced with the daunting task of segmenting what she receives via her senses, and mapping these packages of perceptual information onto sound categories (words). Luckily, there are several perceptual cues that reliably serve to narrow which components of the natural world are better referents than others. For one, the child is unlikely to entertain hypotheses that extend beyond the referential context, the perceptual information directly available to the child in the situation in which novel words are encountered. Within the bounds of the referential context, the child may further narrow her hypotheses by noticing that perceptual information tends to cluster in space, within object boundaries (Baillargoen, 1993; Needham & Baillargeon, 1997). Some scholars have suggested that this straightforward availability of whole objects to perception may, in part, underlie children's early success in mapping novel object labels to their referential counterparts (Gillette et al., 1999; Snedeker et al., 2001). This might explain why children's earliest learned words are primarily object labels, rather than labels for actions, which are less easily perceived (for useful discussions, see e.g., L. Bloom et al., 1993; Gentner, 1982; Landau and Jackendoff, 1993; Pylyshyn, 2000). Others have suggested that the tendency of children to map novel words onto objects as opposed to object properties or the events in which they are participating, is due to the whole-object bias — a cognitive bias that pushes children to interpret novel words as labels for whole objects (Markman, 1990).

There is an ongoing debate about how the whole-object bias arises. While some claim it is already available at birth as a result of cognitive limitations and/or conceptual biases (Markman, 1990), others have claimed that it is acquired via experience with object-label associations (Gillette et al., 1999; L. B. Smith et al., 2002; Snedeker et al., 2001). Scholars who have proposed the latter, emphasize that whole objects strongly cohere in the visual input, and that it is this observation that prompts children to map novel words onto objects. If true, this would imply that word extension is not necessarily a problem of conceptual capacity or biases, but a problem of information availability. Support for this idea comes from the Human Simulation Paradigm, an experimental paradigm in which adults are asked to infer which object in a visually presented scene is most likely labeled by a parent also present in the scene. Importantly, adults are not provided information about the words uttered bye parent, to simulate the situation in which a language learning takes place. Interestingly, adult subjects reproduce the characteristics of early learning in infants, namely slow and errorful learning, with correct inferences largely restricted to nominal terms. Additional evidence in support of an experience-dependent account comes from the observation that the shape bias increases as vocabulary size grows, and infants who know more words that are extended on the basis of shape learn count nouns more quickly.

The primacy of whole objects, whether perceptually or cognitively mediated, supports children's word-to-referent mapping (shown in the top right of Figure 1.1). However, perceptual cues also help guide children's inferences in a different kind of situation, word extension: What other objects does a known word label? In Figure 1.1, this situation is illustrated by the arrow from the known word *house* to the novel referent (a cabin). A perceptual cue that is particularly useful for word extension is object shape (Landau et al., 1988). Children attend to shape at an early age, and and often use shape at the expense of other perceptual dimensions in word extension experiments (Landau et al., 1992). Attending to shape, as opposed to color, texture, or material is a useful strategy, because many labels for objects are in fact based on shape; for instance, dogs have a characteristic shape, and so do cups, and cars. Thus, a child might extend the word *house* to refer to a cabin, based on shape similarity, despite potential large differences in other perceptual properties like size, color, and texture. In addition, the shape bias is useful because it prevents extending known words to objects that are involved in the same event (e.g. referring to birthday presents as birthday cakes). Of course, the shape bias is little more than a heuristic and children must rely on many other perceptual dimensions to acquire object labels that do not pick out concepts characterized by similar shapes. For instance, labels for superordinate category concepts, such as animal and furniture cannot be straightforwardly characterized in terms of commonalities in shape; instead, these concepts appear to refer to commonalities in the functionality or purpose. To deal with issues of this sort, children become more sophisticated in choosing when they apply the shape bias. For instance, S. S. Jones et al. (1991b) found that young children have considerable knowledge about conditional relations between kinds of perceptual properties. The authors found that both 2- and 3-year-old children classified eyeless objects by shape, and objects with eyes by both shape and texture. This demonstrates that children have successfully associated eyes with certain object kinds, which do not lend themselves to shape based classification alone.

1.1.2 Social Cues

Similar to perceptual cues, social cues can also narrow the set of candidate referents that are likely labeled by a novel word. For instance, some scholars have suggested that children implicitly understand that speakers have mental lives, and seek to communicative their intentions (Tomasello, 1999). This prompts children to look for social cues produced by the speaker that might help them narrow their search for referents. One such social cue is the speaker's direction of gaze. Brooks and Meltzoff (2005) showed that 9-month-olds, while sensitive to the orientation of the body of speakers, do not yet follow their gaze. The authors report that a developmental shift in gaze following occurs between 9 and 11 months of age. At 11 months, infants followed adult head turns significantly more often when the adult's eyes were open than closed. Further, the authors also showed a strong positive correlation between gaze-following at 11 months and subsequent language scores at 18 months, which suggests that gaze-following is a useful strategy that helps children tame the word-referent mapping problem. During the second year, children often look at an object at approximately the same time as a speaker is mentioning it. By 18 months, children's eye gaze need no longer coincide with that of a speaker on the same object to benefit acquisition; instead, children are able to map a word to the object a speaker was looking at even when the child herself was examining a different object during that time. In addition, children are sensitive to pointing and gestures, and understand early on that these behaviors serve similar communicative functions as gaze direction (H. H. Clark, 1996).

In addition to external social cues provided by the speaker, there are internal cues. For instance, it is possible that the representations a child already has in mind given the available discourse context may narrow the possible meanings he or she considers when hearing a sentence. Along these lines, L. Bloom et al. (1993) discuss the importance of labeling objects or concepts that are most relevant to the learner. L. Bloom et al. (1993) argue that children benefit the most when a speaker uses language to refer to concepts that are already in the child's mind at the onset of the utterance. That is, word learning is greatest when both the child and speaker consider the social and pragmatic context in which words are uttered, and align their expectations and attention on a clearly circumscribed topical domain that is clearly accessible and relevant to both. L. Bloom et al. (1993) draws on work by (Sperber & Wilson, 1986) to define relevance:

> Relevance is the single property that makes information worth processing and determines the particular assumptions an individual is most likely to construct and process.

On this view, the relevance of a novel word to a child is determined by what he or she is thinking and feeling about the shared focus of attention with the speaker. This means that the internal mental life of the child is itself potentially narrowing his or her word learning hypotheses.

1.1.3 Syntactic Cues

There are times when a child must reverse the direction of inference: In addition to being able to infer which known label can be extended to a novel object, it is equally important to be able to infer which aspects of previously acquired meanings might generalize to novel words. In Figure 1.1, this situation is illustrated by the arrow on the lower right. To do so, learners can, and do, use the syntactic contexts in which a novel word occurs to make inferences about the ballpark meaning of the novel word (P. Bloom & Kelemen, 1995; R. W. Brown, 1957; L. Gleitman, 1990). Syntactic contexts provide at least two kinds of information: First, children are often able to infer the syntactic category of a novel word. This information is systematically, but imperfectly, linked to broad semantic (ontological) distinctions. For example, when presented a word that describes a doll and told that '*this is a zav*', 1-year-olds can use the syntactic category of the novel non-sense word to work out its meaning (Katz et al., 1974). Specifically, children interpret *zav* as a common noun and extend it to other similar dolls. If instead, a word is used in a context such as '*This is zav*', they interpret the novel word as a proper noun that is naming the individual, as in *Stella*. In this case, children refuse to extend the *zav* to other similar dolls, in line with a proper noun interpretation. One of the earliest studies of this kind was carried out by R. W. Brown (1957) who exposed preschoolers to visual depictions of actions being performed on a novel substance with an unfamiliar object. One group of children (in the count noun syntax condition) was told: '*Do you know what a sib is? In this picture, you can see a sib*'; a second group (in the mass noun syntax condition) was told: '*Have you seen any sib? In this picture, you can see sib*'; and a third group (in the verb syntax condition) was told: '*Have you seen sibbing? In this picture, you can see sibbing*'. Children's construal of the novel non-sense word *sib* was in line with the syntactic contexts in which it occurred. Those exposed to count noun syntax guessed a *sib* is an object; those in the mass noun syntax condition judged the word as referring to a substance, and children who heard the verb in a verb frame tended to interpret the word as referring to an action. A primary conclusion drawn from this research is that syntactic cues guide young learners to correct inferences about a novel word's ontological category. Follow-up work showed that by 2-and-a-half years of age, children are also able to extend novel words in adjective frames ('*this is a daxy one*') to objects that share a common property, in line with the ontological status of adjectives as labels for properties (Klibanoff & Waxman, 2000; T. Mintz & Gleitman, 1998; Syrett & Lidz, 2010). The ability of young children to leverage syntactic category cues to infer approximate meanings of novel words (i.e ontological category) has been widely replicated, and is thought to be supported by innate knowledge of universal links between syntactic and semantic categories (Landau et al., 2009; L. Naigles, 1990; Pinker, 1984).

Second, syntactic cues are particularly useful for learning about broad semantic categories of verbs. The reason is that verbs and other predicates tend to occur in richer structural contexts compared to other content words. For example, different semantic classes of verbs require a different number of arguments, and the number of arguments constrains the kinds of meanings that a verb has. For example, verbs related to self-generated motion (e.g., 'dance') often appear with just one argument (the subject), whereas verbs of contact (e.g. *hit*) or caused motion (e.g. *push*) tend to require both a subject and a direct object. In addition, verbs whose meanings are related to transfer (e.g. *send*) more often appear with three arguments (subject, direct object, and indirect object). It is now well known that children use these links between syntax and semantics to rapidly generalize syntactic knowledge to novel words (Fisher et al., 2010; L. Gleitman, 1990). For example, 2-year-olds expect the verb in '*she is blicking her around*' to have a meaning related to 'pushing', and expect the verb in '*they are blicking around*' to mean something like *dancing*. The difference is that the former verb occurs in a transitive utterance, whereas the latter occurs in an intransitive utterance. Transitivity, therefore, is considered a reliable syntactic cue that young learners can use to acquire verb meanings. Knowledge of transitivity is thought to emerge by collecting information about the number of nominal arguments of a verb, and establishing a correspondence between these nouns and participants of the event described by the verb. Fisher et al. (2010) argue that this is helped by an "unlearned bias to map nouns in sentences onto participant roles in events". The ability to use syntactic cues and unlearned knowledge about how syntactic phenomena map onto semantic categories to infer aspects of novel word meanings, is called 'syntactic bootstrapping'. The reason this ability is called 'bootstrapping' is because it allows children to break into language, so to speak, with virtually no prior knowledge about the meanings of words. This implies that children should be able to use the syntactic contexts in which a verb occurs before they know anything about the verb's semantic content. Indeed, Fisher et al. (2010) and many others have found strong evidence that children do make use of abstract structural cues to verb meaning and in the absence of prior knowledge about lexical meanings.

1.1.4 Distributional Cues

A fourth cue that a child can use to support their inferences about word meanings is the lexical context in which a novel word is presented. In contrast to syntactic cues discussed above, in this section, I consider non-syntactic, statistical dependencies between lexical items. The difference between a syntactic cue and a distributional cue is that the former signals membership in a syntactic category (e.g. noun, verb) or phenomenon (transitive vs. intransitive), while the latter is not limited to syntactic class membership. Whereas syntactic cues are derived from the syntactic backbone of sentences, independent of the actual semantic content, distributional cues are based entirely on statistical associations between lexical items, and need not consider the syntactic backbone in any principled manner. The potential advantage of using distributional cues in addition to syntactic cues is that they go well beyond coarse-grained syntactic phenomena and the broad conceptual distinctions that they are correlated with. Identifying that a novel word is a noun is no doubt useful to a learner (syntactic cue), but learning that the novel noun occurs, say, in a distributional context that tends to be filled by animal words provides even more information (distributional cue).

How might a child exploit distributional cues to support their lexical semantic development? There are several steps involved in this procedure: First, the child must identify the linguistic context in which the novel word is presented. Second, the child must retrieve words that tend to occur in similar linguistic contexts. Next, the child must isolate semantic features shared by distributionally similar words. Because shared features tend to be diagnostic of semantic category membership (i.e. they are not unique to any particular word), I refer to these as ‘category-associated features’.[1] Finally, the child can extend those features to the novel word. Ideally, the results of this procedure is that the child correctly infers the semantic category of the novel word. To illustrate, consider a child has heard the word *gorilla* in an otherwise familiar utterance, but does not yet know what *gorilla* means. After performing the above steps, the child should have constructed a new lexical entry for *gorilla* that is populated with semantic features that identify it as a member of a distributionally constructed lexical semantic category that correlates with being an animal. I will refer to this procedure as the ‘**d**istributionally-mediated **e**xtension of **c**ategory-**a**ssociated **f**eatures’ (DECAF). I will use this term to refer to induction that is supported by distributional evidence alone, rather than by more sophisticated conceptual reasoning abilities or innate knowledge about syntax-semantics linkages.

While in principle possible, it is not clear whether DECAF is something that children actually do, or the extent to which distributional cues in the language children hear are sufficiently informative to guide inferences about word meaning. The primary aim of this thesis is to establish the feasibility of this approach using computational modeling: How much information about lexical semantic categories does a cognitively and developmentally plausible algorithm capture when trained on a realistic corpus of transcribed speech to children? There are at least two reasons to think that children make use of distributional cues in word learning: First, prior work has shown that distributional cues exist in input to children, and that they reliably correlate with semantic information that is useful to the language learner (Asr et al., 2016; Riordan, 2007). This thesis builds on this prior work by examining what semantic similarity structure emerges in a more cognitively plausible learning system that is trained in a more developmentally friendly manner. If it turns out that the statistical associations learned by such a system correlate well with semantic category distinctions between words, this would be preliminary evidence that children would benefit by performing a procedure like DECAF. Second, it is well known that children already track linguistic contexts in which words occur to build representations of the syntactic phenomena in their native language. If true, then children are already collecting distributional information about the language they hear. Thus, it is highly plausible that children would use the same machinery to track even finer-grained distributional statistics in order to uncover more subtle distributional correlates beyond broad ontological distinctions.

Distinguishing Syntactic Bootstrapping from DECAF

Given that learning from distributional cues is the primary topic of this thesis, it is worth further exploring the difference between distributional and syntactic cues. While both cues are language-internal, the way in which they are used is very different. In order for a syntactic cue to be useful during syntactic bootstrapping, the child need not have learned what semantic category is correlated with the syntactic cue. This correlation is already available in the form of innate knowledge about universal syntax-semantics linkages. In contrast, in order for a distributional cue to be useful, the child cannot rely on innate knowledge. He or she must construct associations between distributional and semantic information in an experience-dependent manner. Therefore, it cannot be said that distributional cues are universal, or that they hook into innate knowledge about linkages between linguistic structure and broad conceptual distinctions. In fact, distributional cues are by their very nature, language-specific, and require extensive exposure to language input in order to be useful. As a result, DECAF cannot be used to ‘bootstrap’ language acquisition. The use of syntactic

[1] A category-associated feature is a feature that is shared by many or all members of the same or related semantic category. For instance, virtually all animals sleep, breathe, and eat. In contrast, other features are more specific to individual members (e.g. birds ↔ wings, fish ↔ fins).

cues for syntactic bootstrapping is meant to help a learner break into a language with little to no prior knowledge of the meaning of words (the child may need to know a few 'seed word' to get the process started, learned by brute-force). This is in stark contrast to DECAF which is only useful after a learner has broken into a language, and has acquired a respectable inventory of word meanings, that he or she can extend to novel, distributionally similar words. Because syntactic bootstrapping occurs early in acquisition, and DECAF would occur much later, these two inductive procedures are mutually compatible, and in no way are competing theories of word learning. In fact, these two procedures likely share cognitive resources: If sensitivity to syntactic cues is in part guided by distributional analysis of the linguistic contexts in which words occur, does the machinery that underlies children's syntactic bootstrapping overlap with the machinery used to track finer-grained and more open-ended distributional statistics available in children's input? If knowledge of syntactic cues is ultimately cached out in terms of distributional knowledge, this is a plausible hypothesis (see Wojcik and Saffran, 2015 for discussion).

There is another difference worth mentioning. In order to identify syntactic class membership, a thorough distributional analysis is not strictly necessary. Instead, it might suffice to track the relative position of a novel word relative to a high-frequency function word, and pay attention to the overall word-order. The availability of these simple heuristics for identifying syntactic class membership would be especially useful to young children with limited computational resources. Further, making use of such heuristics would allow young children to break into a language faster, and start to make progress earlier compared to patiently collecting a large inventory of distributional statistics. On this view, the usage of syntactic and distributional cues may not share cognitive resources. Investigating the interface between identifying and encoding syntactic vs. finer-grained distributional cues is worth further research. Many questions related to this topic are examined herein.

An Example

It is worth explaining the computational procedure that underlies DECAF in detail. For this purpose, consider Figure 1.2. The two sentences on top are sentences a hypothetical learner has previously been exposed to. Further, our hypothetical learner knows the meaning of each word in these two sentences, which means that he or she has available semantic features for those words. Imagine that this learner encounters a new sentence ('*The lonely gorilla eats bamboo*'), and that he or she knows the meaning of each word in the sentence except for *gorilla*. Further, assume that the learner is not given perceptual information about the novel word. It should be obvious that this kind of situation is not just a thought-experiment or outlier scenario; any language-learning child will frequently find themselves in a situation of this kind. What can he or she do?

First, our learner might collect information about the lexical context in which *gorilla* is presented. For instance, *gorilla* is followed by the 3rd-person verb *eats*. This information alone may be sufficient to induce an approximate meaning. To do so, the learner first retrieves words that have occurred in similar contexts, such as the words *dog* and *bird*, which have both been observed alongside *eats*. Next, the learner might ask: What semantic features do these two distributionally similar words have in common? Because dogs and birds are both animals, the learner would extend semantic features related to ANIMAL membership to the novel word *gorilla*. Voilà, our hypothetical learner has induced a novel lexical semantic representation for a novel word in the absence of direct perceptual information or support from the referential context.

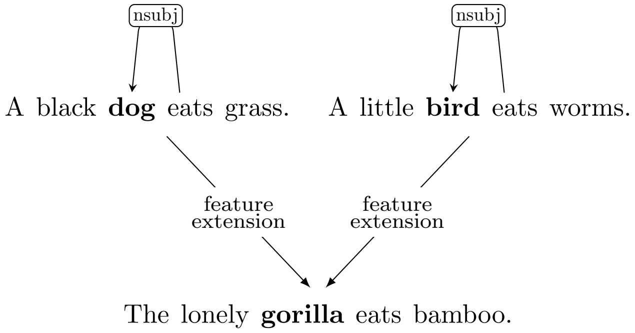


Figure 1.2: An illustration of the distributional extension of category-associated features (DECAF) from known words (*gorilla*, *bird*) to the novel word *gorilla*. By attending to the context in which the novel word occurs, and retrieving words that occur in distributionally a similar context, a child can jump-start his or her semantic interpretation of the novel word. The ability to extend category-associated features (features related to the ANIMAL category) would be especially helpful in the absence of referential information about the novel word *gorilla*. Further, DECAF does not require knowledge about the meanings of co-occurring words (e.g. *eat*); instead, a child need only have previously learned semantic (e.g. perceptual, conceptual) features for words that are distributionally similar to the target word (e.g. *gorilla*, *bird*).

Behavioral Evidence

Is there any behavioral evidence that would support the idea that children actually perform a procedure similar to the one I have proposed above? Most work in statistical and distributional learning has revolved around the discovery of syntactic classes, rather than finer-grained distributional semantic distinctions. However, there is some evidence that children's distributional learning capabilities are at least sufficiently sophisticated to perform DECAF in principle. I briefly review this evidence below, and end this section by discussing a study which provides the most direct support that children use lexical distributional statistics to guide their word-meaning mappings.

First, it is now well established that children track language-internal statistics related to word-order. This is a pre-requisite to being able to perform DECAF, given that distributional similarity needed to extend known semantic category-associated features works best when constrained by word-order. One of the earliest work on infant distributional learning that examined word-order acquisition is a series of studies conducted by Gomez and Gerken (1999). Twelve-month-olds were first exposed to sequences generated by an artificial language, and then tested on several learning and generalization tasks designed to probe their learned knowledge. In particular, the authors were interested in whether the twelve-month-olds could learn transitional probabilities within multi-word sequences. Behavioral testing revealed that children successfully learned the probabilistic regularities in the ordering of words, as evidenced by their differential responding to grammatical and ungrammatical sequences containing familiar words. In a follow-up experiment, infants were also able to differentiate grammatical and ungrammatical sequences when their words were drawn from a novel vocabulary. Gomez and Gerken (1999) concluded that 12-month-olds can and do track sequential statistics, and that doing so yields not just knowledge of individual sequences but also knowledge of a more abstract level of structure.

Lexical co-occurrence associations that are potentially useful for guiding word learning are not limited to neighboring words, but may span across longer distances. To illustrate, consider that the subject noun *dog* and the verb *eats* may be separated by an intervening adjunct, such as the relative clause in '*A black dog that I see on my way home from work eats grass*'. This means that, in order to perform DECAF, children would benefit by being able to track non-adjacent lexical dependencies. Evidence that children can do this comes from work by (Gomez, 2002) on non-adjacent dependency learning in 18-month-olds. Their results showed that children acquired non-adjacent dependencies as long as the intervening slot was highly variable. The experiment was conducted as follows: First, children were exposed to the same non-adjacent dependencies (with transitional probability = 1.0) in one of three conditions. Each condition differed in the number of items that could occur in the intervening slot: In the low variability condition, the set size of intervening items was limited to three; in the high variability condition, the set size was capped at 12; in the high variability condition, as many as 24 items occurred in the intervening slot. Critically, only children in the high variability condition, in which the predictability of adjacent items was very low, learned the non-adjacent dependency, as evidenced by their ability to discriminate grammatical from ungrammatical strings. Additional studies have pinned down the age at which this ability comes online. For instance, Gómez and Maye (2005) showed that 15-month-olds but not 12-month-olds can track non-adjacent dependencies. The authors suggested this developmental pattern is evidence that the tracking of non-adjacent dependencies is supported by increases in memory capacity over the second year.

While there is strong evidence that by their second year, children have the capacity to learn about the statistical correlates of word-order and can track non-adjacent dependencies, do children use this knowledge to infer paradigmatic relations among the words they have heard? There are two lines of research related to this question. First, indirect evidence that children prioritize paradigmatic similarity when inducing novel word meanings comes from observations of the kinds of errors children make when extending known words to novel objects. For example, it has long been known that children rarely confuse an object name with a topically related object label (e.g. calling a gorilla *jungle*, or a soccer ball *goal*). A study of this sort was conducted by Macnamara (1982) who observed that serious naming mistakes as mentioned above are extremely infrequent, and that children are much more likely to confuse labels for objects that are taxonomically related (e.g. calling a dog *cat*). In addition, the authors found that children rarely confuse proper names with common nouns, object names with substance names, or adjectives with verbs. This indicates that children's inductive processes are biased such that labels for ontologically similar concepts are weighted more strongly during word extension. It is possible that this tendency reflects, in part, (i) the tendency of children's learned distributionally constructed lexical categories to be organized by paradigmatic similarity, and that is organization, is in part shaped, by (ii) pressure to construct distributional categories that are useful for performing DECAF.

To date, the most direct evidence that children perform a procedure similar to DECAF comes from work by Lany and Saffran (2010) who showed that 22-month-olds use prior knowledge about language-internal distributional structure to support their word learning. All children were first exposed to sequences of an artificial language with two word categories, X and Y. In the experimental condition, children heard sequences of the form a-X and b-Y, while in the control condition, children also heard an equal number of sequences of the form b-X and a-Y. This means that children in the experimental condition were provided reliable distributional markers that distinguish the two word categories, X and Y.[2] In contrast, in the control condition, these cues were counter-balanced so that they cancel each other out, statistically speaking. Next, children in both groups were trained on pairings between artificial language sequences and pictures of unfamiliar vehicles and animals. More specifically, children in both conditions were shown pictures of animals paired

[2] In this study, category membership was marked by both distributional and phonological cues. In a follow-up study, Lany and Saffran (2011) found that children weight distributional and phonological cues depending on their developmental stage (children with larger vocabularies rely more on distributional cues).

with familiar a-X sequences and pictures of vehicles paired with familiar b-Y sequences. The purpose of this training session was to induce word-meaning mappings between words belonging to category X and animals, and words belonging to category Y and vehicles. The question is whether children that had learned distributional markers that reliably diagnosed category membership in X or Y (experimental condition), would be more successful in mapping the artificial sequences to their corresponding picture. Indeed, only the children in the experimental condition were able to learn the trained associations between artificial sequences and pictures — despite the fact that children in the control group had the same amount of experience.

So far, this is evidence that children are sensitive to the distributional statistics of the language they hear, and that they use this knowledge to support their word-referent mappings. However, DECAF is most useful when encountering a word that a child has never heard before and in the absence of supportive referential context. Is there evidence that prior distributional knowledge can help induction in this situation? The answer is yes. In a follow-up experiment, Lany and Saffran (2010) examined the degree to which children were able to generalize the information they acquired during training to novel sequence-picture pairings. In the generalization portion of their experiment, children heard novel artificial sequences of the form a-Z or b-Z where Z is a set of novel non-sense words that children were never trained on. If 22-month-olds are able to perform distributional extension of category-associated features (DECAF), then the children in the experimental condition should be able to leverage their distributional knowledge to infer that a novel word should be mapped onto a picture of an animal when it occurs in the context a (a-Z) or vehicle if it occurs in the context b (b-Z). Indeed, only children in the experimental condition were able to do this. The authors concluded that experience with statistical cues that mark word categories lays an important foundation for learning the meanings of those words. More broadly, these results demonstrate that children leverage their experience with language-internal distributional statistics that pick out lexical semantic categories to induce novel word meanings. In sum, DECAF appears to be a procedure that children use. However, how the statistical information that children gather over the course of language experience is used and transformed into a format that can support induction is much less understood, is the topic of this thesis.

A more recent study by Wojcik and Saffran (2015) examined how grammatical knowledge influences children's ability to exploit different statistical cues for inferring distributional semantic properties of novel nouns. To answer this question, the authors first exposed 2-year-olds to familiar sentences that featured novel non-sense nouns. Crucially, novel nouns had been experimentally grouped into form-based classes; category-membership was either marked by (i) occurrence in identical sentences, or by occurrence in identical syntactic positions across different sentences. Using a variant of the headturn preference paradigm, the authors then tested whether children have encoded the form-based semantic similarity relations among the novel nouns. While all children were able to infer the semantic similarity structure when category-membership was cued by occurrence in the same sentences, only children with more advanced grammatical knowledge learned the positional similarities of novel words across sentences. Overall, this work demonstrates that the ability to perform form-based semantic inferences about novel words — a precursor to DECAF — draws on and is supported by grammatical knowledge. In a subsequent section, I discuss how grammatical knowledge is implicated in DECAF.

Related Work

More broadly, the idea that categorization may aid in knowledge generalization from similar to novel words in the same category is not new. For instance, Borovsky and Elman (2006) observed that variation in language input related to semantic category cues can facilitate lexical acquisition in a computational model.

In particular, the authors found that language-internal statistics that are related to the coherence of lexical categories (e.g. frequency distribution of words that belong to the same category) can influence word learning, independently of the richness of the vocabulary or total amount of language exposure. Taking this idea to the next level, a randomized trial of a pre-school intervention that emphasizes children's understanding of taxonomic associations among words was conducted. In spirit to the ideas discussed here and by Borovsky and Elman (2006), the authors found that the intervention improved children's ability to use semantic categories to identify the meaning of novel words (Neuman et al., 2011). These gains in word learning persisted at least until six months post-intervention for children in the treatment group. The authors concluded that a program targeted to learning words within taxonomic categories may "act as a bootstrap for self-learning and inference generation" (Neuman et al., 2011).

Desiderata

What are the computational pre-requisites or abilities that must be present for a child to perform DECAF? Apart from plenty experience with language (i.e. accumulation of distributional knowledge), and a sizeable inventory of learned word-meaning mappings, there are important constraints on the kind of distributional similarity that would be most useful for extending category-associated features. Consider, for example, Figure 1.2. The semantic features that should be extended are those features that *dog* and *bird* have in common. A potential difficulty is distinguishing these two words from other words that also occur in the same sentences. For instance, the words *grass* and *worms* also co-occur with *eats* — in fact, they are also neighbors of *eats* just like *dog* and *bird* are. What, then, prevents a child from extending features that are common to these words? The answer is that in order for DECAF to work properly, the child must compute a particular kind of distributional similarity, namely 'paradigmatic similarity' (De Saussure, 1989; Rapp, 2002; Ruge, 1992). This kind of similarity takes into consideration word-order so that words that are topically related (occur in the same event and/or sentence) are judged less similar relative to words that are taxonomically related (words that are members of the same superordinate category, such as ANIMAL or NOUN). To be useful in DECAF, a measure of distributional similarity must prioritize paradigmatic similarity over topical similarity. If this were not the case, then a child might — incorrectly — infer that semantic features common to grass and worms (features related to FOOD or NATURAL-KIND) should be mapped onto the lexical representation of *gorilla*.

There are a number of ways to tune a distributional semantic model to pay particular attention to paradigmatic similarity (M. N. Jones et al., 2015; Sahlgren, 2006; J. A. Willits et al., 2007). One way is to track co-occurrence associations within a restricted window surrounding the target word. By tracking nearby lexical items only, a distributional semantic model is more likely to encode paradigmatic similarity. If the window is small enough, a model primarily learns similarity relations within part-of-speech (POS) classes, given that POS classes are defined in terms of the distribution of adjacent words (i.e. neighbors). In contrast, by expanding the window size, distributional semantic models tend to become more tuned to topical similarity: Which words tend to co-occur in the same document? If documents describe situations or events that are topically similar — which is usually the case — then the observation that two words are more likely to co-occur in a given document (compared to chance) indicates that they are topically related (Landauer & Dumais, 1997; Sahlgren, 2006).

That said, there is another way to promote learning of lexical semantic representations such that paradigmatic instead of topical similarity relations are prioritized: Next-word prediction. This approach is the one that is adopted in this thesis. By learning to predict upcoming words given some prior context (a window of, say, seven consecutive words), a distributional model can learn which words are substitutable given prior context. Models that learn via next-word prediction are often called 'language models' in the technical literature (Bengio et al., 2003; De Mulder et al., 2015; Futrell et al., 2019; Schwenk & Gauvain, 2005), and they are said to learn by minimizing the 'language modeling objective'. While it is well known that language models quickly acquire knowledge about part-of-speech membership, and finer-grained substitutability relations, it is not yet known whether these relations are also suitable as the basis for the distributional extension of semantic category-associated features (DECAF), and whether such models are sufficiently robust to handle transcribed speech to children. There are several potential obstacles that might impede the emergence of useful lexical semantic representations in language models, and all of them are related to the tendency of next-word prediction to produce lexical representation that are too specific (i.e. not sufficiently categorical). The resultant representations may, therefore, not be useful in situations or tasks that are not closely aligned with the next-word prediction task that was used to construct lexical semantic representations. Because the distributionally-mediated extension of category-associated features (DECAF) cannot be straightforwardly operationalized as a next-word prediction task, it remains to be seen whether representations and relations learned in the service of next-word prediction can be used to perform DECAF.

A second, crucial desideratum is that a distributional model be able to encode paradigmatic similarity relations at the lexical level as opposed to just the chunk-level. For instance, a model should be able to assign a high similarity to the lexical pair *dog-cat* without needing to observe each word in identically distributed contexts. It is much more likely that each word occurs in distinct contexts (e.g. '*this hungry dog*', '*a fuzzy cat*') more often than in shared contexts. Due to the idiosyncratic nature of language use, and the strong reliance on integrating sequential dependencies at the chunk-level, a language model is likely to struggle with this. Because language models are not explicitly optimized at the lexical-level, there are no explicit constraints in place that would guarantee that a language model learns that two semantically similar lexical items should be assigned similar lexical representations. Under what conditions this assumption is violated, and its negative consequences for supplying lexical semantic representations to DECAF, is also a major topic that is investigated in this thesis. The discussion of desiderata is picked up in Chapter 4.

When is DECAF Useful?

Inferences based on language-internal distributional statistics would be particularly useful when inferring aspects of word meanings in the absence of perceptual information about the labeled object. That said, one may wonder: What is the point of constructing approximate meanings in these perceptually uninformative situations, when the child could instead forego the construction of approximate meanings and wait until a more felicitous situation arises in which referential information about the novel word is actually available? Because DECAF can only provide approximate meanings, why not simply delay learning until a more precise meaning can be pinned down? From a theoretical perspective, DECAF is preferred because the coarse-grained lexical representations it produces may prepare or give an edge to children when they next encounter the same novel word in the presence of perceptual information. That is, armed with an approximate lexical semantic representation, a child would be more likely to identify the referent of the novel word compared to a child who begins the mapping problem without additional supportive information. That said, this does not imply that DECAF is useful only in situations in which perceptual information is absent, nor that DECAF can be applied only when other word learning strategies do not. Rather, tracking distributional cues, and using them to constrain one's inferences is can provide important supplementary information even when other cues are more readily available and/or alternative strategies are more applicable. For example, a labeled object may be present in the referential context but the child may have difficulty isolating that object among other novel objects. DECAF may be useful in narrowing the scope so that the child's word-to-referent mapping strategies can get a head start.

It is very likely that the utility of DECAF only increases with age during the first several years of life. The reason is that the external semantic reference of discourse reduces with age, as caregivers increasingly discuss concepts outside of the 'here and now' and the topics of conversation become increasingly broader (and more abstract). All this means that children must increasingly shift their reliance on the referential context to linguistic and/or conceptual knowledge to keep up the pace. In particular, the perceptual correlates of abstract words tend to be subtle and difficult to isolate; it is not implausible that a child would make heavy use of distributional cues to lay the foundation for learning about what these words mean. DECAF therefore presents a potential strategy for bootstrapping the learning of abstract words.

Furthermore, I suspect that DECAF is most helpful when the novel word is a noun, and least helpful when it is a verb. There are at least two reasons for this: First, the meanings of verbs are much more relational than nouns. As a consequence, to know the meaning of a verb, a learner should know what verbal arguments are, how they function in the syntax of their native language, and which words tend to fill which argument slot. This separation between the different functions of different slots and which slots tend to be filled by which words is not easily accounted for by statistical knowledge about the contexts in which verbs tend to occur. Second, as mentioned previously, the number of verb arguments plays a crucial role in determining aspects of the meaning of a verb. However, an exclusively distributional analysis of verb contexts does not straightforwardly lend itself to keeping track of argument number — this would require counting co-occurrences among verb arguments in addition to verb-argument co-occurrences. Keeping track of these higher-order statistics and how they relate to syntactic structure is possible (Erk & Padó, 2008; Mitchell & Lapata, 2010), but goes well beyond distributional modeling as traditionally conceived (Bullinaria & Levy, 2007; Landauer & Dumais, 1997; Lund & Burgess, 1996).

Another area in which DECAF might be useful is when the child has gained a basic proficiency in reading, but does not yet know many of the words he or she encounters during reading. Due to the nature of written material, a budding reader cannot rely on perceptual information or the referential context to support inferences about novel word meanings. Upon encountering a novel word on the page, a reader may be especially reliant on syntactic and distributional cues, in addition to world knowledge. Further, reading is not just a situation where DECAF could be performed, but also a source of knowledge that could enrich inferences based on DECAF in the future. Interestingly, this raises the possibility that contexts for novel words encountered during reading may induce approximate meanings, and these meanings may influence how a listener interprets novel usages of the novel word. Indeed, careful manipulation of lexical contexts during reading can influence how a reader interprets novel (and known) words in a subsequent language task (Dye et al., 2017)[3].

It should be kept in mind that, at the point in a child's life when they have learned to read, it is safe to assume that he or she can leverage sophisticated knowledge about how the world works to make plausible guesses about a novel word's meaning. For example, a child who reads or hears '*I woke up yesterday, turned off my alarm clock, took a shower, and cooked myself two grimps for breakfast*' (Granger, 1977) would probably know a lot about the meaning of the word *grimp*, provided he or she already has a basic command

[3] This experimental paradigm may be used to experimentally validate many of the claims made herein about DECAF and children's statistical learning. A downside is that subjects must have basic reading proficiency which limits the age range of children that can be investigated

of the English language as well as an understanding of the meanings of the other words in the sentence. In the absence of perceptual information, or relevant information provided by the referential context, a school-age child could infer that a *grimp* is a common noun that labels a type of breakfast food. Interestingly, implemented systems that can perform such inferences have a long history in computational linguistics (Berwick, 2014; Granger, 1977; P. S. Jacobs & Zernik, 1988). This raises an important question, namely how much knowledge about how the world works and knowledge about linguistic contexts contribute to inferences during written and spoken language comprehension, and how they interact. It is plausible, that as the child grows older, he or she is less likely to rely on distributional cues, and instead consults her growing experiential knowledge (e.g. the likelihood and plausibility of certain events happening in the world, knowledge of cause-effect relationships, etc.). To better understand DECAF, it will be useful to know how much distributional information continues to be of use to the child as they grow older. It is not implausible that distributional information is altogether discarded in late-stage semantic development, at which point it would be impossible for researchers to be able to study or find behavioral evidence for DECAF. I leave these questions for future work.

1.2 Research Questions

Nobody denies or doubts that the language that children hear is essential for learning about semantic categories. As E. V. Clark (2017b) remarks:

> ...child-directed speech plays a critical role in the construction of [semantic] categories during acquisition. And what adults say to children in conversation offers them a major source of information of information about the semantic categories of the language they are learning.

Linguistic conventions segment the natural world into useful semantic categories, and via exposure to language use, the child is able to tune into these conventions in order to add to, extend, and refine those semantic categories that he or she already knows from birth (e.g. animacy). What is less well understood is whether the *statistics* of language use — as opposed to how words are mapped onto concepts in a language — can be used to supplement children's lexical semantic development, and whether, children actually use statistical cues to do so. Investigations of this sort are relatively new, and have benefited by recent efforts to compile and make available large child-directed language corpora, and the availability of computing resources that can exploit the statistics available in such corpora (Asr et al., 2016; Riordan, 2007). In this thesis, I continue this line of work by training a neural language model to predict next-words in a large corpus of language directed to children and examining what the model has learned about the semantic similarity structure of common nouns (*boat* and *bicycle* are both members of VEHICLE; *dog* and *cat* are both members of MAMMAL). If the model is able to capture the taxonomic similarity among related nouns, this would suggest that (i) distributional cues about lexical semantic category membership are present in child-directed input, and, more importantly, (ii) lexical semantic representations useful for performing DECAF emerge in the service of predicting upcoming words in child-directed input.

There are a number of challenges ahead. First, the choice of computational model is an important factor when modeling psychological processes implemented in biological systems. The chosen model must not only be (i) sufficiently complex to be able to discover arbitrary statistical cues to semantic category membership, but also (ii) sufficiently simple to reduce doubts about cognitive, neural, and developmental plausibility, and to be interpretable by non-experts. In addition, the chosen model should preferably have been well established and accepted by the psycho-linguistics and cognitive-science community. If not, the link between DECAF and the kinds of computations that language-learning children are thought to perform would be weakened. One of the overarching goals of this thesis is to examine whether DECAF is a procedure that is compatible with the kinds of behaviors and computations that children are already known to perform, and to determine whether another, more specialized system is needed instead. In effect, this thesis is a feasibility study: Can we take an existing model that is well-understood in psycho-linguistics and use it to model a novel task — the construction of form-based lexical semantic representations useful for performing DECAF? If so, this would lend additional support to the psychological reality of both the computational model, and the distributionally-mediated extension of category-associated features (DECAF). My approach follows the premise of a recent line of investigation that considers model building in the psychological sciences as a need to satisfy constraints that arise at multiple levels of analyses, and in multiple disciplines that may have been historically separated (Martin, 2020; Van Rooij, 2008).

1.2.1 A Fundamental Trade-off

The type of model that is investigated in this thesis is the neural language model — a neural network based system trained to predict next-words in a large corpus of language data (Bengio et al., 2003; J. L. Elman, 1990; Mikolov, 2012). An in-depth discussion why this choice was made, details concerning the architecture of the model are given in an upcoming section. For now, it suffices to say that neural language models are not only (i) highly capable in their ability to acquire arbitrary sequential dependencies between words, but also (ii) keep track of word-order explicitly, and (iii) prioritize grouping of lexical semantic representations by paradigmatic similarity. As such, neural language models appear to satisfy the minimal requirements for producing lexical semantic representations in a format useful for performing DECAF. They are also widely known and researched in the cognitive science and psycho-linguistics community (T. A. Chang & Bergen, 2022; Cleeremans et al., 1989; Ettinger, 2020; Futrell et al., 2019; Gulordava et al., 2018; Servan-Schreiber et al., 1991; van Schijndel et al., 2019).

A second reason why I have opted for existing neural language models is the unique opportunity to examine whether the distributionally constructed lexical semantic representations needed to perform DECAF can emerge in a model that can already account for other phenomena in children's language acquisition, such as the discovery of part-of-speech classes. By being able to accomplish multiple learning tasks, this lends additional credibility to the model, and by extension the proposal that children perform DECAF. If children already have the computational capacity to perform DECAF, why wouldn't they? Further, such a unified model would also be able to account for aspects of syntactic bootstrapping (e.g. the discovery of syntactic cues) which is extremely helpful for early word learning. A primary question is whether the same system that can account for aspects of established and well accepted proposals for how children break into language can also be used (with no or minimal modification) to learn finer-grained distributional statistics that go beyond syntactic category membership, and which can be used as the basis for DECAF. This possibility should be examined first before a more specialized model — separate from previously established accounts of computational underpinnings of language acquisition — is developed and tested.

However, there are caveats, which are potentially of theoretical interest. It is important to note that neural language models were developed to learn sequence-level (as opposed to lexical-level) representations and are almost always used to perform sequence-level language tasks (e.g. sequence categorization, grammaticality judgment, ambiguity resolution, etc.). At first blush, this would make models of this type less favorable for modeling the acquisition of lexical semantic representations needed to perform DECAF. However, my aim is not to develop a special-purpose model and optimize it on a single task — the relevance of the obtained results to psycho-linguistic research would be questionable, and would require much more insight into the inner workings of the model to become widely accepted by scholars interested in children's semantic development. Instead, I have opted to use an existing and well-established model that is known to account for a diverse array of language tasks (e.g. contextual processing, discovery of POS classes, etc.). While this comes with limitations on the degrees of freedom available to the modeler, I take this as a good thing. In fact, the tug-of-war between learning sequential statistics and word-level statistics that this setup invariably forces upon the modeler raises interesting questions that are relevant to the psycho-linguistic community (e.g. how do constraints at the sequence-level inform relations between individual words?).

That said, the unified account proposed above presents a unique challenge. I will argue that the goal of learning about the sequential structure of language is only partially compatible with the goal of acquiring lexical semantic representations, and by extension, paradigmatic similarity judgments, that are useful for performing downstream tasks such as DECAF. In fact, at the limit, it appears that these two objectives are fundamentally at odds with each other. Determining whether this is actually the case — and taking into consideration the unique language environment of children — is the purpose of this thesis. The answer to this question is important for language acquisition research as it can inform decisions about the architecture of the human language system: For instance, how modular are the sub-systems of the human language system? What are the computations that each sub-system performs? Which learning problems can be accounted for by a unified system, and which require qualitatively different computations?

Upcoming sections and chapters provide much greater detail and support for this argument. To preview, I briefly outline the steps of the argument. As mentioned previously, a pre-requisite for performing DECAF is the availability of lexical semantic representations that capture paradigmatic similarity among individual words. This means that a distributional learning system must in one way or another track word-order statistics to reduce the impact of topical similarity on learned representations. Neural language models are well suited to this task, given they track word-order explicitly and better than other distributional systems. Second, learned semantic representations must capture distributional semantic information at the word-level as opposed to the chunk-level. However, and this is the crux of the argument, the same mechanism that enables sensitivity to word-order in neural language models (i.e. next-word prediction), also promotes the formation of knowledge at the chunk-level and at the expense of the lexical-level. If this argument is solid, then there would be restrictions on the conditions under which lexical semantic representations useful for performing DECAF can emerge in neural language models. A primary aim of this thesis is to determine precisely what these conditions are.

Many of the questions posed above are relevant to ongoing work in psycho-linguistics and cognitive science more broadly. For example, the question about the size of linguistic units stored in semantic memory, and to what degree storage is influenced by word-order, is related to previous work that has examined what computational processes and representational systems children need to succeed in language acquisition (Ervin-Tripp, 1973), and the degree of functional specialization the brain regions responsible for keeping track of word-order and integrating semantic information (Brouwer et al., 2017; Federmeier, 2007; Kuperberg et al., 2007). In addition, if it turns out that the distributionally mediated extension of category-associated features (DECAF) is something that children can perform, then this would provide fresh insights about the interface between language-internal distributional similarity and other language systems that construct meaning from extra-linguistic experience.

1.2.2 Hypotheses

In sum, the most important questions considered in this work are the following: Can neural language models be used as a source of distributional knowledge useful for extending known semantic features to novel words? In particular, can such models capture distributional regularities that can be used to diagnose the semantic category membership of common nouns in child-directed input? Further, do the learned representations capture semantic knowledge at the lexical-level or the chunk-level, and how is this trade-off influenced by the statistical properties of the input? Could a statistical learning engine based on next-word prediction be used to identify and encode distributional cues in a language-agnostic manner? What are the cross-linguistic considerations of learning lexical semantic representations in a system that is optimized for next-word prediction?

There are some reasons based on first-principles that would suggest one outcome is more likely than another, and it deserves mentioning here. Based on insights into learning dynamics of neural networks and error minimization (Saphra & Lopez, 2020; Servan-Schreiber et al., 1991; Shah et al., 2020), and how credit is assigned across time steps in recurrent systems (Bengio & Frasconi, 1993; Werbos, 1990), a reasonable working hypothesis is that much of the distributional semantic information that would be useful for performing DECAF will be encoded at the chunk-level by neural language models, and is therefore not made available in the form of standalone lexical representations. In other words, lexical semantic information useful for performing DECAF is likely to be 'trapped' in the internal contextualized state of the processor, where it is of little use to an external system. The reason is that contextualized knowledge may be difficult to extract without additional specialized machinery. Thus, it is possible that while knowledge relevant to performing DECAF has has been acquired by the network, it may not be available in a format that provides ready access to this knowledge to downstream systems. This hypothesis is bolstered by decades of work by psycho-linguists who have showed that semantic phenomena are tightly intertwined with context and syntactic properties of sentences (J. L. Elman, 2011; McRae et al., 2005; Tabor et al., 1997a; Trueswell et al., 1994). The presence of these interactions, the context-dependent nature of language, and the many spurious and idiosyncratic dependencies in natural language sentences make the prospect of encoding static and readily accessible lexical semantic representations improbable, but certainly not impossible. This wisdom has morphed into an unspoken assumption, namely that next-word prediction is too unconstrained to disentangle lexical-level semantic information from the their syntactic backbone and/or other (potentially spurious) lexical interactions among sentence components. This issue is related to the claim that people perform structured semantic decomposition of the utterances they hear, in addition to (or potentially at the exclusion of) prediction-based processing (Abend et al., 2017; Gershman & Tenenbaum, 2015). This possibility is discussed in more detail in an upcoming section. While this topic is central to many lines of research in computational linguistics and formal semantics (R. Jackendoff & Jackendoff, 2002; Pustejovsky, 1998), I am not aware of any work in the psycho-linguistics literature that has tested this assumption using a combination of computational modeling and a realistic sample of of child-directed input.

In the next two sections, I provide a brief overview of the field of distributional semantic modeling, evidence for distributional learning in children, criticisms of a purely distributional approach to lexical semantic development, and a micro-survey of corpus-based distributional semantic models used in psychological research. The purpose is to provide the reader a basic familiarity with computational models that have been proposed in the past, and to set the stage for discussing why I have chosen to depart from the standard formulation for modeling aspects of DECAF.

1.3 The Distributional Hypothesis

Recent research efforts to better understand children's lexical semantic development have focused on the distributional hypothesis (Firth, 1957; Harris, 1954), the claim that the similarity, class membership, or relations between linguistic units or concepts can be inferred from language-internal statistical contexts in which those units occur. In the computational realm, this idea was formalized in a range of different models of adult semantics, such as LSA (Landauer & Dumais, 1997), HAL (Lund & Burgess, 1996), BEAGLE (M. N. Jones & Mewhort, 2007), and Probabilistic Topic Modeling (Steyvers & Griffiths, 2007). These models use distributional information to construct semantic feature vectors for words. Feature vectors can be composed of co-occurrence associations between words, or they can consist of abstract or latent features that are formed over the course of learning. The semantic similarity of two words can then be calculated by measuring the similarity of the feature vectors of two words (Kahneman & Tversky, 1972; E. E. Smith et al., 1974). Considerable research has since shown that these and related procedures for representing semantic similarity predict a wide range of adult psycho-linguistic variables, such as semantic priming and explicit similarity judgments (Bullinaria & Levy, 2007; M. N. Jones et al., 2006; Lund & Burgess, 1996; McDonald & Lowe, 1998; Olney et al., 2012; Pereira et al., 2016; Plaut & Booth, 2000). Other work has shown that distributional semantic models trained on large corpora can predict variation in multivariate fMRI activity patterns of people engaged in language tasks (Kaiser et al., 2022; Pereira et al., 2018).

Although one goal of research in distributional semantic modeling is to understand how people acquire and represent lexical semantics directly (Lenci, 2018), another important goal is to characterize how forces that govern the accumulation of distributional information (e.g. quantity and quality of language input) influence the operation of downstream tasks that leverage this information. The distributionally-mediated extension of category-associated features (DECAF) is one of many such tasks; others include word recognition during reading (Amenta et al., 2020; Patterson et al., 1996; Staub et al., 2010), verb argument processing (Lenci, 2011; McRae et al., 2005), paste-tense formation (Ramscar, 2001), production (MacDonald, 2013; Thothathiri & Braiuca, 2021), word segmentation (Goldwater et al., 2009; McCauley & Christiansen, 2019), and sentence parsing (D. L. T. Rohde, 2002; Trueswell et al., 1994).

1.3.1 Computational Work

Numerous corpus and computational studies have provided evidence for the usefulness of distributional information for an initial categorization of words in the absence of referential information (Asr et al., 2016; Cartwright & Brent, 1997; Chrupała, 2012; Feijoo et al., 2015; Fremann & Lapata, 2016; M. N. Jones & Mewhort, 2007; T. H. Mintz et al., 2002; Redington et al., 1998).[4] Much computational and corpus work related to the distributional hypothesis has focused on the discovery of part-of-speech categories in large corpora of child-directed input, and especially the ability to distinguish nouns from verbs based on the lexical contexts in which they occur (Brusini et al., 2021; Freudenthal et al., 2013; J. A. Willits et al., 2014; Zhu & Clark, 2022). There is considerably less work on the role that language-internal distributional information might play in children's acquisition of semantic category distinctions, such as knowing that the word *dog* refers to an object in the category MAMMAL, and not BIRD. That said, the few studies that have examined this question, have found that distributional cues useful for clustering common nouns into semantic categories

[4] In agreement with the idea that distributional learning is a domain-general cognitive capacity, recent work has shown that distributional cues are also informative in the visual domain, in which co-occurrence between objects, rather than words, can provide useful information about visual concepts (Sadeghi et al., 2015a).

are widely available in the language statistics available to English-learning children (Asr et al., 2016; Riordan, 2007; Unger & Fisher, 2021; Unger et al., 2020).

1.3.2 Behavioral Work

While computational and corpus studies cannot demonstrate that children actually use distributional statistics, numerous behavioral experiments have demonstrated that human learners are sensitive to distributional information from an early age (Gomez and Gerken, 1999; T. H. Mintz, 2003; Pelucchi et al., 2009; for a review, see Lany and Saffran, 2013). In the work I review below, I specifically consider distributional as opposed to related kinds of statistical learning, such as learning transition probabilities between syllable or word-boundaries[5] which is useful for word recognition and speech segmentation (see, for example, Cartwright and Brent, 1997; Hockema, 2006; Saffran, Aslin, et al., 1996).

One of the most comprehensive investigations of distributional learning in the language domain was conducted by Reeder et al. (2013). Specifically, the authors investigated which specific distributional variables influence the ability of university-students to discover grammatical form-classes in a carefully controlled sample of artificial language. The authors observed that people are not only sensitive to the contexts of individual words, but also track the degree to which contexts overlap across words. It turns out that the latter is a crucial determinant of how much people extend class membership to novel words takes place vs. allow for lexically specific exceptions. Interestingly, context overlap, and number of shared contexts interact and together determine how readily participants generalize learned distributional properties. For instance, Reeder et al. (2013) observed that when a new word shares just one context with the category members of a learned category, participants readily extended category membership as long as the context was shared by many other members. As context overlap is lessened, learners become more conservative with regards to extending class membership to a novel word. Importantly, the authors found that participants' judgments were sensitive to the degree of consistency in context overlap, such that statistically less likely gaps in overlap are treated differently than those that are expected due to chance. Based on this observation, Reeder et al. (2013) concluded that people use distributional cues in a "principled way to determine when to generalize and when to preserve lexical specificity".

Is distributional evidence similarly useful for an initial clustering of words into semantic categories, as it is for grammatical categories? There is preliminary behavioral evidence to support such a view.

For instance, it has been demonstrated that the context of a noun in the same sentence provides cues about how likely it is that the noun refers to an animate object. In a recent study, 9- and 24-month-olds exposed to sentences like *The vep is crying* inferred that *vep* referred to an animate entity (i.e., a novel animal), while participants who heard sentences in an animacy-neutral condition like *The vep is right here* showed no such preference at test (B. Ferguson et al., 2014, 2018). To exhibit this behavior, participants must have learned that *vep* appeared in the subject position of a familiar verb that requires an animate agent. Acquisition of the selectional restrictions on arguments of the verb *cry*, in turn, can be explained in terms of the acquisition of how *cry* patterns with other words in the same sentence.

Other work has shown that, by their second birthday, children have begun to appreciate the selectional restrictions of at least some verbs (Friedrich & Friederici, 2005; L. R. Naigles et al., 2009; Valian et al., 2006).

[5] This type of statistical learning differs from distributional learning, as it requires children to track frequencies rather than transitional probabilities. That said, these two types of learning are intimately related, given that sequential statistics can be used as the basis for distributional (i.e. co-occurrence) statistics.

Selectional restrictions are boundaries on the kinds of arguments a verb takes. Because these boundaries tend to circumscribe semantic and conceptual domains, the observation that children possess this knowledge at an early age is indirect evidence of a basic understand of which words are semantically related to which other words. Similarly, adults also use their knowledge about how familiar verbs pattern with other words to predict the semantic properties of verb arguments (G. T. M. Altmann & Kamide, 1999; Fernald et al., 2008; Valian et al., 2006).

There is also a sizeable literature on how distributional factors influence or change how people interpret familiar and novel words. For instance, McDonald and Ramscar (2001) found that participants' similarity judgments of novel and nonce words are affected by the linguistic contexts in which they were read. In their study, young adults read short text passages in which contextual cues to word meaning were manipulated. In accordance with these manipulations, the authors found that reader's interpretations of target words were 'pushed' towards the meanings of context words. A similar study was conducted by Dye et al. (2017) who examined distributional statistics related to word-order influence people's interpretation of the meaning of familiar and novel words during reading. The critical manipulation was whether a word that is semantically similar to a target word is presented in the left or right context in text read by participants. Interestingly, participants rated the similarity between target words and words occurring in the left context higher compared to words occurring in the right context. This result confirms that the distributional semantic representations learned by people are influenced by the order in which word pairs were presented in sentences. Together, these results suggest that people are sensitive to lexical distributional statistics, including word-order statistics, and that these factors influence how meaning is constructed during reading.

More evidence for the potential role that distributional learning plays in lexical semantic development comes from a large literature that has examined what factors influence which words children learn first. Among these factors, measure of linguistic context consistently show up, and in many cases explain more variance than non-contextual factors such as word frequency, orthography, and semantic difficulty such as concreteness (B. T. Johns et al., 2016). In addition to measures of learning, contextual factors also account for large amounts of variance reaction time variability in in lexical decision tasks (M. N. Jones et al., 2012). In related work, Hills et al. (2010) found that a word's contextual diversity significantly influences how that word is learned and remembered. Words that are presented in a more diverse set of linguistic contexts are acquired more rapidly in early lexical acquisition.

In two separate lines of work, a correspondence was established between children's behavior and co-occurrence associations in child-directed language corpora. For instance, Unger et al. (2020) have linked distributional information with infants' semantic category formation in eye-tracking studies, in which patterns of looking-behavior are thought to reveal aspects of children's word knowledge. Second, Fourtassi (2020) found that individual difference between children's free word association responses could be accounted for by their distributional similarity in the linguistic environment. This statistical relationship was maintained even after other factors such as phonological similarity, word frequency, and word length were included in their analysis.

Taken together, the accumulated findings are converging on the idea that distributional regularities are readily available in language to children, and that children can and do rely on these regularities to accelerate their lexical semantic development.

1.3.3 The Symbol-Grounding Problem

As with all widely advertised theories in the psychological sciences, the distributional hypothesis has not gone unchallenged. Since the beginning, distributional models have been criticized for relying exclusively on language-internal information (Harnad, 1990; Perfetti, 1998). This means that the only window on meaning such models are provided with is the relations between the symbols themselves, but not their content, which includes perceptual information from the external world, and internal mental states (e.g. feelings, sensations, intentions, etc.) associated with a given symbol. Unlike people, distributional semantic models are not provided experiential information that would 'ground' the structural information they acquire in the real world. On this view, the criticism is not just that distributional semantic models (including neural language models) have much less information to work with than people, but that their information makes no contact with perceptual experience. This idea has been around since the development of the first neural language model: In J. L. Elman (1990), J. McClelland described language modeling as learning language "by listening to the radio". The same idea has been surfacing in computational linguistics and machine learning, where large neural network models trained on large corpora of text have made big headlines in recent years (Y. Liu et al., 2019; Radford et al., 2019). Despite their enormous successes, Bender and Koller (2020) argue that these models do not learn language in the way that children do, and therefore cannot be said to understand the meaning of the sentences in their input in the same way that a person would.

One response to these criticisms is to discard the idea of amodal symbols (e.g. lexical items, morphemes, atomic semantic units) as the central component underlying human semantic memory, and replace them with patterns of activation that directly relate to the sensory modalities. A particularly well known proposal of this sort is known as perceptual symbol systems theory (PSS; Barsalou et al., 1999, and is part of the larger grounded cognition movement in psychology (Barsalou, 2008). In PSS, the mental representation of a word is based on the perceptual states associated with experiences with the word's referent. Across many such experiences, the corresponding neural activations are thought to stabilize, thus giving rise to perceptually grounded representations. Some scholars have claimed that PSS is a competing approach to modeling and understanding mental representations in human semantic memory; others think that the two may be reconciled via a common interface (Riordan & Jones, 2011). Many scholars agree that the two theories are not mutually exclusive. For instance, PSS lacks a convincing account of words that do not label physically detectable objects and concepts, such as abstract nouns (Lupyan & Lewis, 2019). More importantly with regards to the aims of this thesis, PSS is silent concerning the fact that children are quite capable of constructing lexical semantic representations even in the absence of perceptual information needed to ground those representations. This is precisely the situation in which distributional semantic models excel.

In light of the discussion above, consider potential areas where distributional semantic models might fall short. Some research in this area has demonstrated that distributional models do not account for a variety of semantic phenomena in the realm of embodied cognition (Glenberg & Robertson, 2000). This should not come to anyone's surprise, given that distributional semantic models were not developed to account for human semantic memory in its entirety. Only few people would claim that language-internal distributional information is at the core of human concepts; rather, on many accounts, distributional information is considered supplementary to existing representations grounded in perceptual experiences. The distributional hypothesis does not argue that perceptual information is unimportant; what is needed is a way to integrate these two sources of information so that each can complement the other's weaknesses. Models that integrate linguistic and perceptual information in a unified distributional model have been proliferating in recent times (Andrews et al., 2009; M. Jones & Recchia, 2010; Sadeghi et al., 2015b). Finally, it should be noted that perceptual information is not totally inaccessible to models that derive lexical semantic representations exclusively from corpus data. For example, some distributional models are capable of inferring modality (B. Johns & Jones, 2011); others have shown that, when trained on child-directed speech, distributional models perform as well as sensorimotor-based feature representations in a semantic categorization task (Riordan & Jones, 2011). The latter finding is evidence that information diagnostic of lexical semantic category membership is redundantly encoded in perceptual and linguistic experience (Riordan & Jones, 2011).

The critique that word meanings must be grounded in experiential knowledge is related to a theoretical debate in concept acquisition, namely the abstractness of knowledge. Essentially, the question is whether knowledge consists primarily (or exclusively) of a rich set of associations between sensory-motor features, or instead also consists of abstract, amodal concepts that bind those features together. Waxman and Gelman (2009) succinctly describe this as a debate between two metaphors. The first is 'child as data analyst', whereby language acquisition occurs because of children's amazing statistical learning skills and their ability to build webs of associations of a wide variety of perceptual inputs and motor actions. This is contrasted with the 'child as theorist' metaphor, whereby children begin with and/or build up theories about the world involving rich conceptual knowledge structures, and these knowledge structures play a critical role in structuring language acquisition. Waxman and Gelman (2009) accept a role for statistical learning, but reject an exclusively 'child as data analyst' perspective, arguing that abstract concepts play a critical role in language acquisition and knowledge representation. Statistical learning algorithms, such as distributional semantic models, are often lumped into what Waxman and Gelman (2009) call 'child-as-data-scientist' explanations.

In sum, the question to what extent existing distributional semantic models represent viable psychological models of human semantic memory is an ongoing one. Despite the criticisms that have been launched at the distributional hypothesis, cognitive scientists continue to study, extend, and improve theories and models of distributional semantics (Bullinaria & Levy, 2012; Erk & Padó, 2008; M. N. Jones & Mewhort, 2007; Kabbach & Herbelot, 2020; Riordan & Jones, 2011; St Clair et al., 2010). A more recent examination of this issue concluded that distributional semantic models continue to be useful as "serious contenders as psychological theories of semantic representation..." (Günther et al., 2019).

Extending vs. Constructing Meaning

Concerns such as those raised by Barsalou et al. (1999) and Waxman and Gelman (2009) raise important questions about the interface between linguistic knowledge, the extra-linguistic conceptual domain, and perceptual experience. Historically, distributional semantic models have focused on the regularities within the former, and often at the expense of the latter. I argue, that to better understand the interface between language-internal and language-external knowledge structures, it can be useful to examine how language-internal knowledge is used in tasks that people perform naturally (i.e. outside the laboratory). While it can be useful to study the organization of distributionally constructed semantic knowledge formally and/or in isolation, questions about the nature of semantic representations should be informed by constraints on tasks and psychological processes that make use of that knowledge. Unsurprisingly, I argue that DECAF is one such task, and that better understanding the computational desiderata needed to perform DECAF can help constrain our thinking about how distributionally constructed semantic knowledge may interface with extra-linguistic knowledge in the brain.

Let me put this somewhat differently. Discovering regularities in the input is not likely something that people do without some end-goal. For instance, being able to predict upcoming words in a foreign language is not useful in and of itself, but can be enormously helpful if one's goal is to arrive at an understanding of the meaning of the words. While researchers have made enormous progress developing powerful new systems that can discover and acquire arbitrary statistical associations from natural data and without supervision, the next generation of models should take into consideration developmental and neurophysiological constraints, and, more importantly, task demands. How is the distributional information that is amassed by children over many years actually put to use beyond the passive integration into semantic memory? In this thesis, I ask what we might learn about children's semantic development if we consider language-internal distributional regularities not just as a direct source of information about word meanings, but as a tool that children might use to extend — rather than to construct — meaning. In addition to exploring what structure emerges via distributional analysis of language corpora, I also ask what representational structures are needed to account for the way in which distributional knowledge is deployed during word learning. This subtle re-framing allows one to gently avoid the symbol-grounding problem, and make progress characterizing the interface between passively acquired language-internal statistics, conceptual knowledge, perceptual experiences, and how these domains are actively combined in language learning and language use.

1.3.4 Many Models and Parameters

Both computational and experimental work has shown that substantial semantic information exists in the co-occurrence patterns of words, that human learners are sensitive to this information, and use this information to learn the meaning of words. Despite enormous progress in this field, scholars are still uncertain about how to best instantiate the distributional hypothesis in a computational model — the possibilities are virtually endless. A variety of model classes have been proposed, each of which differs in what statistics are captured, and how they are encoded. Current models vary widely in architecture and aims. While some focus on modeling qualitative aspects of the organization of lexical semantics, others emphasize a good fit with behavioral data such as laboratory tasks (e.g. semantic priming, word similarity judgments) and/or explaining aspects of real-world language use (e.g. reading, production). In this section, I provide a brief overview of psychological models that have been used to research the structure of human semantic memory. There are many others used in adjacent fields such as computational linguistics and machine learning which I do not review here (see e.g Mikolov, Chen, et al., 2013; Pennington et al., 2014). Further, the models I review below operate exclusively on corpus data, rather than data derived from adult norming studies. This is a deliberate decision, considering that the goal of this thesis is to model the contribution of children's experience-dependent language-internal distributional knowledge to their lexical semantic development.

At the core of most corpus-derived distributional semantic models is the co-occurrence matrix. In general, the rows correspond to target words, and the columns correspond to context elements. Upon observing that two words co-occur in the input, the corresponding element in the matrix is incremented. After a large number of such co-occurrences have been collected, the vectors in the rows can be considered the distributionally constructed meaning representations of words. Given that these representations are vectors, one can view them as 'distributed representations' and associated with a point in a high-dimensional vector space. In that space, semantically related words tend to be closer to each other than semantically unrelated words.

There are two broad classes of matrix-based co-occurrence counting models: paradigmatic and syntagmatic (i.e. topical relatedness) models. Systems that capture paradigmatic similarity tend to track co-occurrences within a small context window (Sahlgren, 2006). A small context window roughly corresponds to the number of words that occur in an average sentence or clause; in this way, relations between words that occupy different syntactic positions are tracked. When representations contain information about these syntactic relations, they tend to emphasize paradigmatic similarity. Models of this type produce high similarities for words that occur in the same context (i.e. are substitutable) but not at the same time (i.e. same event or topic). Models of this type have been used to discover part-of-speech classes in child-directed language (Keibel, 2005; T. H. Mintz et al., 2002; Redington et al., 1998).

One of the earliest paradigmatic models used to fit psycholinguistic data was the The Hyperspace Analogue to Language (HAL; Lund and Burgess, 1996). HAL counts co-occurrences within a bi-directional context window of 10 words, and uses linear weighting to de-emphasize co-occurrences spanned by greater distances. The original model was trained on a 300 million word corpus composed of text from Usenet, an online discussion forum. HAL has been used to account for many different priming phenomena, including distinguishing semantic and associative word priming (Lund & Burgess, 1996), priming of abstract and emotion words (Burgess & Lund, 1997), and mediated priming (Burgess et al., 1998). Many improvements of the original HAL algorithm have been proposed. For instance, D. L. Rohde et al. (2006) suggested down-sampling high-frequency context words, converting raw co-occurrence counts to correlation coefficients, and performing dimensionality reduction. The resulting model, COALS, performed comparably or better than its predecessor on all evaluation tasks, including the Test of English as a Foreign Language (TOEFL), a standardised test to measure the English language ability of non-native speakers.

While small context windows tend to produce clusters that respect part-of-speech distinctions and de-emphasize topical similarity, large windows tend to produce clusters of topically related words. To capture topical similarity, window sizes often span entire documents or paragraphs. Given that documents tend to describe events and situations that belong to one or a small number of topics, words that tend to co-occur in the same documents are likely topically related. One of the earliest models to be proposed along these lines is Latent Semantic Analysis (LSA; Landauer and Dumais, 1997). LSA was trained on 4.6 million words that combined Grolier's American Academic Encyclopedia and the Touchstone Applied Science Associates (TASA) corpus of educational materials. The raw co-occurrence counts were normalized and transformed by the row-wise entropy. Next, similar columns in the co-occurrence matrix were collapsed using Singular Value Decomposition (SVD) to produce, what Landauer and Dumais (1997) called, 'latent dimensions'. This step can be viewed as inferring 'indirect co-occurrences', with the effect of making more similar words that occurred in similar — despite not identical — contexts. This abstraction step, Landauer and Dumais (1997) argue, allows LSA to capture word similarity judgments in a more human-like fashion. Thus, we have seen there are many ways to build co-occurrence matrices. An outstanding challenge in distributional modeling is the development of systems that can capture both paradigmatic and topical similarity in a unified model.

One of the downsides of working with distributional semantic models is the large number of parameters that need tuning, and little theoretical guidance for how to set parameters or how they influence the performance of the resulting model on a specific task. Instead, practitioners tend to tune models on a task-by-task basis, which raises questions about the task-specific nature of human semantic memory. A persistent concern is how to best normalize co-occurrence counts, and how to reduce the dimensionality of the co-occurrence matrix. Raw co-occurrence matrices tend to be very big (in the thousands or tens of thousands), and dimensionality reduction can help with efficiency. Further, dimensionality reduction, when performed using singular value decomposition (SVD), can help remove noise and often produces composite dimensions that code higher-order co-occurrence relations. These composite dimensions are sometimes considered to code abstract knowledge that goes beyond raw co-occurrence, and which takes into consideration the similarity among multiple words and multiple contexts simultaneously. While practically and theoretically of interest, Bullinaria and Levy (2012) note that "there appears to be no robust theoretical underpinning to SVD, and the optimal dimensionality appears to have to be determined empirically for each new application domain". Another downside is that the co-occurrences matrices are probably too big and too precise to be realistic models of children's semantic development. While the co-occurrence matrix is a convenient computational-level metaphor, it is unlikely that an instantiation at the algorithmic level will look anything like HAL or LSA. While this is not a major concern for theorizing about the organization and structure of human semantic memory, it is when we attempt to explain the kinds of computations that a young word learner is able to perform under realistic conditions, as I am in this thesis.

There are other kinds of models that are not based on storing counts in a large co-occurrence matrix. For instance, (M. N. Jones & Mewhort, 2007) proposed BEAGLE which is based on random-vector accumulation. At initialization, random patterns of activation are assigned to each word in the vocabulary, supposed to approximate unique differences in meaning, phonology, and orthography. During training, as words are encountered, vectors corresponding to each word are added together to form new patterns of activations. In addition to breaking free from reliance on a large co-occurrence matrix, BEAGLE is also much more sensitive to word-order than previous models. To capture word-order information, the vector representation of words that occur at different distances from a target word are integrated into the vector representation of the target word using slightly different operations. This is accomplished using holographic vector combination methods. The authors show that the integration of word-order information provides the model with a much better fit to human data in a variety of semantic tasks (M. N. Jones & Mewhort, 2007; Recchia et al., 2010).

I consider the aforementioned systems as 'traditional' distributional semantic models because they were specifically developed with modeling psychological aspects of lexical semantics in mind. In contrast, one might consider a different class of models that were not explicitly designed for such purposes, but which are nonetheless based on similar principles. As mentioned before, language models also learn from language-internal lexical co-occurrence, and, in particular, neural language models also learn distributed representations of word-like units from unannotated corpus data and without expert supervision. I consider the latter class of models to be 'distributional' in a broader, non-traditional, sense. In what follows, I characterize one such model, the simple RNN.

1.4 The Simple Recurrent Neural Network

The model that is investigated in this thesis is the simple Recurrent Neural Network (simple RNN, J. L. Elman, 1990). It stands out from other distributional semantic models on many dimensions, as discussed below. First, a brief introduction is warranted. A more detailed discussion of how learning takes place in the network, and potential limitations for learning lexical semantic representations is provided in Chapter 3.

Traditionally considered a model of language processing, the simple RNN operates over sequences of consecutive words, and is tasked with learning statistical regularities in language that are useful for predicting upcoming words. Over the course of training, the model learns to expect which words are more likely to occur next, and by so doing, encodes both grammatical and semantic constraints useful for next-word prediction. I will use the term 'language modeling' and 'language models' in the technical sense, to refer to the use of the next-word prediction objective for training, and models that are trained using this objective, respectively. Language models are of theoretical interest to cognitive scientists as they can be used to study learners that exclusively rely on statistics over language (Boleda, 2020; Lenci, 2018).

For brevity, I will refer to the simple RNN as the RNN, unless otherwise noted. The work herein does not draw a theoretical distinction between different varieties of RNN models; in fact, I will show that many of the observations about the simple RNN also hold in its more sophisticated cousin, the Long Short-Term

Memory (LSTM).

1.4.1 Previous Work

The first studies of the simple RNN showed that it could learn to predict next-words reasonably well, and that by doing so, the network encodes grammatical and semantic constraints implicit in the input language (Cleeremans et al., 1989; J. L. Elman, 1990, 1991). For example, J. L. Elman (1991) showed that the RNN could learn the regularities of an artificial corpus composed of thousands of pseudo-English sentences generated by a simple grammar that featured embedded clauses and number agreement. Once trained, the RNN was able to correctly predict next-words in the training data, as well as generalize learned processing dynamics to novel input sequences. Early work showed that the RNN performed well in a long-distance grammatical task that evaluated the networks ability to predict inflectional verb-endings. Strikingly, the model was able to use number (singular vs. plural) marking on a subject to infer the correct inflection on a subsequent verb — even in cases where the verb was separated from the subject by one or more embedded clauses. More specifically, principal component analysis of the hidden representations revealed that during the course of training on the next-word prediction task, the model encoded a hierarchical constituent structure that neatly separates input corresponding to different grammatical categories and their location in the parse tree (e.g. distinguishing between a noun in a main vs. embedded clause). The RNN's success at this task was due to its ability to compress information about past words into a compact distributed representation at the hidden layer. [6]

Follow-up work demonstrated the potential of the RNN for learning challenging grammatical structures. In particular, linguists have long claimed that some aspects of grammar are too difficult to acquire in an experience-dependent manner and provided only finite samples; this is the well-known 'poverty of the stimulus argument' (Chomsky et al., 1976; Gold, 1967; Marcus, 1993). Instead, innate knowledge is invoked to explain children's ability to make rapid progress acquiring the grammatical structures of their native language. A center piece of this innate knowledge is the ability to perform recursion, the successive application of identical operations on arbitrarily nested input. However, the simple RNN has been used to successfully challenge this proposal. For example, Rodriguez et al. (1999) showed that the network can produce correct completions for portions of unfamiliar samples drawn from a context-free grammar of the form $a^n b^n$.[7] The mastery of a context-free grammar by the simple RNN is an important counterargument to the claim by Chomsky (1957) that that recursion in natural language in principle rules out associative and finite state models of language processing like the RNN. Analyses of the internal dynamics of the model, based on non-linear dynamical systems theory, revealed that the simple RNN, trained on finite samples of $a^n b^n$, learned to count, a stepping-stone to recursion and the ability to process non-finite and context-free grammars. The demonstration that the simple RNN can in fact approximate recursive behavior when processing linguistic input, has cast doubt on whether innate constraints beyond those already encoded in the network's architecture are in fact necessary for mastery of human grammar (see also Christiansen and Chater, 1999a; Tabor, 2002).

[6] In a distributed representation, a concept is represented by a pattern of activations across an ensemble of units. Typically no single unit can convey a learned concept on its own; most of the time, each unit works in tandem with many other units to convey a single concept.

[7] In this grammar, the symbol b is repeated as many times as the symbol a, and a system is said to have successfully recognized (i.e. learned) the grammar only if it successfully predicts exactly the same number of b symbols as there are a symbols.

1.4.2 Reasons in Favor of the Simple RNN

There are several theoretical and empirical reasons why I chose the simple RNN as a potential model of how children might construct lexical semantic representation from distributional evidence. While the simple RNN has not been previously regarded as a distributional semantic model, I will show in this thesis, that it can be used in the same way that traditional distributional semantic models have been.

There are two primary reasons I decided to investigate the potential of the RNN as the supplier of lexical semantic representations as the input to a larger system responsible for performing DECAF. First, in contrast to traditional distributional semantic models, the simple RNN provides an opportunity to study lexical semantic development in the context of sequence-learning, which also enables the acquisition of grammatical knowledge. In other words, the simple RNN not only tracks co-occurrence frequency, but also information about co-occurrence distance, and the sequential structure in which words co-occur. This means that the learned representations will emphasize paradigmatic similarity between words, which is a crucial first step for being able to perform DECAF. Second, the simple RNN has the potential to account for both lexical semantic and grammatical development under one roof, which is an advantage from the perspective of a unified model of early language acquisition. Further, this would be in line with the proposal that children bootstrap the acquisition of grammar by first noticing sequential dependencies in the surface statistics of the language they hear (Bowerman, 1973). On this view, children learn what they can about word-order, before replacing their initial statistical hypothesis with more sophisticated grammatical hypotheses. Notice, also, that the simple RNN is compatible with children's discovery of syntactic cues used to preform syntactic bootstrapping. In sum, the simple RNN has a broader coverage of phenomena in language acquisition, and should therefore be the first model to be investigated before more specialized (i.e. traditional, lexical-level only) models are considered.

In addition, there are several other reasons why the RNN stands out among other distributional semantic models. I discuss each reason, in turn, below. This discussion is picked up in Chapter 4, where I consider theoretical desiderata a system should satisfy, or at least approximate, if it is to be considered a viable model of the supplier of lexical semantic representations needed to perform DECAF.

Rich History in Psycho-Linguistics

Another advantage of the simple RNN is its rich history in psycho-linguistics and cognitive science. In particular, the RNN is often invoked as a a model of theories of human language processing that emphasize constraint-based, probabilistic, or expectation-driven learning and processing (G. T. M. Altmann & Kamide, 1999; G. T. Altmann, 1998; Ford et al., 1982; John & McClelland, 1990; MacDonald et al., 1994; MacWhinney & Bates, 1989; McRae et al., 2005; McRae et al., 1998; Pannitto & Herbelot, 2022; Tanenhaus et al., 1989; Trueswell et al., 1994). In addition, variants of the RNN have been integrated in numerous models of language processing, comprehension and production (F. Chang et al., 2006; J. L. Elman & McRae, 2019; Fitz & Chang, 2017; Kirov & Cotterell, 2018; MacDonald & Christiansen, 2002; Misyak et al., 2010; D. L. T. Rohde, 2002). The incremental learning and processing dynamics of the RNN also fit well with empirical findings concerning online and incremental language processing. One of the earliest demonstrations that the information provided by individual words is used immediately during processing comes from a study by Marslen-Wilson et al. (1988) who measured the time it takes for participants to detect different types of linguistic violations within spoken sentences. Their findings showed that listeners use both a verb's semantic and syntactic information immediately and incrementally to buildup a representation of sentence meaning. More recently, reaction time and eye-movement studies of participants processing sentences with local syntactic ambiguities have validated and extended this idea (Boland & Tanenhaus, 1991; MacDonald et al., 1994; Stowe, 1989; Trueswell et al., 1994), among others.

It is worth nothing that the RNN is also the pre-cursor to many ground-breaking achievements in computational linguistics, including in speech recognition (Chan et al., 2016; A. Graves et al., 2013), text generation (Lu et al., 2018), text summarization (Andhale & Bewoor, 2016), and sequence classification (Chen et al., 2020; Warstadt et al., 2019).

Broad Coverage of Empirical Findings

As mentioned before, variants of the RNN have been used to explain a broad range of findings in the language acquisition literature. For brevity, I do not attempt to provide an exhaustive review. I have already mentioned that the RNN has been used to account for phenomena in online language processing, the discovery of part-of-speech classes, acquisition of long-distance morphosyntactic regularities, but there are many others. For instance, the RNN has been used to model human speech recognition (Magnuson et al., 2020), sequencing during production (F. Chang et al., 2006), reading time contrasts (Tabor et al., 1997b), and speech segmentation (Christiansen et al., 1998; Mirman et al., 2010). It is worth briefly discussing the latter: The learning algorithm that underlies the RNN is compatible with proposals in word segmentation research where it is often claimed that tracking syllable transition probabilities is useful for the discovery of word boundaries in fluent speech (Aslin et al., 1998; Saffran, Newport, et al., 1996). While the simple RNN is often used to perform next-word prediction, it is equally compatible with the prediction of characters, phonemes, and syllables. It is likely that children track transition probabilities at multiple levels that would enable the simultaneous discovery of sub-lexical, lexical, and super-lexical (e.g. phrasal boundaries) regularities. A challenge in this area is how to build a model that tracks regularities at different temporal scales (Carta et al., 2020; Gluck & Bower, 1988), and how to integrate regularities at each scale to construct a single meaning interpretation (Martin & Doumas, 2017). This discussion also raises the possibility that by supporting the identification of words in the speech stream, an RNN-like statistical learning system operating on syllable-like units can support word-referent mapping (for preliminary evidence, see Mirman et al., 2010). Thus, DECAF is not necessarily the only way in which language-internal statistical information can be put to use in word learning.

Cognitive Plausibility

From the perspective of cognitive plausibility, the simple RNN stands out relative to many other distributional semantic models: The simple RNN makes no assumptions about its input prior to learning, and is trained directly on natural language with little or no pre-processing.[8] The next-word prediction objective is grounded in cognitive theory (A. Clark, 2013; Friston, 2005; Kuperberg & Jaeger, 2016; Pickering & Garrod, 2007; Rao & Ballard, 1999; Suddendorf & Corballis, 2007) and makes contact with a rich neuro-scientific literature on neural correlates of prediction (G. T. M. Altmann & Mirković, 2009; Grisoni et al., 2017; Hubbard et al., 2019), and the possibility that prediction-like computations may arise from basic neural principles (Falandays et al., 2021). Furthermore, learning in the simple RNN is entirely self-supervised given that the ground-truth for next-word predictions are readily available in the training data. Further, learning in the simple RNN does

[8] However, an important determination that must be made by the researcher is how language is segmented (i.e. tokenized) into lexical units prior to training.

not rely on, for instance, pre-engineered contrastive loss functions (Mikolov, Chen, et al., 2013), linguistic analysis (etc. syntactic and/or morphological parsing), or the storage and explicit factorization of large co-occurrence matrices. Lastly, recent work comparing computational models to neural recordings of humans engaged in language tasks has revealed a close correspondence between the outputs of predictive models and neural activation in human brains (Schrimpf et al., 2021; Wang et al., 2020). Interestingly, models that perform better at predicting the next word in a sequence also better predicted brain measurements — providing computationally explicit evidence that predictive processing shapes the language comprehension mechanisms in the brain.

In addition, the RNN is considered a process model rather than a representation model. That is, the computations that the RNN performs are more closely aligned with the computations performed by the human brain, at Marr's algorithmic level of analysis. This does not make the RNN inherently superior relative to other distributional semantic models; instead, it means, that by studying the RNN, we are studying computations that, to the best of our current understanding, are likely similar to those actually performed by people learning and processing language (for a critical discussion, see Lillicrap and Santoro, 2019; Xie and Seung, 2003). This point is particularly important for the purpose of this thesis, because my aim is not only to establish DECAF as a procedure that children can perform, but a procedure that likely emerges in the kind of machinery that they already use to perform other language tasks. Demonstrating the latter would lend additional credence to DECAF as a veridical psychological construct.

That said, it should be kept in mind, that artificial neural networks — the RNN among them — are not models of biological processes. In the words of Hinton (1990), artificial neural nets "emphasize computational power rather than biological fidelity". That is, their superficial similarity to biological networks should not be mistaken as an attempt to account for happenings at Marr's implementational level (in this case, the neural hardware). Neural networks are still cognitive models, and are not intended to simulate computations performed by individual neurons. There are many aspects of cell biology, anatomy, and electrophyisiology that are completely ignored by most artificial neural networks (but see Gluck and Rumelhart, 2013).

Connections to Neural Dynamics

Unlike traditional semantic models based on co-occurrence counting, the simple RNN makes contact with principles of neural computation. For example, recurrence is a computational primitive that shows up time and again in studies of neural computation and systems neuroscience. For example, Yang et al. (2019) found that recurrent units can develop into clusters that are functionally specialized for different cognitive processes, and that the specialization that emerges in a recurrent network is similar to the task selectivity of prefrontal neurons in people. This specialization was observed under developmentally plausible conditions, in which tasks were learned one after another using a continual-learning technique. Similarly, recurrent systems trained to perform navigation tasks in 2D arenas converge on a functional specialization between border-cells, band-cells, and grid-cells that has been observed in the human entorhinal cortex (Cueva & Wei, 2018).

The kind of computation that the RNN performs is also compatible with the 'predictive coding' framework for understanding computation in biological systems. Predictive coding theories assume that the brain continuously forms top-down predictions about upcoming events, and propagates these from higher to lower cortical levels (Aitchison & Lengyel, 2017; A. Clark, 2013; Friston, 2005, 2010). Recurrent neural networks have been used to explain how predictive coding might emerge in neural system. For instance, Ali et al. (2021) observed that predictive coding does not require architectural hard-wiring in the form of specialized hierarchical connectivity (i.e. excitatory bottom-up signals and inhibitory top-down feedback), but readily emerges in a recurrent neural network provided metabolic constraints on energy usage (e.g. regularization imposed on the magnitude of activity patterns).

The RNN makes additional contact with neural dynamics by virtue of having been used to successfully account for patterns of scalp-recorded neural activity during language processing. In particular, EEG recordings of humans engaged in language comprehension tasks are marked by a characteristic event-related signal — called the N400 — that has been associated with retrieval and/or activation of semantic features of words (but also other non-linguistic stimuli) encountered in the task. Federmeier (2022) argues that the N400 arises due to a mixture of factors that are not limited to predictability of a word. Federmeier (2022) writes that prediction is a “separate, non-obligatory factor that can shape activation states in semantic memory and thereby affect the N400” rather than an error-signal related to mis-prediction alone. That said, aspects of the N400 that can be accounted for by prediction error have been successfully modeled by an RNN trained to predict message-level information during online sentence comprehension (Rabovsky et al., 2018; Rabovsky & McClelland, 2020).

Further, the RNN is a discrete-time non-linear dynamical system, and as such can be viewed as an approximation of continuous-time non-linear dynamical systems known to capture aspects of computation in biological systems (Izhikevich, 2007; Meyer-Lindenberg et al., 2002). The nonlinear and continuous valued nature of dynamical systems allows them to respond flexibly; responses may be graded under some circumstances, while in others, responses are more binary. Dynamical analyses figure prominently in the cognitive science literature: L. B. Smith and Thelen (1993) and Spencer and Schöner (2003) examined the role of dynamical systems in cognitive development; Spivey (2008), Tabor et al. (1997a), and Tabor et al. (2004) proposed dynamical accounts of sentence processing (for a review, see McClelland et al., 2010). The dynamical systems perspective has proven useful for characterizing the emergence of complex adaptive behavior in biological systems faced with enormous uncertainty and a need to flexibly respond and reshape learned routines when environmental conditions change (Arbib & Érdi, 2000; Rosenfeld et al., 1987). For an example of an analysis of the simple RNN from the perspective of dynamical systems theory, see Rodriguez et al., 1999.

1.4.3 The Next-Word Prediction Objective

By considering the RNN as the supplier of lexical semantic representations to DECAF, one is implicitly testing the hypothesis that next-word prediction can construct lexical semantic representations. In this section, in an attempt to defend my choice of model, I examine (next-word) prediction more closely.

Prediction, as J. L. Elman (2011) points out, is:

> ...a simple behavior with enormously important consequences. It is a powerful engine for learning, and provides important clues to latent abstract structure (as in language). Prediction lays the groundwork for learning about causation. [...] It should not be surprising that prediction would also be exploited for language learning and play a role in on-line language comprehension

Further, Dell and Chang (2014) argue that prediction is ubiquitous in language. According to the authors, prediction-error is not only employed to mold the language production system, but it is also important for language comprehension and acquisition.

But not all are convinced. There are many that have criticized the notion that prediction and error-based learning alone are sufficient to explain aspects of language acquisition. However, almost nobody argues that next-word prediction is the primary — let alone, the sole — mechanism responsible for the acquisition of language abilities. In most theories in which prediction takes the front-row seat, the role of prediction tends to be supportive rather than the sole basis on which linguistic knowledge is built. With so much interest and debate surrounding this issue, a growing number of studies have examined the cognitive and neural plausibility of predictive processing in people engaged in language tasks, some of which I review below. It should be noted, however, that existing evidence is is often mixed, and that strong conclusions about the role of prediction cannot be made at this point in time. To arrive at stronger conclusions, psychologists and neuroscientists require a more nuanced understanding of computation in human brains than is currently available. That said, emerging evidence from a variety of experimental paradigms suggests that there are deep similarities between language processing in people and in prediction-based artificial neural networks like the simple RNN.

One line of evidence involves correlating the behavior of neural language models with neural recordings during language processing. For instance, Schrimpf et al. (2021) examined the relationships between a diverse array of neural language models (including word-embedding models, RNNs, and Transformers) across three neural language comprehension datasets (two fMRI, one ECoG), as well as behavioral signatures of human language processing in the form of self-paced reading times. The authors found that both Transformers and RNNs generalized to held out neural data, lending credibility to the claim that prediction-based models trained on language data alone can explain variance in brain activation patterns involved in language processing. Second, the authors found that Transformers provided a better fit than RNNs, and strikingly, explained virtually all of the variance in neural responses to sentences. Third, only models trained on the next-word prediction task were predictive of neural or behavioral data. Schrimpf et al. (2021) remark that:

> ...these findings provide strong evidence for a classic hypothesis about the computations underlying human language understanding, that the brain's language system is optimized for predictive processing in the service of meaning extraction.

Another line of evidence for computational similarities between RNN-like mechanisms and language processing in people comes from studies that investigated the sensitivity of people to next-word prediction error. Like many connectionist systems, the simple RNN learns via backpropagation of the next-word prediction error (J. L. Elman, 1990). Thus, to ask whether the RNN is a good model of learning in children, we may ask whether there is any evidence that children use error-based learning to learn new knowledge and update existing knowledge. Some behavioral evidence exists to support such a notion. For instance, (Ramscar et al., 2013b) found that 2- to 3-year-old children learn words in accordance with error-driven principles. Rather than simply counting positive co-occurrences of words and potential referents, the authors found that children also tracked negative co-occurrences, which would allow them to compute expectations (quantified using 'informativity') about word-referent relationships. The authors accomplished this by exposing children to ambiguous object-label pairings across two presentations in which only one of the objects was repeatedly presented. Importantly the target object (i.e. the object presented in both situations) was paired with two different labels, and alongside competing objects. If children are indeed sensitive to the informativity of the object-label pairings, their responses should be systematically biased toward the label that is consistent with their experiences across both presentations, and prefer that label over a novel label, which, albeit an equally logical alternative, is less consistent with their experience. In accordance with error-based learning, children's pattern of matching objects to labels matched well with the informativity of each object. Interestingly, this was less true of the responses made by older participants (undergraduate students). The latter observation suggests that error-based learning is greatest in the developmental stage during which language acquisition takes place — perhaps not a coincidence.

A third line of support for the idea that prediction of upcoming words plays a role in lexical semantic development comes from eye-tracking studies of people engaged in online language processing. For instance, G. T. M. Altmann and Kamide (1999) observed that people predict upcoming event participants in line with previously presented semantic cues. Having heard the start of a sentence that cues an eating event, participants' eyes tend to fixate on visual depictions of items that are semantically congruent (i.e. an image of a cake, rather than a carousel). This phenomenon is often referred to as 'predictive processing'. Follow-up work has found similar evidence, such as prediction of semantically congruent sentence components in L2 listeners (Dijkgraaf et al., 2017) and both skilled and unskilled simultaneous translators (Amos et al., 2022). Many other eye-tracking studies have converged on similar conclusions, with the inclusion of visual context (Mani & Huettig, 2012; Rommers et al., 2013), and others that do not involve a visual context (Grisoni et al., 2017).

There is a debate in psycholinguistics surrounding what prediction is good for. Some have proposed that prediction of upcoming words facilitates processing, and bootstraps comprehension via emulating production (Federmeier, 2022; Pickering & Garrod, 2007). Similarly, some scholars have suggested that people do indeed often predict words to come, but only when there are predictions to be made. For instance the pre-activation theory proposed by Huettig et al. (2022) considers prediction as a consequence of fully activating previously pre-activated linguistic representations. On this view, when no representations have been previously pre-activated, prediction does not occur. Others, have proposed a much more comprehensive role for prediction. For instance, some scholars consider the formation of expectancies during language processing as a primary source of information about event knowledge (J. L. Elman & McRae, 2019; McRae et al., 2005). While there is considerable evidence for the role of prediction in semantic and grammatical processing, it is less clear whether learning is in part due to prediction. Under an account of predictive processing, learning is shaped by error minimization, and this means that costs associated with mis-prediction should be detectable. While evidence from event-related potential (ERP) studies demonstrates that people predict upcoming words and incur mis-prediction costs when predictions are not correct (Hubbard et al., 2019), others have not. For instance, Luke and Christianson (2016) did not find evidence of mis-prediction costs associated with eye-movement during silent reading. Instead, the authors found that when a strongly expected word was replaced by a less expected word (using Cloze probabilities as a measure of expectancy), several eye-tracking related measures showed a facilitative effect rather than a slow-down in processing. Combined with their finding that highly predictable words of the sort used in most prediction experiments occur rarely, Luke and Christianson (2016) argued that over-active reliance on prediction would be disruptive and misleading instead of facilitative. Clearly, the role of prediction and error-based learning in language learning, and comprehension requires further work. As Luke and Christianson (2016) suggest, it is possible that prediction only occurs under certain circumstances, especially when the listener has available additional information, such as speaker-listener interactions, for making more accurate next-word predictions.

1.4.4 Potential Obstacles

While there are many advantages of using the RNN in modeling how people perform language tasks, the RNN presents unique challenges that may make it unsuitable as a supplier of lexical semantic representations that children can use to perform DECAF.

At Odds: Uncertainty Maximization and Minimization

In principle, there appears to be a fundamental trade-off between 1) learning lexical categories, and 2) minimizing next-word prediction error. On the one hand, the discovery of lexical categories is supported by boundaries that are marked by high uncertainty, such as between the semantically vacuous determiner *the* and a subsequent noun. Not knowing which word comes next is precisely the basis on which category knowledge is built: There are many plausible continuations, and many of them are nouns. The set of plausible next-words (after the determiner *the*), can therefore be considered members of the same lexical category. It is precisely because nouns are substitutable in the same contexts that one can speak of nouns as members of a coherent category. On the other hand, high uncertainty, the corner stone of lexical category formation, is exactly what learning systems based on next-word prediction error are supposed to minimize (Onnis et al., 2003). Thus, it appears that uncertainty maximization needed for the discovery of lexical categories is therefore at odds with uncertainty minimization needed to learn the sequential structure of language. The idea that such a trade-off exists is corroborated by evidence from the influence of prefixes and suffixes on learnability and processing. While suffixes shared by members of the same lexical category can facilitate the discovery of grammatical categories by young children (St. Clair et al., 2009), prefixes reduce uncertainty about what comes next, and therefore sacrifices category learning. Instead, the presence of prefixes promote online comprehension and production. A version of this argument also appears in (Dye et al., 2017) and Ramscar et al. (2013a).

The reason that the discovery of POS classes is possible in the RNN at all is because natural languages tend to be structured such that the precise identity of words in content-word slots cannot be accurately predicted. The RNN can perform error minimization as much as it wants, but there tend to be few contextual cues to latch onto in order to minimize uncertainty at content-word slot boundaries. At least, this is the case when extremely small windows are considered. The word that immediately precedes or follows a noun typically provides very little information about the identity of the noun (or its semantic category). This is especially true in child-directed input, where nouns are frequently framed by semantically vacuous lexical context (Chapters 7, and 8). This works well for the discovery of POS classes, but is this also true for finer-grained semantic categories? Probably not. In order to discover semantic categories, we require semantic cues, and these are often found much farther away from a target word than neighboring items. By increasing the window size over which statistical associations are tracked, however, we run into potential issues that might impede the formation of lexical semantic category information in the RNN.

Essentially, by increasing the window size, the RNN is given more room to memorize idiosyncratic multi-word sequences at the expense of lexical-level statistics. By memorizing which words follow one another in frequent chunks (i.e. *the* predicts *dog* which, in turn predicts *I*, which in turn predicts *saw*, and so on), the RNN can avoid learning long-distance, structure-dependent associations that are potentially useful for clustering words into semantic categories. The tendency to form chunk-level knowledge at the expense of discovering and encoding structure-dependent long-distance dependencies has been attested by previous work on the RNN (Cleeremans et al., 1989; D. L. Rohde & Plaut, 1999; Servan-Schreiber et al., 1991). The possibility for minimizing prediction error across chunk-level units is something that is unique to the RNN and other language models; traditional models of distributional semantics typically do not attempt to track word-order statistics explicitly, or have ways to mitigate their effect on the quality of learned representations (M. N. Jones & Mewhort, 2007). Prediction-error minimization thus presents a unique challenge for modeling the construction of form-based lexical semantic categories. With a wide enough window needed to exploit semantic cues, the possibility of minimizing error associated with predictable chunk-level sequential statistics de-emphasizes the discovery of uncertainty boundaries that might signal semantic category membership. This chunk-level knowledge also has the effect of mixing and conflating category-relevant with category-irrelevant statistics, which need to be separated in order to produce lexical semantic representations that are useful for performing DECAF.[9] Simply put, the emphasis on chunk-level error minimization introduces noise into lexical semantic representations, and therefore make them less useful for the distributionally-mediated extension of category-associated features (DECAF). Mixing category-relevant with category-irrelevant statistics, while potentially useful for other aspects of language acquisition, increases the likelihood of making inference errors during DECAF. Next, I further explore this potential conflict of interest, and what can be done to mitigate potential negative consequences on word learning.

Chained Conditional Probabilities

Why does the RNN encode chunk-level statistics at the expense of potentially more generalizable lexical-level statistics?

The short answer is the network's (over-)reliance on accumulated context during next-word prediction. In essence, the RNN is forced to integrate information from previous time steps with the information provided at the current time step, and has no straightforward way to ignore (or delay the processing of) information that has already been accumulated in memory. A randomly initialized RNN does not have specialized units dedicated to time-delayed processing that would permit the model to make decisions about when information accumulated in memory should be combined with new information. While an RNN may learn to ignore information or remove information from its memory, this ability, however, must be *learned*, and is not readily available to a randomly initialized network. Further, a principled organization of this type does not readily emerge even when trained on lots of data (Linzen & Baroni, 2021). Without this ability, the representations that are learned by the RNN early during training are influenced by all of the information that has accumulated in the hidden layer up to the time step at which a target word is input to the network. Much of this information is often irrelevant for learning about semantic category membership or may result in the encoding of spurious, maladaptive statistical associations.

A more technical answer can be found in the formal description of the objective that the RNN language model is attempting to solve. Technically speaking, the language modeling objective is to minimize the error associated with the estimation of a joint probability distribution over multi-word sequences, $P(w_{t-k}, ..., w_t)$, where k is the window size, and $w_{t-k}, ..., w_t$ is a multi-word sequence starting at time step $t-k$ and ending at t. Importantly, this joint probability is factored into a chain of conditional probabilities, $P(w_{t-k}, ..., w_t) = \prod_{i=1}^{t} P(w_i | w_{i-k-1}, ..., w_{i-1})$, each of which estimates the probability of the next word conditioned on the previous context. Here, w_t is the next-word, and $w_{i-k-1}...w_{i-1}$ is the sequence of words that has been observed at the point at which the next-word prediction is made. What this means is that the RNN language model is attempting to estimate a k-order conditional probability distribution.[10] Speaking more plainly, the RNN does not fundamentally operate on individual words, nor does it model how individual words

[9] I do not mean to suggest that chunk-level knowledge should be avoided by the language system altogether. In fact, such a proposal would not hold up to behavioral work that has demonstrated that children not only store information about individual words, but also chunk-level knowledge of multi-word sequences. Specifically, Bannard and Matthews (2008) observed that 2- and 3-year olds were significantly more likely to repeat frequent multi-word sequences correctly than to repeat infrequent sequences correctly. In fact, Arnon and Christiansen (2017) showed that chunk-level knowledge can *support* the acquisition of aspects of grammar such as agreement patterns and verb–preposition pairings.

[10] Contrast this with Markov models of order $k = 1$, commonly called Markov chains (Baum & Petrie, 1966; Shannon, 1948). In contrast to the RNN, these models assume that the next state of a system (e.g. next-word) only depends on the current state (e.g. current word).

relate to their contexts independently of intervening items.[11] Specifically, the RNN models the probability that a given target word occurs *conditional* on the items that have already occurred in the target word's left-context — and the probability that an item in the left-context occurs is itself conditionally dependent on the sequence of prior words. A consequence of this is that much of the predictive uncertainty that is needed to discover lexical semantic categories can be soaked up by words that appear in the left-context of a target word. In particular, I will show that the more conditional information that is encoded in the RNN's lexical representation, the less useful the learned lexical representations are for downstream tasks, such as the distributional extension of category-associated features (DECAF). That is, the more redundant information that is provided by the left-context about an upcoming semantically informative cue, the more likely it is that the semantic information is encoded in (or 'soaked up by') the representations of items in the left-context relative to the target word itself.

Lexical Atomicity: A Heuristic for Young Word Learners

From the perspective of learning lexical semantic representations that are useful for performing DECAF, chunk-level memorization is problematic for several reasons. Not only are chunk-level statistics (i) less generalizable than lower-order statistics (e.g. binary relations as opposed to ternary relations), they also (ii) present unique challenges related to retrieval, and (iii) raise questions about how multiple relevant chunk-levels statistics should be integrated prior to performing DECAF. To illustrate all these issues, consider, for example, that '*your pink bunny*' predicts *eats*, and '*her green bunny*' predicts *jumps.* By noticing that each word is predictive of the next, the RNN is able to minimize prediction-error by learning to stitch together one word after another in each sequence. However, by doing so, the RNN is at risk of trapping semantic information about the word *bunny* in two separate chunk-level sequence representations rather than a single representation dedicated to *bunny.* Without integration, the chunks '*your pink bunny*' and '*her green bunny*' are stored separately, and without knowledge that each has the noun *bunny* in common. The risk of foregoing integration of chunk-level statistics at the lexical-level, is that a child may not be able to make strong inferences about novel words that are distributionally similar to *bunny.* While the child may draw upon his or her chunk-level knowledge, it is less clear how chunk-level knowledge would be accessed relative to an integrated representation of *bunny* which could be stored as a standalone pattern in a dictionary-like data structure.[12]

Further, access to chunk-level representations would be accompanied by category-irrelevant statistical baggage, such as the preceding contexts '*your pink X*' and '*her green X*'. This additional category-irrelevant context would encourage the child to make more exact matches between stored chunk-level knowledge and novel linguistic material. Because exact matches between longer strings are much less likely to occur that matches between shorter strings, a child would be better off with shorter chunks, or, ideally, with lexical-level representations.

[11] The RNN does not model the probability that a target word co-occurs with an arbitrarily distant context (word or multi-word sequence).

[12] Certainly, chunks may be stored as static entries in a dictionary data structure too. However, the RNN itself does not have chunk-level storage. To compute a chunk-level representation, one would have to supply the RNN with the words that make up the chunk, and then extract the pattern of activation at the hidden layer. Further, given that multiple chunk-level representations may be relevant for performing DECAF, this procedure would have to be repeated multiple times. That said, one might argue that these challenges are unique to the RNN, and that children do store chunk-level representations that can be readily accessed with limited computational overhead. While this is a possibility, this would raise questions about whether static storage of linguistic chunks is cognitively plausible given the need for enormous memory resources.

One response to the issues raised above is to fully embrace language statistics, and to trust that once the statistics have converged, the child will be able to extract generalizable lexical-level statistics given that spurious associations will have canceled out or been smoothed into non-existence. While I agree that language statistics can be incredibly powerful, especially with lots and lots of data, I am less hopeful that a language-learning child prior to his sixth birthday has accumulated anywhere near a sufficient amount of linguistic experience to consider his or her collected statistics to have 'converged'.[13] For the purpose of this thesis, which aims to understand the influence of the early language environment on semantic development in children, it is especially important to consider that young learners have relatively limited experience with language statistics, and that a passive accumulation of statistics may need to be supplemented with subtle heuristics that help them succeed in the tasks those statistics are supposed to facilitate. For instance, to get the most out of their limited experience (e.g. performing DECAF), young learners would be ill-advised to rely on chunk-level sequential statistics. Instead, a useful rule-of-thumb would be to preferentially attend to statistical associations between an individual word (e.g. *bunny*) and an upcoming semantic cue (e.g. '*eats carrots*') while de-emphasizing the influence of surrounding and/or intervening context (e.g. '*your pink X*'). Alternatively, one can understand lexical atomicity as a force that promotes integration: Assume that learned associations between the complex noun phrase '*your pink bunny*' and upcoming category-relevant semantic cues are also equally applicable to just the head of the complex noun phrase, *bunny*. Because paying special attention to the lexical-level, and de-emphasizing category-irrelevant associations are so important to the work presented herein, I will refer to this heuristic as a preference for 'lexical atomicity'. I discuss this concept in much greater detail in Chapter 4. Suffice to say, that lexical atomicity appears to be a useful heuristic for young learners hoping to use their accumulated distributional knowledge to guide their extension of category-associated semantic features to novel words. Importantly, lexical atomicity does not necessarily require specialized machinery or innate knowledge, is not a prerequisite for performing DECAF, and should not even be considered a long-term learning goal. I consider lexical atomicity as a quick and dirty strategy to help make the most of distributional statistics right away, and to get the distributional extension of category-associated features (DECAF) off the ground as soon as possible.

Entanglement: An Obstacle to Atomicity

As previously mentioned, the foremost obstacle to lexical atomicity in the RNN is the network's tendency to learn chunk-level sequential statistics at the expense of more generalizable lexical-level statistics. Underneath this tendency, however, lies a deeper reason, namely entanglement. Entanglement is a complicated topic in neural network research and machine learning more broadly; it concerns the tendency of some learning systems to dedicate overlapping computational and representational resources to the processing of information that is causally related to independent components (i.e. factors) in the input data. In the context of neural networks, entanglement arises due to the nature of the distributed code used to represent concepts. In distributed representations, the same computational units are dedicated to processing all dimensions of the incoming input simultaneously; while this promotes the discovery of complex interactions, it often occurs at the expense of separating features that are causally unrelated. In contrast to sparse or localist codes, distributed codes are vulnerable to the formation of maladaptive conjunctive representations, which intermix separable factors of the input data, and thereby reduce their generalizability to novel data (O'Reilly, 2001). The formation of maladaptive conjunctive codes (i.e. entanglement) is further promoted by interactivity, such as recurrent

[13] Note, that whether this is even possible is widely debated. See, for instance Chomsky, 1957.

feedback connections, as is the case in the RNN. Entanglement is also more likely to occur in conditions in which there is a large number of input dimensions (e.g. large vocabulary), and low signal-to-noise ratio (e.g large proportion of category-irrelevant vs. category-relevant examples). Language modeling with recurrent neural networks, is therefore, a breeding ground for entanglement.

To demonstrate why entanglement is potentially maladaptive for learning lexical semantic representations, consider Figure 1.3. For each sentence, I have marked the semantically informative lexical dependency using a dependency-arrow. The point is to demonstrate that not all items in the input are relevant to a target semantic relation (membership in ANIMAL). Consequently, a learning system that knows virtually nothing about the structure of the input language, will have a difficult time separating dependencies that are relevant from those that are not. In a highly interactive neural network, like the RNN, it may take a long time to converge on a sparse internal organization of computational units that neatly separate different kinds of structure-dependent relations from each other. For instance, it would be useful to dedicate computational units specifically to the extraction of dependencies between the head of a subject noun phrase, and the main verb (e.g. *gorilla-eats*, *bird-eats*). This is difficult, however, given that surrounding and/or intervening context must be ignored (e.g. '*big black X*', '*X with red wings*'). The ability to perform this is often termed 'predication'. Predication is closely related to the notion of lexical atomicity. A system that can perform predication, should have little trouble learning atomic lexical semantic representations.

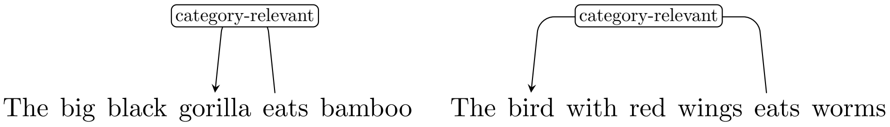

category-relevant
The big black gorilla eats bamboo
category-relevant
The bird with red wings eats worms

Figure 1.3: Category-relevant dependencies are relatively sparse in natural language sentences. In the sentences '*The big black gorilla eats bamboo*' and '*The bird with red wings eats worms*', only the dependency labeled by the arrow is informative about the semantic category membership of the subject nouns *gorilla* and *bird*. Entanglement occurs when this dependency is not encoded independently from other possible associations between words in the same sentence (subject and adjective, and/or subject and relative clause).

Predication is unlikely to emerge in language models such as the simple RNN because such systems are neither given privileged access to the syntactic structures of sentences nor cues to the structure-dependency of semantic relations between words (Gershman & Tenenbaum, 2015). Instead, the network is only provided strings as input, which do not straightforwardly advertise their syntactic or semantic structure. If predicates (i.e. structure dependent relations) are to emerge during training, they are almost certainly going to do so gradually, after having observed multiple instances of a relation (e.g. between head of subject and verb) and applied to different combinations of arguments (e.g. *dog*, *cat*, *airplane*, and *under the table*, *on the window*, *in the sky*). While the RNN can, in principle, discover abstract relations and learn to predicate those relations, it would likely take a long time. For instance, in order to predicate the semantic category-relevant relationship between *bird* and *eats* in Figure 1.3, the RNN would have to recognize that this relationship is (i) only one instance of a more general structural relation between heads of subjects and main verbs, and that it is (ii) independent of the content of intervening material. To recognize such a structural relationship, the RNN would have to detect featural invariants of this relation, and isolate these invariants from other properties that co-occurred in the input. This is a slow and and arduous process, given that the RNN cannot immediately rely on principled decomposition strategies such as constituency parsing or semantic-dependency parsing. Instead, the RNN must rely on the gradual dampening of irrelevant (idiosyncratic) features over the course of training on large amounts of diverse language data.

1.4.5 Accessing and Abstracting over Contextualized Representations

It should be noted that whether or not entanglement is actually problematic depends on the target task. If the goal is to capture multiple simultaneous constraints on online language processing, entanglement would be considered an advantage over less open-ended, structured systems. It is only when the RNN is used to learn lexical semantic representations, that entanglement is potentially an obstacle. On this view, we might stop to consider whether it even makes sense to coerce the RNN to learn something it was not developed to perform well. Is the learning of lexical as opposed to chunk-level contextualized representations antithetical to the purpose the RNN was developed for? Proponents of probabilistic, graded, and constraint-based approaches to language might argue that it is. Indeed, J. L. Elman (1990), the creator of the simple RNN says this about attempting to extract knowledge about an individual words from the hidden layer of the RNN:

> ...it is incorrect to speak of the hidden unit patterns as word representations in the conventional sense, since these patterns also reflect the prior context.

Despite this, J. L. Elman (1990) extracted contextualized representations[14] and showed that by averaging across contexts, one can study lexical-level phenomena. More precisely, J. L. Elman (1990) showed that the similarity structure between learned contextualized representations can be interpreted as a measure of grammatical and semantic similarity between the words they represent. While this is a useful way to study and interpret the RNN, this procedure is questionable from the perspective of modeling how children might access lexical knowledge from an RNN-like system. Not only does this procedure require an external system to access precise knowledge about previous contexts in which a target word has occurred, but it also requires re-playing those sequences so that the experimenter can extract the resulting hidden layer states. In addition, the resulting collection of hidden states must be averaged to smooth over the activation patterns contributed to by the context. This raises several concerns: First, the experimenter is doing the integration/abstraction himself, rather than the RNN. More importantly, if the goal of this ad-hoc abstraction procedure is to 'remove' information due to the contexts in which a target word has occurred, why insist on a complicated procedure that starts with contextualized representations and ends with removing context information? Wouldn't a more direct approach — simply extracting the static lexical semantic representation in the input-to-hidden weights — be more straightforward and cognitively more plausible? Accessing the contextualized representations at the hidden layer is easier said than done; it requires an external system that tracks precisely which contexts a child has heard every word in and then feeds those sequences back into the RNN. It strikes me that this process may not be sufficiently expedient to support word learning, and may be computationally untenable (but see Arnon and Clark, 2011; Bannard and Matthews, 2008; C. L. Jacobs et al., 2016; Rubino and Pine, 1998 for evidence that children do store multi-word sequences). It is possible that all this external machinery is not needed to generate useful lexical representations, and that the static lexical semantic representations that emerge at the input-to-hidden weights are sufficient. Whether this turns out to be the case is a major question pursued in this thesis.

1.4.6 Promoting Atomicity with Developmentally Plausible Training

The demonstrations presented in this thesis show that the simple RNN does not fully meet several of the desiderata discussed above in order to be considered a viable model of children's distributional category-based

[14] Contextualized representations are hidden layer activations in response to a target word in a particular context

induction. In particular, the network struggles to acquire atomic lexical semantic representations under all but the most restricted circumstances, in which the training data has been carefully counterbalanced. To promote the formation of more atomic lexical semantic representations, I propose a novel training regime, in which data is presented to the RNN in stages. In particular, this proposal takes advantage of the observations that the lexical contexts in which nouns occur are less idiosyncratic in language spoken to younger compared to older English-learning children. Based on this, I propose that in order to learn more atomic lexical representations of nouns, the RNN must concentrate on those structural regularities *before* learning finer-grained lexical regularities that distinguish semantic classes of nouns from each other. By so doing, the RNN should become preferentially tuned to structural relations during early training, and this should promote the extraction of category-relevant relations with less interference from category-irrelevant co-occurring items. The results show that training incrementally on an age-ordered corpus of English child-directed input has exactly this effect, and has long-lasting advantages for learning more atomic lexical semantic representations. This observation highlights the need (i) to move away from general claims about the advantages and disadvantages of context-sensitivity and towards a more subtle understanding of the interplay of different constraints on learning at different times during training, and (ii) to consider the developmental trajectory of children's language environment when training and evaluating models of language acquisition.

The novel training regime proposed above takes advantage of the fact that predictive processing — a type of associative learning — is susceptible to a phenomenon called blocking, whereby a previously learned association between a stimulus and target reduces the likelihood that another, equally predictive, stimulus can elicit prediction of the target in the absence of the originally paired stimulus (Waldmann & Holyoak, 1992). A detailed discussion of the staged training regime is provided in Chapter 9.

1.5 Overview and Contributions

This thesis is organized as follows: In Chapter 2, I characterize longitudinal statistical changes in a corpus of child-directed transcribed speech, and demonstrate that input to children is not static but non-stationary across developmental time. This finding is used in a subsequent chapter to inform the development of an improved strategy for training RNNs on child-directed input. In chapter 3, I show that an RNN trained on child-directed input learns lexical representations that capture knowledge about semantic category membership of a large set of common nouns. In Chapter 4, I discuss desiderata that a computational model should satisfy if it is to used to supply lexical semantic representations for performing DECAF during language acquisition. In particular, I discuss the need to avoid or reduce the impact of entanglement of category-relevant and category-irrelevant information at the hidden layer, a phenomenon that arises due to the strong tendency of the RNN to encode chunk-level as opposed to lexical-level statistics. Further, I describe a property of computational systems that allows learned representations of atomic units (e.g. lexical items in natural language) to be flexibly re-used outside the context of the task and architecture in which the representations were acquired. I have termed this 'lexical atomicity', and argue that atomicity is, in principle, possible in the RNN, but is difficult to achieve when trained on natural language. In Chapter 5, I use artificial language corpora to study what kind of statistical properties influence the formation of lexical semantic representations, and what properties are most useful for learning atomic representations. In Chapter 6, I use these findings to motivate a theoretical account that seeks to explain what kind of linguistic input is needed to achieve lexical atomicity. This account, termed 'Semantic Property Inheritance' (SPIN) theory, provides insight into when and how atomic lexical representations are learned in the RNN. The theory states that the RNN cannot acquire atomically organized lexical semantic representations when trained on input where the target semantic category structure is highly fragmented. In Chapter 7, I delve deep into what I mean by 'fragmentation', and why fragmentation in natural language data is an obstacle to lexical atomicity in the RNN. Specifically, fragmentation is a measure of how coherent a category is (e.g. NOUN, VERB) from the perspective of a distributional learner whose aim is to discover clusters composed of distributionally similar words. Using this method, I demonstrate that distributional evidence for the noun category is more fragmented in input to older as opposed to younger children. In Chapter 8, I further develop this idea from an information-theoretic perspective, and compare and contrast tools for studying corpus data relevant to language modeling. Based on these findings and the insights provided by SPIN theory, I propose a novel training regime, based on staged/curriculum learning, to promote the formation of more atomic lexical semantic knowledge in the RNN. In particular, I suggest that lexical atomicity should be greater when the RNN is trained in age-order on child-directed input. In Chapter 9, I scrutinize this claim from the perspective of error minimization in neural networks, and distributed representations. In Chapter 10, I empirically confirm this prediction. I term this result the 'age-order' effect, and use follow up analyses of RNN learning dynamics to further support that SPIN theory is in fact a viable explanation of this effect. In Chapter 11, I connect my findings to existing work in a variety of disciplines, from machine learning, to language acquisition, and computational linguistics. I discuss implications of my findings for theories of language acquisition in Chapter 12, limitations and future directions in Chapter 13, and provide concluding remarks in Chapter 14.

1.6 Summary

The experiments reported herein (i) showcase the potential of the simple RNN for modeling the construction of form-based lexical semantic category knowledge, (ii) demonstrate how and when processing based on next-word prediction is at odds with learning atomic lexical semantic representations, (iii) highlight challenges that limit the model's applicability for modeling children's construction of form-based lexical semantic representations, and (iv) provide recommendations, and predictions for how to leverage the organization of children's non-stationary language input to address the model's shortcomings. Broadly, this work is an attempt to bridge the study of language learning in computational systems — in particular neural networks — and language learning in children. The approach adopted in this thesis is that in order to better understand whether existing neural network models are a good fit for modeling aspects of language acquisition in children, we must combine formal analyses of model learning dynamics, empirical investigations of learning outcomes under conditions that most closely resemble those of children learning in the real world, and corpus analyses of the quantity and quality of information present in the input available to children. Finally, a major contribution of this thesis is computational evidence for the idea that learning about the distributional semantic properties of words and predicting upcoming words in a sentence are partially at odds with each other, and that intelligent systems might be better off dividing the labor among two functionally distinct sub systems — one dedicated to acquisition and storage of distributional lexical semantic properties of individual words, and another specializing in prediction-based processing of word sequences.

Chapter 2

The AO-CHILDES corpus

In this chapter, I introduce the AO-CHILDES corpus, a custom compiled and pre-processed collection of transcripts of child-directed American English purposefully created for the work presented in this thesis. This corpus plays a central role in this thesis, not only because it is a psychologically plausible and developmentally realistic sample of the kind of language children are exposed to, but because it exhibits longitudinal structure that appear to be well suited to the construction of form-based lexical semantic knowledge in the recurrent neural network.

I will use this corpus in two experiments presented in this thesis: In Chapter 3, I demonstrate that the simple RNN captures lexical statistics that are diagnostic of semantic category membership in the network's contextualized representations at the hidden layer. Second, in Chapters 6 and 10, I demonstrate that the network also captures some of that information at a lower level in the network, where lexical information is statically stored, and where it can be more readily accessed for use in downstream tasks (e.g. DECAF).

The age-ordering of transcripts in AO-CHILDES presents an opportunity to analyze how basic statistical properties of input to children changes as they grow older. Specifically, in this chapter, and much of the remainder of this thesis, I explore the possibility that there are special properties in transcribed speech to younger children that facilitate the learning of statistical cues that can support the grouping of nouns into semantic category clusters compared to input to older children. I begin this investigation by documenting diachronic changes in the surface-level properties of input to children. First, I provide a brief overview of previous findings regarding unique properties of child-directed speech, with an emphasis on properties evident in transcribed data as opposed to at the acoustic or phonological level. Next, I introduce the AO-CHILDES corpus, and confirm it has the same longitudinal properties that have been observed in previous studies. Of particular interest are age-related changes in the statistical properties of lexical co-occurrence data across developmental time. I investigate such changes both globally, considering all words in the corpus, and in a more targeted manner, by restricting a subset of analyses to frequent nouns in Chapter 7. I did not measure changes that require linguistic knowledge, such as the number of discontinuities, embedded clauses, or inversions, or other factors that require construction of parse trees, given that the RNN has no a priori knowledge of English syntax. Most of the analyses reported in this chapter are tailored to properties that a distributional learning system such as the RNN is likely to be most influenced by.

I would like to note that this chapter is not meant to provide an exhaustive study of differences between child-directed and adult-directed language. My analyses are restricted to the textual domain, thereby eliminating my ability to detect age-related differences in, say, prosody, gestures, and referential context. Most of the beneficial properties of child-directed input on language acquisition are related to prosody (Fernald et al., 1989) and are thus not captured by the text-based representation of AO-CHILDES. Enhanced prosodic cues in child-directed input are known to facilitate children's vowel discrimination (Trainor & Desjardins, 2002), word recognition (Singh et al., 2009), and speech segmentation (Nelson et al., 1989; Thiessen et al., 2005). Moreover, unique social factors associated with child-directed input but not adult-directed language have been shown to influence language learning (Ramirez-Esparza et al., 2014). These benefits are worth mentioning because they provide existence proofs that the quality of data children are exposed to early during development can influence learning outcomes at later ages.

To preempt any confusion relating to the term 'child-directed speech' or 'child-directed language', I use these terms to broadly refer to language that is directed to or overheard by children. I do not use the term to refer to speech that has been specifically adapted or simplified for (implicitly or explicitly) pedagogical purposes. To make this point very clear, I use 'child-directed' to refer to any speech that is directed at a child, rather than speech that has been simplified for the purpose of facilitating acquisition. AO-CHILDES contains a mixture of input, such as utterances produced by caregivers that heavily skew their use of language, utterances produced by caregivers that do not adapt their language, and over-heard speech between adults.

2.1 Data and Pre-Processing

The AO-CHILDES corpus was first described by P. A. Huebner and Willits (2021b). The reason for the name is to differentiate the corpus from other corpora based on the CHILDES database, and to indicate that it is age-ordered (AO is short for age-ordered). Detailed analyses of pre-processing can be found in P. A. Huebner and Willits (2021b). Briefly, raw transcripts in the American -English section of the CHILDES database were retrieved (MacWhinney, 2000; Sanchez et al., 2019). All transcripts of speech to children older than 6 years of age, and transcripts without age information were removed. The text was lower-cased, spell-corrected, and periods in questions incorrectly marked with periods were replaced with question marks. When used for analysis, text was tokenized by splitting on white spaces and contractions.[1] Finally, transcripts were ordered by the age of the target child.[2] After pre-processing, I obtained 3,251 transcripts containing 272,250 types, and 5,245,298 tokens. Considering that a typical working-class American child receives approximately 6.5 million words per year (Hart & Risley, 1995), the training corpus represents approximately one tenth of the amount of lexical input of an average 6-year-old child — that said, there are large individual differences largely predictable by socio-economic status. The code for building AO-CHILDES from raw data is available at https://github.com/UIUCLearningLanguageLab/AOCHILDES.

There are some caveats concerning the use of AO-CHILDES for training models that simulate aspects of children's language acquisition. First, the CHILDES database is not perfect as a representative sample of the full range of activities that parents participate in with their children or the variety of language used during those activities, but is instead a useful approximation. Indeed, the relatively constrained set of activities that occur in CHILDES ought to hinder learning of a diverse range of lexical associations in model trained on AO-CHILDES, and thus make positive results all the more impressive. Second, AO-CHILDES is comprised of speech from many hundreds of speakers, and is thereby not ideal for studying diachronic trends

[1] However, when used for training, the corpus was tokenized using the Byte-Level Byte-Pair Encoding proposed by Sennrich et al. (2016).

[2] Transcripts associated with the same age were ordered randomly amongst themselves.

in the co-occurrence structure of input to a single child. Thus, all analyses based on AO-CHILDES likely underestimate the extent of developmental shifts in the lexical co-occurrence structure in children's input. Third, different transcriptions of the same word were left intact, and as such were treated as completely separate words. Thus the corpus is noisier than the input to a single child who is exposed to more homogeneous speech from a much smaller number of adults. Using this corpus as input, a distributional learning model is presented a greater challenge than that faced by an individual child.[3]

A question that is frequently asked in the context of child-directed input and the early learning environment of children is what factors drive the changes in linguistic factors of caregiver's language? Are caregivers modifying their language output based on age, as is assumed here, or the linguistic competence of the target child? If the latter were the case, this would put into question the utility of a corpus like AO-CHILDES where age is the criterion used to order transcripts. The corpus would be of little use for studying developmental changes as such changes would not depend on age, but linguistic competence. Fortunately, this is not the case; for instance, Newport et al. (1977) found that many features of caregiver speech change in accordance with the child's age, and not linguistic competence.

2.2 Input to children starts small

Previous investigations of how language changes across developmental time has focused on binary comparisons between child-directed and adult-directed speech. While this is no doubt useful, such analyses have left unanswered how language to children changes as they grow older. Classic studies comparing child-directed to adult-directed speech have uncovered differences along multiple dimensions, such as larger pitch contours, lengthened vowels (Fernald & Kuhl, 1987; L. R. Gleitman et al., 1984) and restricted use of complex constructions (Broen, 1972; Furrow et al., 1979; Pine, 1994; Richards, 1994). Caretakers producing child-directed input, are also more likely to restrict the range of conversational topics, choice of words and grammatical abstractions (Lieven, 1994; Snow & Ferguson, 1977), and make longer pauses between utterance boundaries (Gallaway & Richards, 1994). A key finding in this literature is that speech to children is less lexically diverse compared to adult speech (Kirchhoff & Schimmel, 2005). Moreover, Foushee et al. (2016) found that lexical diversity gradually increases over the first three years of life before merging with adult-level lexical diversity soon after. This means that a gradient in lexical diversity should also be present in AO-CHILDES. Similarly, it has been known for a long time that mean-length-of-utterance (MLU) is smaller in speech to younger children and increases with age of the target child (Broen, 1972; Fernald & Morikawa, 1993; Snow, 1972). Another important difference between child-directed input and adult speech is that nouns occur more frequently and tend to refer to more concrete objects (Tardif et al., 1997). Interestingly, novel nouns, and other words a child is not yet familiar with, tend to be preferentially introduced in utterance-final positions, which appear to be especially salient to young children, and promote the best learning outcomes (Fernald et al., 2010). While numerous benefits of child-directed input have been found on early language acquisition (Golinkoff & Alioto, 1995), others have shown that children in some cultures appear to learn language just as well when their primary caregivers do not adapt their speech when talking to children (Schieffelin & Ochs, 1986). Whether age-related changes in any of these factors can benefit semantic category learning from distributional information has not previously been investigated.

[3] Though, a child must content with enormous variation in the acoustic signal, unlike distributional semantic models which start with representations based on orthographic identity.

2.3 Lexical versus Combinatorial Diversity

An overarching goal of this chapter, and others, is to identify lexical statistics that might make it more or less difficult to discover grammatical and/or semantic categories via distributional analysis, and how these factors change with age. I distinguish between two classes of statistical phenomena in language that can influence the construction of form-based lexical categories: Lexical and combinatorial diversity. I will argue that an increase in one or the other can, independently, impede the discoverability of category-relevant distributional information. To illustrate, assume a distributional learner has encountered one of the following utterances:

(a) Do you want some X ?

(b) Do you want some more X ?

(c) Do you want some additional X ?

Assume that the left-context of the target word X, '*Do you want some*', perfectly predicts the occurrence of related words that belong to the semantic category DRINK. Utterance (b), however, contains an additional word between the semantically informative context and the target word, and this reduces the predictive utility of the otherwise perfectly predictive context. Utterance (b) exemplifies what a distributional learner might encounter in a corpus with higher combinatorial diversity relative to (a). In such a corpus, predictive contexts occur in more variable positions — sometimes closer and at other times farther away — relative to a target word, making them less reliable and more difficult to discover. The same is true of utterance (c), except that something slightly different is going on here. I would target that utterance (c) is an example of an increase in lexical, as opposed to combinatorial, diversity relative to the base utterance (a). The reason is that the newly introduced word, *additional*, occurs in the same syntactic position in which *more* occurs. While the mechanism by diversity has increased in utterances (b) and (c), the results is the same: The search for language-internal cues that are diagnostic of semantic category membership has become more difficult. This happens because greater diversity reduces the likelihood that a systematic association between a category-relevant context and a target word will be found. If utterance (b) and (c) both occur in the training data, then the strengths of the association between *more* with a target word's membership in DRINKS, and *additional* with a target word's membership in DRINKS are weakened by the presence of the other. While each context is equally informative about a target word's category membership, the fact that each context is half as frequent compared to a unified cue makes the category-informative signal more difficult to acquire.

2.4 Lexical and Noun Phrase Diversity

I computed the lexical diversity of 15 consecutive (age-ordered) partitions of AO-CHILDES to confirm that lexical diversity increases with age of the target child. While age-related increase in lexical diversity have previously been observed (Foushee et al., 2016; Jiang et al., 2020a; Kirchhoff & Schimmel, 2005), I wanted to confirmed that this holds true in the corpus used in this thesis. To compute lexical diversity, I opted for a procedure that takes into consideration various confounds that may arise when computing simple type-token ratios, called MTLD (McCarthy, 2005). Further, I examined the diversity of base noun phrases across corpus partitions. I did this by treating each noun phrase as a single unit, and counting the number of unique noun phrases that occur in each partition. Considering age-related increase in fragmentation of the noun category, I predicted an increase in noun phrase diversity with age. To isolate base noun phrases (base NPs), I used the python library *spacy* v2.3.7, which defines a base NP as a phrase that does not permit other NPs to be nested within it. This definition excludes NP-level coordination, prepositional phrases, and relative clauses.

The results are shown in Figure 2.1. The left panel shows that lexical diversity steadily increases with age (i.e. age-ordered corpus partitions), as reported by previous studies. The right panel shows that the diversity of base NPs is relatively stable across partitions, but is much lower in the first two partitions, which contain input to the youngest children in AO-CHILDES. This observation was the starting point for much of the work presented in this thesis, and the inspiration for many of the ideas proposed and tested in this work. Specifically, this observation is compatible with the idea that distributional learners exposed to input to younger children need to integrate over fewer unique stimuli to form a proto-noun category compared to learners exposed to input to older children.

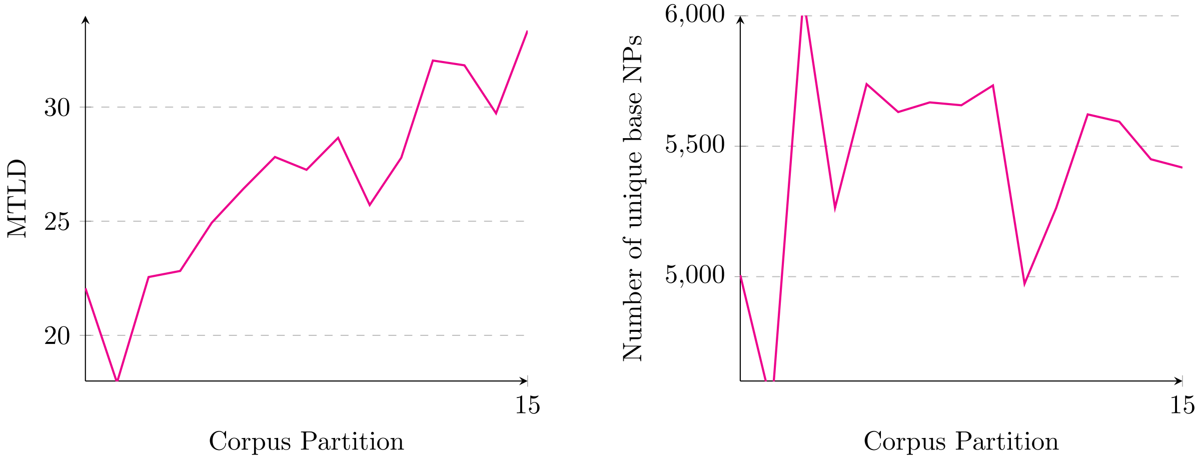


Figure 2.1: Lexical and NP diversity across partitions of AO-CHILDES. The left panel shows MTLD (a measure of lexical diversity), and the right panels shows the number of unique base NPs (a measure of NP diversity).

2.5 Combinatorial Diversity

2.5.1 Mean Utterance Length and Function Word Density

To confirm that combinatorial diversity increases with age of the target child, I quantified the mean utterance length (MLU) across 15 consecutive partition of AO-CHILDES. I predicted a steady age-related increase, as has been documented previously using different data (Newport et al., 1977). Further, I predicted that an increase in MLU is associated with an increase in the syntactic complexity of input, such as increased usage of function words to form modifying prepositional phrases (requiring prepositions such as *on*, *over*, *by*, *etc.*), clausal conjunction (requiring the function words *and*, *because*, *but*, etc.), and complex clausal complements (requiring complementizers *that*, *if*, etc.). I tested this hypothesis by quantifying the density of function words across corpus partitions[4].

The results are shown in Figure 2.2. While there is a considerable amount of variance, there is a noticeable and gradual upward trend in MLU (left panel) from approximately 5 words per sentence between ages

[4] For the complete list of function words, see https://semanticsimilarity.wordpress.com/function-word-lists/

1-3 years to just short of 8 words/sentence between the ages of 3-6 years. This is a clear indicator of an expansion in combinatorial diversity in input to children across developmental time. The right panel shows a gradual increase in function word density across all partitions in AO-CHILDES and confirms the hypothesis that sentences uttered to children become increasingly more complex by making greater use of syntactic operations such as modification, adjunction, conjunction, embedding. Distinguishing these different types of operations is certainly interesting but is not directly relevant to this work. For the purpose of this thesis, it makes little difference *how* or *why* distributional signals diversify with age; my aim is to take the first step towards demonstrating that this trend exist in the first place, and that it has downstream consequences for distributional learning (and by extension, DECAF).

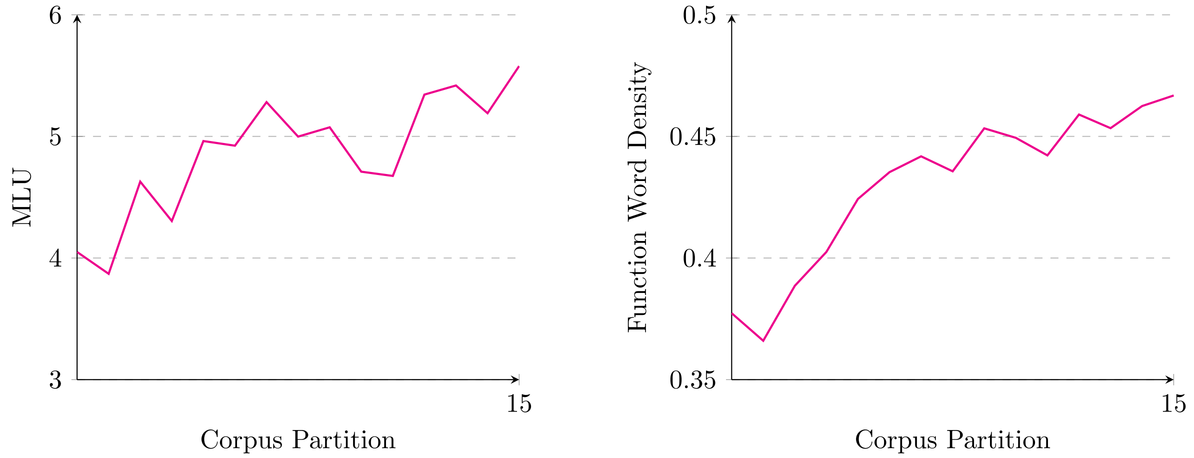


Figure 2.2: Mean length of utterance (MLU) and function word density across partition of AO-CHILDES.

That said, to provide a glimpse at what types of words contribute most to the growth in MLU, I conducted a follow-up analysis that tracks the density of the major part-of-speech (POS) classes across partitions of AO-CHILDES. To do so, I first divided the corpus into 64[5] equally sized consecutive partitions, and then counted the number of words from each POS class that occur in each. Next, for each POS category, I correlated the vector of frequencies (the number of vector elements corresponds to the number of partitions) with a vector containing the partition numbers of the age-ordered corpus (e.g. 1, 2, 3, ..., 64).[6] The results are shown in Table 2.1. I found that the density of all POS classes except for interjections and adjectives are significantly correlated with corpus partition (a proxy for age).

Interestingly, only nouns and determiners are negatively correlated with partition number, meaning that they are the only POS classes whose members significantly decrease in frequency across the age-ordered corpus. Given that density is defined as relative frequency, an increase in one or more POS classes with age requires a decrease in one or more other classes (not unlike resource allocation in a zero-sum game). It is not surprising therefore that as other POS classes become more prevalent, nouns and determiners are the classes that make room to enable their expansion. It is likely that the decrease in determiners simply tracks the decrease in nouns with age. Importantly, however, this says nothing about a potential distributional shift within the class of determiners across age. In fact, the class of determiners is relatively rich, and includes, among

[5] The number of partitions was increased from 15 to 64 in order to compute reliable statistics.

[6] Spearman's rank correlation was used.

Part-of-Speech	Spearman's Rho	p-value
ADJECTIVE	0.1950	0.1226
ADPOSITION	0.8334	0.0000
ADVERB	0.8373	0.0000
DETERMINER	-0.6229	0.0000
INTERJECTION	-0.0523	0.6817
NOUN	-0.8617	0.0000
PRONOUN	0.8544	0.0000
VERB	0.7920	0.0000

Table 2.1: Rank-Correlation between Token Frequency and Corpus Partition

others, articles (e.g. *a*, *the*), demonstratives (e.g. *these*, *those*), possessives (e.g. *his*, *their*), and quantifiers (e.g. *some*, *any*). It is therefore possible that there is an internal redistribution, from, say, semantically vacuous articles to the more semantically informative quantifiers, which can be used to discriminate among mass and count nouns, or pronouns. Indeed, I found that the determiner *the*, which provides virtually no information about semantic category membership, is in the top-5 words with the greatest decrease in density across age-ordered partitions of AO-CHILDES.[7] Given that the words whose density increased the most across AO-CHILDES are possessive pronouns (e.g. *her*, *my*), it is likely that there is an age-related shift from articles to pronouns in the pre-nominal position. This non-stationary aspect of the data would make it difficult for a distributional learner to form a stable representation of the noun category, and therefore could negatively impact the formation of lexical semantic representations of nouns.

A graphical depiction of the frequency and density (relative frequency) of POS tags in AO-CHILDES is shown in Figure 2.3. The left panel shows that most transcripts in AO-CHILDES contain transcribed speech to children around their second birthday. The right panel shows the same distribution in terms of

[7] The item with the greatest decrease in density is the punctuation symbols which frequently occur after nouns in input to children.

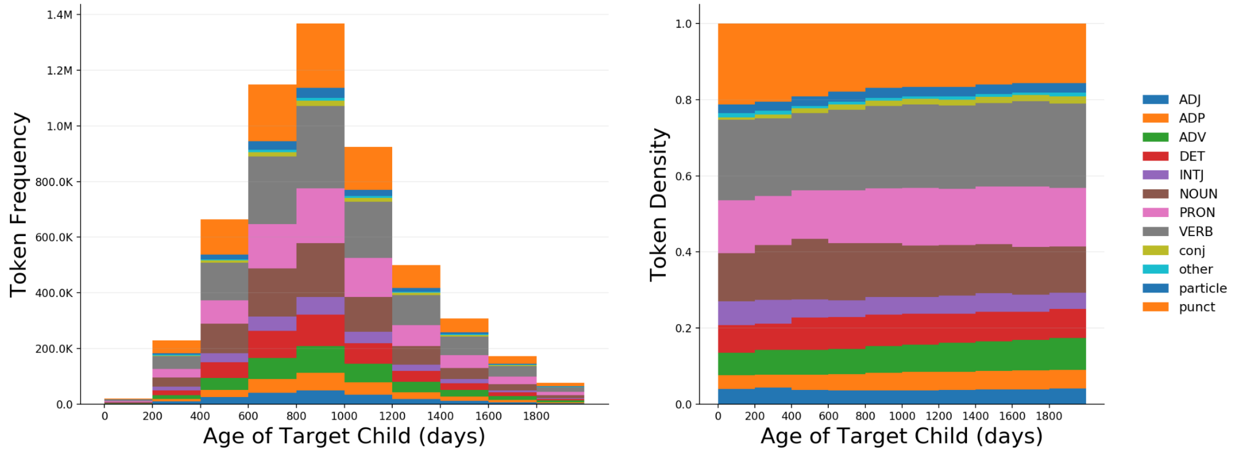


Figure 2.3: The distribution of POS tags over consecutive age bins in AO-CHILDES. The bars in the left panel represent token frequency, while the bars in the right panel represent density (frequency normalized by bin size). Legend entries labels strata in reverse order (i.e. the top-most entry labels the bottom-most stratum).

N-gram size	partition 1	partition 2
1	142.75	130.89
2	11.51	10.46
3	4.01	3.81
4	2.66	2.57
5	2.25	2.21
6	2.10	2.08
7	2.05	2.03

Table 2.2: Average number of times an n-gram is repeated.

relative frequency to make age-related trends in the density of POS tags more visible. One of the most clearly visible trends is the decrease in the density of punctuation tokens (top-most stratum in orange). Further, in accordance with the analysis above, the figure clearly demonstrates that the density of conjunctions, adverbs, and pronouns increases gradually with age.

2.5.2 Repetition of n-grams

A hallmark of less complex language input is repetition, both at the single-word and multi-word level. To quantify the amount of repetition in a given partition of AO-CHILDES, I computed the average number of times an n-gram of a particular size is reused in the partition. The results are shown in Table 2.2. Not surprisingly, the average number of times a sequence is repeated is inversely proportional to the length of the sequence. More importantly, the number is consistently larger for partition 1, at all n-gram sizes (1-7) examined.

2.5.3 Taylor Exponent

A recently introduced measure of structural complexity of linguistic sequences is the Taylor exponent (Tanaka-Ishii & Kobayashi, 2018). It is the exponent in a power-law relationship between the variance of word frequency and the average word frequency per unit of time. This relationship, known as Taylor's law, was first discovered in ecology where the variance of the number of individuals of a species per unit area is related to the average number of individuals per unit area according to a power law. Taylor analysis has since been applied in numerous other fields (Eisler et al., 2008) to demonstrate systematic relationships between events. Theoretically, an independent and identically distributed (iid) process must have a Taylor exponent of 0.5, and larger exponents indicate processes in which events depend on each other. Human linguistic sequences exhibit a Taylor exponent above 0.5 (Tanaka-Ishii & Kobayashi, 2018), indicating that words co-occur systematically. Moreover, child-directed input is characterized by a larger Taylor exponent than adult speech, which suggests that child-directed input is more template-like. Because I am interested in category learning from distributional information, I asked whether speech to younger children is more systematic, and less structurally complex than speech to older children.

I computed the Taylor exponent separately for partition 1 and 2 of AO-CHILDES using the same method used by Tanaka-Ishii and Kobayashi (2018). First, I split each partition into chunks of 5,600 words and computed the frequency of all words in each chunk. Then I obtained the average and standard deviation of the frequency of each word across the chunks and fitted the resulting data to a linear function in log-log coordinates by the least-squares method. The best-fit line represents the relationship between the standard

AO-CHILDES		Newsela				
Partition 1	Partition 2	Level 5	Level 4	Level 3	Level 2	Level 1
0.635	0.614	0.604	0.584	0.587	0.591	0.598

Table 2.3: Taylor Exponent computed on AO-CHILDES partition 1 and 2, and sections of Newsela grouped by simplification level. Level 5 corresponds to the highest level of simplification.

deviation and mean of each word's frequency, and the Taylor exponent, represents the slope of the line in log-log coordinates.

To make comparison between the two partitions of AO-CHILDES more interpretable, I also computed the Taylor exponent for texts with known reading difficulty. I chose the Newsela corpus [8] for this purpose. It includes 1, 911 news articles written in English, and for each article there exist 4 or 5 simplified versions, rewritten by professional annotators for children with different reading proficiency. For each simplification level 1-5 (mapping on to grades 2 through 12), I obtained approximately 1M words, and computed the Taylor exponent in the same manner as explained above.

The results are shown in Table 2.3. The Taylor exponent associated with partition 1 (0.635) is larger than the Taylor exponent associated with partition 2 (0.614). This indicates that partition 1 contains a greater number of word sequences with fixed forms compared to partition 2. As mentioned before, a Taylor exponent calculated for an *iid* process is 0.5, and the larger the value the more template-like it is. This finding is consistent with that of Tanaka-Ishii and Kobayashi (2018) who found that child-directed input speech (in addition to programming languages and music) was found to have a larger Taylor exponent than adult language. Moreover, the Taylor exponent behaves as predicted for texts with different reading difficulty. As texts become more simplified (from low to high simplification level), the Taylor exponent becomes smaller. If one were to align the two series of Taylor exponents, beginning with partition 1 of AO-CHILDES, and ending in Newsela texts with simplification level of 1 (representing the most difficult texts), one would observe an ordered sequence of numbers from highest to lowest. This result lends strong support for the validity and utility of the Taylor exponent as a tool for assessing language complexity.

2.5.4 N-gram Model Perplexity

No analysis of combinatorial diversity would be complete without evaluating the fit of an n-gram language model on a corpus. Specifically, I split AO-CHILDES into 2 equal sized partitions and trained Kneyser-Ney language models of varying n-gram sizes (3 to 6) separately on each partition. I performed this analysis twice. In one analysis, I trained n-gram models on input where any out-of-vocabulary word had been replaced by an out-of-vocabulary symbol. I repeated the analysis with the full vocabulary (all words left intact). N-gram language models were trained with the KenLM Language Model Toolkit (Heafield et al., 2013) and scored using the python module *kenlm*. The results are shown in Figure 2.4. Using the reduced vocabulary (all but the 4,096 most frequent words), average perplexity scores for all three n-gram sizes were smaller when trained and evaluated on partition 1 compared to partition 2 (4-grams: 9.1 vs 9.6, 5-grams: 5.4 vs 5.9, 6-grams: 4.2 vs 4.6). The same pattern is observed when training n-gram language models on partitions with the full vocabulary (4-grams: 8.8 vs 9.4, 5-grams: 5.4 vs 5.8, 6-grams: 4.2 vs 4.6). Perplexity is a measure of

[8] Newsela. (2016). Newsela Article Corpus. Version: 2016-01-29, https://newsela.com/data.

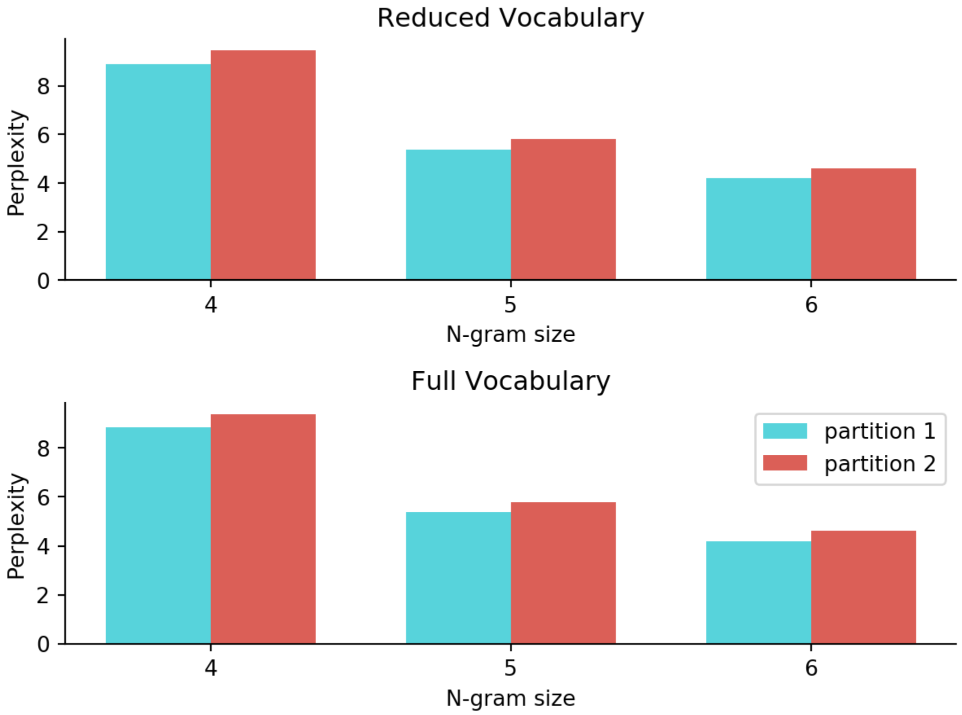


Figure 2.4: Perplexity of n-gram models trained on AO-CHILDES partition 1 (blue) or 2 (red). Lower perplexity indicates the model makes more accurate next-word predictions.

sequence prediction error. As such, a lower perplexity indicates that it is easier to predict the next word given the words that come before it. Perplexity can also be viewed as a measure of how unlikely a model judges a sequence of words, given the knowledge it has accumulated during training. All six n-gram models judge sequences in partition 2 to be less likely than an equivalent n-gram model trained on partition 1. Again, this confirms that combinatorial diversity is smaller in transcribed speech to younger compared to older English-learning children.

2.5.5 Number of n-gram types

Another crude method for quantifying combinatorial diversity is the number of unique n-grams. A greater number of unique n-grams can indicate a number of phenomena, including that a speaker is sampling words from a larger vocabulary, uses longer utterances (e.g. multi-clause utterances, multiple noun phrases, more frequent use of prepositional phrases, insertion of adjectives and/or adverbs), or makes more frequent use of infrequent or alternate constructions. It could also indicate reduced semantic richness, reduced lexical diversity, or perhaps conversation about a more restricted set of topics. This measure is clearly not quantifying a single dimension of the input, and cannot be said to be purely a reflection of combinatorial diversity. The first step of this analysis involved counting the number of unique n-grams per partition, and in the complete corpus. I included n-grams of size 1 through 7 to maximize the likelihood of detecting any differences between speech to younger and speech to older children. Next, I conducted two analyses: In the first, I asked what percentage of all n-gram types occur in partition 1 and what percentage occur in partition 2. Put differently, how representative is the set of unique n-grams in each partition of the complete inventory of n-grams in

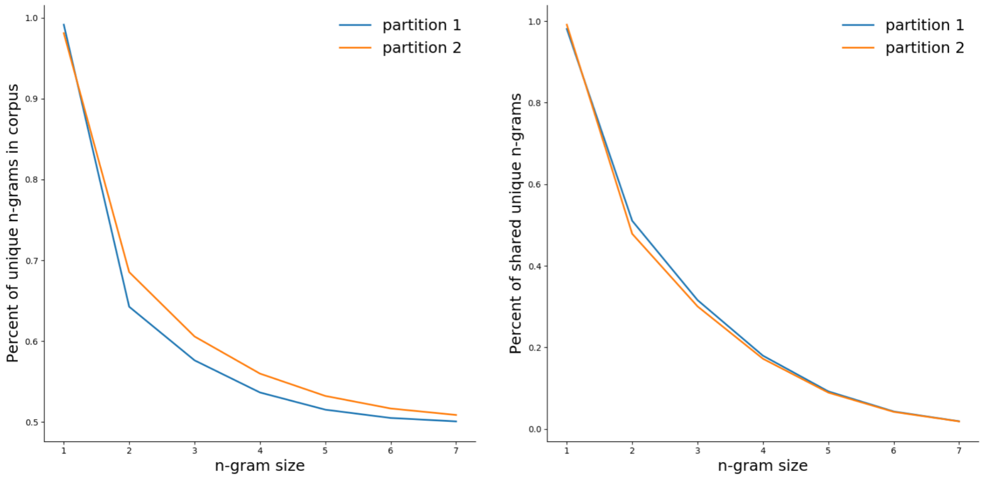


Figure 2.5: Quantifying n-grams in AO-CHILDES. Left panel: Percent of n-gram types in partition 1 (blue) and partition 2 (orange) of the n-grams in the complete corpus. Right panel: Percent of n-gram types in partition 1 that also occur in partition 2 (blue); percent of n-gram types in partition 2 that also occur in partition 1 (orange line).

the corpus? In a second, related analysis, I asked what percentage of n-grams in a partition are also present in the other partition. The usage of more high-frequency, canonical constructions will result in a higher percentage, as they are more likely to re-occur in speech to children at all ages.

The results of the first analysis are shown in the left panel of Figure 2.5. I found that partition 2 contains a greater proportion of the total number of n-gram types (of size 1 through 7) compared to partition 1. For example, nearly 72% of the total number of 2-gram types can be found in partition 2, while only 66% can be found in partition 1. The results of the second analysis are shown in the right panel of Figure 2.5. Partition 1 consistently scores higher than partition 1 for all n-gram sizes evaluated, but the difference drops off with n-gram size. In interpret this result as follows: Word-sequences that occur in input to younger children are more likely to re-occur in input to older children, compared to the other way around.

2.6 Summary

In this chapter, I introduced the AO-CHILDES corpus, which will be used extensively in this thesis as a representative sample of language to English-learning children below the age of 6. Importantly, the corpus has been ordered by the age of the target child, which opens the door to researchers interested in studying the diachronic structure of transcribed speech to children across developmental time. I have conducted such preliminary analyses, focusing on surface-level lexical statistics that a distributional learning system such as the RNN is likely sensitive to. I distinguish between two distinct statistical phenomena, lexical and combinatorial diversity, each of which may, independently, contribute to age-related changes in language statistics across developmental time.

The age-related changes in input to children that were identified in this chapter are the starting point for many of the questions asked in this thesis. Many of them are picked up in Chapters 6 and, in particular 7, 8, and 9. Such diachronic changes are of interest because they may make it more difficult for a distributional learner (artificial or not) to identify structural commonalities across sentences, such as the syntactic relation between the main clause subject and verb. As the diversity of multi-word combinations increases (e.g. greater use of embedding, and modifiers), and sentences grow longer to communicate richer messages, the deeper structural commonalities between sentences are likely to become increasingly opaque to statistical learning systems. If a distributional learning system were trained on input where important relations are buried beneath long-distance dependencies across complex intervening spans, the system might struggle to isolate relations that are most useful for learning about semantic category membership of nouns. I test this prediction explicitly in chapter 10.

Chapter 3

Next-Word Prediction Constructs Lexical Semantic Category Knowledge

In this chapter, I advocate for a graded, open-ended approach to modeling children's acquisition of form-based lexical semantic category knowledge.[1] Toward this end, I empirically demonstrate the feasibility of learning semantic category knowledge in the simple RNN trained on a corpus of child-directed language. More broadly, my aim in this chapter is to establish the RNN as a potential model of how children discover and represent fine-grained statistical regularities among words as a stepping stone to performing the distributionally mediated extension of category-associated features (DECAF). Many parts of this chapter have been previously published in a peer-reviewed journal (P. A. Huebner & Willits, 2018).

3.1 Drawbacks of Rule-based Information Extraction

First, I would like to motivate the need for an open-ended system that can capture arbitrary dependencies such as the RNN. I do this by illustrating the limitations of a more traditional and less open-ended method for discovering semantic category knowledge from corpus data. For example, in traditional and non-distributional information extraction, the semantic relations to be extracted are pre-defined by abstract patterns created by domain experts. While the advantage of this approach is that learning can be reduced to a search problem, such algorithms are notoriously brittle. The burden is on the creators of such algorithms to identify useful patterns that work well in a variety of situations. Even when patterns are identified by domain-experts, they often do not generalize well outside the conditions for which they were developed. To illustrate this in the domain of language, consider a traditional information-extraction algorithm tasked with identifying semantic-categories in a sample of natural language. One solution that has been proposed is to search for 'Hearst patterns' which identify hypernym-hyponym (e.g. *dog*, *animal*) pairs in raw text. All five patterns are shown in Table 3.1.

To illustrate how well this rule-based search for Hearst-patterns would work when applied to transcribed speech to children, I identified all Hearst-pattern matches in AO-CHILDES, and reported them in Table 3.2. The results are not promising; there are 5M total words in the corpus, and the rule-based algorithm

[1] In this thesis, I speak exclusively of form-based (i.e. distributional) semantic knowledge, except when otherwise noted. For brevity, I will refer to 'form-based semantic knowledge' as simply 'semantic knowledge'.

Pattern	Example
X and other Y	churches, and other building.
X or other Y	dogs, cats, or other pets
Y such as X	a vehicle, such as a plane
such Y as X	such furniture as tables, sofas
Y including X	electronics, including TVs

Table 3.1: Hearst patterns used in traditional rule-based approaches for extracting hypernym-hyponym pairs.

only identified 20 patterns. These 20 patterns are an extremely impoverished window into semantic category structure implicit in child-directed input. Of the 5 possible patterns, only 3 patterns were matched in the corpus. Further, the most frequently matched pattern (X and other Y) mostly returned matches that are not useful for diagnosing semantic category membership; the phrases '*and other things*', and '*and other objects*' are nearly semantically vacuous.

ribbons and other things
animals and other toys
books and other things
scooters and other things
bicycles and other things
children and other mommies
scooters and other things
food and other things
pebbles and other treasures
names and other things
plants and other objects
balls and other stuff
straw and other things
foods and other people

(a) X and other Y

foods such as flour
crops such as wheat

(b) X such as Y

products including cheese

(c) X including Y

Table 3.2: All hypernym-hyponym pairs extracted from a corpus 5M words of American-English child-directed input using Hearst-Pattern matching.

3.2 The Simple Recurrent Neural Network

As an alternative to rule-based approaches to extracting lexical semantic category knowledge from natural language, I consider the simple Recurrent Neural Network (simple RNN), a connectionist model that has been widely used by the psycho-linguistic community for modeling the sequential structure of language (J. L. Elman, 1990, 1991). A schematic of the architecture, used in a language modeling task, is shown in Figure 3.1.

Learning in the RNN is based on minimizing the error between the predicted and actual next item in the corpus, given the model's encoding of previous items. Its memory of previously seen items is encoded into a fixed-size low-dimensional vector, and is typically referred to as the activations at the model's hidden layer. This pattern of activation, h_t, is defined as:

$$h_t = f(h_{t-1}, x_t), \tag{3.1}$$

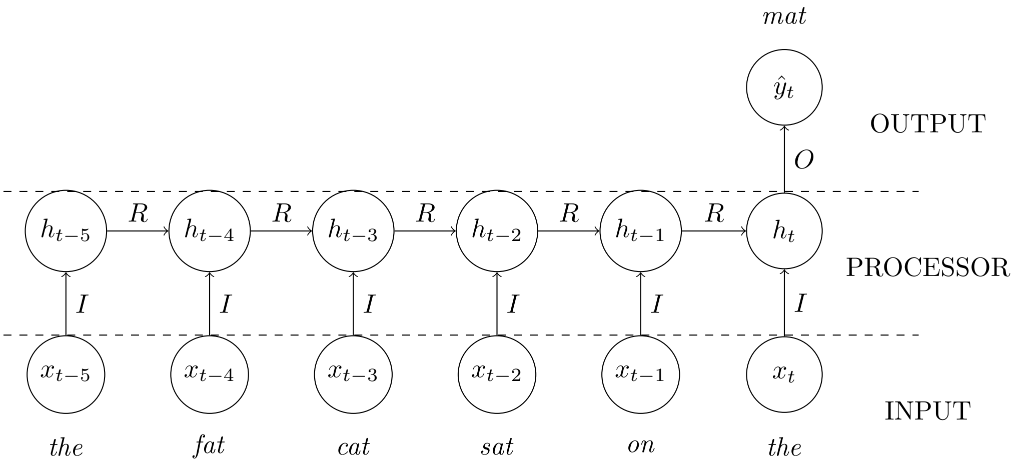


Figure 3.1: A schematic illustrating the architecture of the simple RNN. Each circle is a multi-dimensional activation state. Circles separated by horizontal arrows are activation states separated by time steps. While the input vectors, x, use a localist code, all hidden activation states, h, are dense and distributed. The localist vectors, one for each token in the vocabulary, are first processed by the weight matrix I, the input-to-hidden weights. The dot product of I and the localist input vector is equivalent to looking up a word's low-dimensional lexical semantic representations in a dictionary-like word embedding space. R labels the recurrent weight matrix, and O labels the hidden-to-output weights. The same weight matrices are reused across time steps, and updated simultaneously. $\hat{y}_t$ is the predicted probability distribution of next-words, which is the size of the model's vocabulary.

where h_{t-1}, is the value of the hidden layer at time $t-1$, x_t is the input feature vector at time t, and $f(.)$ is a nonlinear function (e.g. tanh or sigmoid). There are three sets of weights (i.e. parameters that are adjusted during training): The connections between the input and hidden layer, the connections between the hidden and output layer, and the connections from the previous hidden layer state to the current hidden layer state. The latter set of weights is also called the 'transition function', or the 'recurrent weights'. Each set of weights is randomly initialized at the start of training and updated using Stochastic gradient Descent (SGD) or some variant thereof. Importantly, unlike feed-forward networks, the error is back-propagated across time steps, and used to update all the weights that contributed to a prediction (Rumelhart et al., 1986; Werbos, 1990). The simple RNN is only one instance of a large class of recurrent networks (RNNs), that vary in the number of layers, activation functions, gating units, self-attention, and other machine-learning tricks (see e.g. Mikolov et al., 2011, Dyer et al., 2016). In this work, I will often drop the prefix 'simple' when referring to the simple RNN, given that many of my findings and ideas are relevant to the RNN class, rather than to the simple RNN only.

In language acquisition research, the RNN has been primarily used to examine what kind of linguistic dependencies can be captured by the network and under what circumstances (e.g. size of training data, distance between dependent items), and whether the resultant knowledge generalizes to unseen examples (G. T. M. Altmann & Mirković, 2009; Dienes et al., 1999; Linzen & Jaeger, 2016; D. L. Rohde & Plaut, 1999). Early studies using artificial grammars and pseudo-English showed that the network learns to predict the next word probability distribution reasonably well, even in situations requiring it to learn long-distance dependencies across potentially uninformative sub-sequences, such as embedded clauses (J. L. Elman, 1990, 1991; Servan-Schreiber et al., 1991). In addition to evaluating the network's success at the next word prediction task, the network's hidden layer activations have been studied. Such analyses have shown that activation patterns over hidden units capture the syntactic and semantic structure implicit in the training corpus (Christiansen & Chater, 1999b; J. L. Elman, 1990, 1991; D. L. Rohde & Plaut, 1999; Tabor et al., 1997b). For these reasons and others, the RNN has become the model of choice for researchers who consider predictive processing as fundamental to language comprehension and production (G. T. M. Altmann & Kamide, 1999; Dell & Chang, 2014; J. L. Elman & McRae, 2019; Hubbard et al., 2019; Huettig et al., 2022; Kukona et al., 2011; Kuperberg & Jaeger, 2016; Linzen & Jaeger, 2016; Pickering & Garrod, 2007; Rabovsky et al., 2018).

3.2.1 Learning Lexical Semantic Category-Membership in the RNN

In this work, I will examine whether the RNN can be pushed beyond the purpose for which it was originally developed. Apart from capturing sequential dependencies among linguistic units in natural language, does the RNN also learn lexical semantic category knowledge in a format that can be readily extracted from the network? Because that this question not been investigated before, I briefly review what is known about lexical category learning in the RNN. What criteria must be met for category learning to take place in the RNN, and how can we identify whether category learning has actually taken place?

I start by considering the statistical properties of the input. A set of words can be said to belong to a category when there exists a signal in the training corpus that systematically differentiates same-category members from all other words in the network's vocabulary, but not from each other. But what exactly does such a signal look like? A signal useful for lexical categorisation must meet two criteria. The signal must result in lexical representations:

1. of different-category members to move farther apart in the representational space, and
2. of same-category members to move closer together in the representational space.[2]

For example, determiners are a reasonably diagnostic signal for the noun category in English. Because determiners are more likely to occur in front of nouns compared to non-nouns, their distributional signal satisfies both criteria. Because they are more likely to occur before nouns, the lexical representations of nouns move closer together, and this satisfies criterion 2. Second, because determiners occur less frequently before non-nouns, the lexical representations of non-nouns are pushed farther away from those representing nouns, and this satisfies criterion 1. However, determiners are not perfect in this regard because they also often precede adjectives in English. I do not consider this to violate criterion 1, because determiners are still useful for discriminating nouns and non-adjectives. This shows that no single signal is by itself perfectly diagnostic; even the best distributional signal typically only provides fuzzy evidence. In order to learn sharp category boundaries around clusters of same-category members, a network must integrate across a large number of fuzzy distributional signals. Integration requires an abstraction mechanism, which is made possible by the compression of features at the RNN's hidden layer.

Contrary to syntactic categories (e.g. noun, verb), lexical semantic categories (e.g. MAMMAL vs. FISH) are based on much finer grained distributional evidence, typically involving longer distance dependencies, and more complex (e.g. higher order), and less frequently observed relations. For example, when talking about a dog, a sentence almost always provides a diagnostic cue about the grammatical category of *dog*,

[2] Representational space can refer to any learned parameter (i.e. connection weight) space in the network. In this chapter, I consider the hidden layer where semantic information is accumulated in a contextualized manner. However, in subsequent chapters, I examine the input-to-hidden weights (i.e. embedding matrix).

but need not include much semantic information (e.g. is_a ANIMATE, is_a MAMMAL) that would allow a listener to infer the semantic category of *dog*. For example, the sentence *I often do that with my dog.* provides clear evidence for membership of *dog* in the noun category, but relatively impoverished semantic information — other than the notion that an activity can be performed with a dog. This example illustrates that the problem of learning lexical semantic category membership from distributional evidence alone is more involved, and will require additional care than those that are typically performed in studies of distributional learning of grammatical categories (Freudenthal et al., 2013; Keibel, 2005; Redington et al., 1998).

3.2.2 Separation and Integration

In order to develop a precise and theoretically grounded understanding of the learning dynamics underlying the construction of form-based semantic category knowledge in the RNN, I adopt terms from information theory, such as entropy and redundancy (see also Chapter 8). I will use the term 'information' in the information-theoretic sense, which enables quantitative work with lexical co-occurrence data and orients our thinking towards predictive uncertainty. An example of prior work that has examined category learning in the context of language acquisition from an information-theoretic perspective is that of Cassani et al. (2018) who showed that words are easier to categorize when learners are faced with some degree of predictive uncertainty. Specifically, the authors demonstrated that, while both contextual diversity and frequency positively affect category formation, the average conditional probability was negatively predictive of category learning. Cassani et al. (2018) state that

> ...categorization works best for words which are frequent, diverse, and hard to predict given the co-occurring contexts. This shows how, in order for the learner to see an opportunity to form a category, there needs to be a certain degree of uncertainty in the co-occurrence pattern.

This is an important insight because it means that category formation can not be reduced to a simple gathering of facts that distinguish the usage of a given word from another; rather lexical categories are useful precisely when there is little reason to separate concepts that perform similar linguistic and/or communicative functions. Moreover, this result confirms my earlier claim that successful category formation from distributional evidence not only requires separation between semantically different lexical items (criterion 1), but also integration of semantically similar lexical items (criterion 2).

When thinking about category formation, it is easy to forget the importance of criterion 2, integration. The driving force behind integration is within-category similarity, whereas the driving force behind criterion 1, separation, is between-category similarity. While the former tends to result in a tight clustering of same-category members in a representational space, the latter ensures that clusters are well separated from each other. Many algorithms in machine learning operate on the principle of finding decision boundaries between data points (e.g. lexical semantic representations) that belong to different categories. The same idea applies here, except that the RNN is not explicitly trained to perform clustering of lexical semantic representations. Any clustering that emerges at the RNN's hidden layer, does so as a by-product of (in the service of) tuning parameters that work well for next-word prediction. This means that the set of upcoming words that are predicted by a target word determine the target word's location in representational space. When the set of upcoming words is highly diverse and varies between same-category target words, the representations of same-category members will be driven apart from each other, resulting in a lower quality clustering. With the idiosyncratic nature of language statistics, it is likely that this teasing apart of same-category members (i.e. separation) can overpower the opposing force, integration, needed to pull same-category members back together into tight clusters. Category learning (operationalized as a clustering problem), therefore, requires careful tuning to prevent useful separation from morphing into over-fitting — an extremum of separation that results in dispersion of data points without concern for what makes them similar. This concern is an ongoing theme in this thesis.

3.3 Next-word Prediction Constructs Lexical Semantic Category Clusters from Child-directed Input

Previous work has shown the feasibility of using an RNN to model the sequential regularities of natural language using large-scale corpora as input (Jozefowicz et al., 2016; Merity et al., 2017; Mikolov, 2012). Such findings demonstrate the potential of the RNN to rival more formal approaches used in computational linguistics (e.g. n-grams for modeling sequential structure) for estimating the joint probability distribution over natural language sequences. In this thesis, I extend the use of RNN language modeling from sequence learning to the domain of learning lexical semantic representations from noisy, naturalistic child-directed speech. In particular, I trained the RNN to predict next-words in the AO-CHILDES corpus, described in Chapter 2. The goal is to test whether the RNN can learn to group common nouns into proto-semantic categories, when the only source of information is the language-internal lexical statistics of child-directed input.

3.3.1 Methods

Before reporting the results, I outline the steps taken to train and tune the RNN, and how learned representations were probed.

RNN Vocabulary

The RNN was trained using one of two vocabularies. The first consisted of the 4,096 most frequent words after white-space tokenization. All items not in the vocabulary where replaced by an 'UNKNOWN' symbol. Second, a custom sub-word vocabulary with 8,000 tokens was created using the Python package *tokenizers*. The construction of the vocabulary is based on the Byte-Pair Encoding introduced by Sennrich et al. (2016). This eliminates the need for vacuous 'UNKNOWN' symbols, and is therefore a more ecologically plausible representation of the input. The same tokenization strategy is used in recently proposed Transformer based language models, such as GPT-2 (Radford et al., 2019). The choice of vocabulary did not affect the qualitative nature of the results. In subsequent chapters, all RNN simulations used the sub-word vocabulary.

RNN Training and Hyper-parameter Tuning

Ten simple RNNs were trained in a standard language modeling task on AO-CHILDES. A single weight update (i.e. training step) consisted of the following routines: Sixty-four consecutive word sequences of length 7 were drawn randomly from AO-CHILDES and grouped into a batch. Next, the entire batch was presented to the model, and for each sequence in the batch, the model predicted the next word. Finally, the average prediction error was computed across the batch, and the weights were updated using SGD with a constant, empirically determined, learning rate. By ensuring that there were always exactly seven items in the RNN's memory before updating the weights, the training methodology was equivalent to backpropagation-through-time with gradient truncation applied every seven time steps.[3] This process was repeated until all sequences in AO-CHILDES have been exhausted. At this point, a new epoch began, and the entire process was repeated once more for a total of 12 passes over the corpus (i.e. 12 epochs).[4]

Instead of tuning the model's hyperparameters exclusively on the language modeling objective as is commonly done, I chose hyperparameters that resulted in a compromise between good performance on next-word prediction and a downstream semantic categorization task. Because I am primarily interested in the discovery of lexical semantic categories, this tuning strategy reduced any bias that would have resulted from exclusively optimizing the language modeling objective. Hyper-parameters are shown in 3.3.

window size	7
hidden layers	1
hidden units	512
learning rate	0.01
epochs	12
optimizer	AdaGrad
batch size	64
non-linearity	*tanh*
initialization	uniform between $\pm\sqrt{\frac{1}{512}}$

Table 3.3: Hyper-parameters used to train the RNN.

Usually, next-word prediction is both the task used to train the RNN and the target task used to evaluate the RNN. However, in this work, I am primarily interested in the representations the model has learned at the hidden layer rather than its ability to predict the next word — which reveals little about what the network has learned about semantic category membership. That said, I used the language modeling loss (i.e. next-word prediction error) on a held-out test set as a guide for tuning the parameters of the RNN. Tuning results show that the RNN trained on AO-CHILDES successfully learned the sequential regularities of child-directed input, achieving a mean per-word perplexity of 43.8 ± 0.05 (M ± SEM), given a vocabulary of 4,096 words, and 10 replications. After tuning, and verifying that all 10 RNNs have learned the sequential structure of the input, I examined their learned representations at the hidden layer.

Obtaining Contextualized Representations

I focused my examination on 700 common nouns, which I will refer to as probe words. Each belongs to exactly one of 28 semantic categories such as MAMMAL, PLANET, and VEHICLE. Details regarding probe words can be found in the published paper (P. A. Huebner & Willits, 2018) and in Appendix A.

An important question is how distributional semantic knowledge pertaining to these words can be accessed in the RNN and studied. There are multiple locations in the network where this knowledge might be stored, including in some combination of the three weight matrices (input-to-hidden, hidden-to-hidden, hidden-to output), or in the pattern of activations that is dynamically generated at the hidden layer as the RNN is

[3] Hidden states were not saved across training steps, and were initialized to a zero-vector.

[4] Random sampling was used to guarantee that the model has been exposed to the full corpus at every epoch. After each epoch, batches are sampled in a different random order. This is standard practice in neural network training. In Chapter 10, I examine the potential of an incremental strategy that (i) is cognitively more plausible, and (ii) also respects the longitudinal structure of children's language environment across developmental time.

PC 1	PC 2	PC 3
hah	quack	prince
quack	oink	arrive
oink	woof	merrily
meow	baa	ago
who	meow	grew
zoo	finish	stir
forest	get	fork
bathtub	find	napkin
mirror	share	spoon
couch	play	tissue

Table 3.4: Results of Principal Components Analysis (PCA) of learned contextualized representations of all words in the RNN's vocabulary. The top-5 most strongly (top rows) and most weakly loading words (bottom rows) are shown for principal components 1-3.

processing multi-word sequences. Following previous work (J. L. Elman, 1990), I chose the latter option. Specifically, I examined the dynamically generated activation states at the hidden layer in response to all sequences in the training data that end with a probe word. There are about 200,000 such sequences in AO-CHILDES. To obtain the probe word representations at the hidden layer, all 200,000 sequences were fed back into the RNN at the end of training, and without updating the network's connection weights. During this procedure, the resulting patterns of activation at the hidden layer at the final time step were collected. Next, for each probe word, all the hidden layer representations pertaining to a given probe (i.e. those produced in response to sequences ending with the same probe word) word were averaged. I will refer to the resulting representations as 'contextualized representations'. Importantly, these representations not only contain knowledge about probe words themselves, but they also store information about the average context in which they have occurred in the training data.

Notice that the procedure to dynamically generate contextualized representations is much more computationally expensive compared to accessing statically available information in the network, such as that stored in the connection weights. The procedure used to obtain contextualised representations may be a useful analytical tool for researchers studying the RNN from an engineering perspective, but, as I will argue in great detail in Chapter 4, it is not a plausible model of how children might access their own distributional semantic knowledge if constructed via next-word prediction.

3.3.2 Results

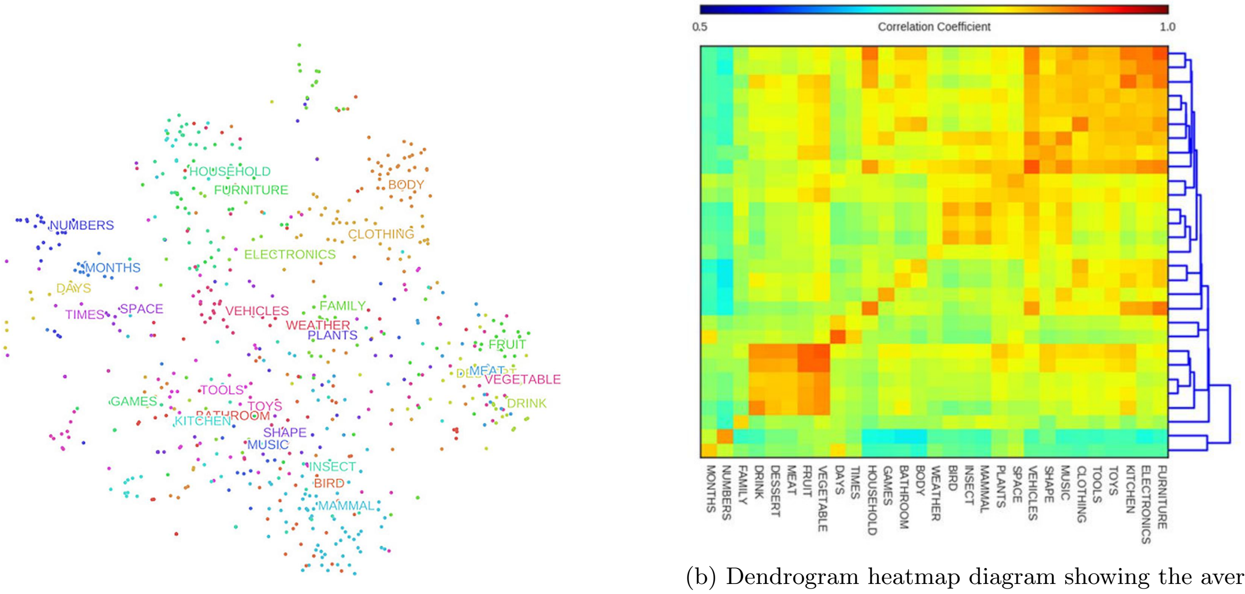
I conducted three types of analyses of learned contextualized representations: First, what are the most important linguistic features used by the RNN to organize representations at the hidden layer? Second, to what extent are the learned representations hierarchically organized? Third, and most relevant to the overall goal of this thesis, how well do the network's semantic similarity judgments of probe word pairs respect the semantic category boundaries assigned by people?

Most Important Dimensions

Prior to evaluating what the RNN has learned about probe words, I first examined how the network has organized its learned knowledge of all the words in its vocabulary. Specifically, I asked what kind of linguistic features the RNN has become most sensitive to over the course of training. Toward that end, I conducted a principal components analysis (PCA) of the contextualized representations for each word in the RNN's vocabulary after training. Here, I only evaluated the first three principal components (PCs). The first three components are especially useful given that the PCs returned by PCA are ordered by the amount of variance each accounts for (from most to least amount of variance explained). To assess what each PC codes for, I inspected the words that loaded most strongly (both positively and negatively) on each PC. The top-5 words that load strongest on each components are shown in Table 3.4. I found that the first two PCs code for grammatical features that appear to be correlated with part-of-speech. The first PC looks as if it codes for noun category membership, and the second principal component sharply distinguishes between verbs and words that tend to appear in isolation such as onomatopoeia and interjections. After the first two PCs, the next began encoding fine-grained semantic distinctions. For instance, PC 3 was effectively coding for the activity context, specifically whether the context is eating, compared to something more akin to playing. Nouns and verbs relating to playing, singing, reading, watching television, and the locations where those events occur, have highly positive activations, whereas nouns and verbs relating to eating have highly negative values on this component. This is not surprising as these are likely two of the most frequent and coherent events in young children's lives, and are also orthogonal in the sense that they rarely occur together. I consider this evidence of abstract knowledge given that in order to arrive at such a grouping, the RNN would need to integrate information and identify commonalities across many situations that potentially differ widely in surface-level language use.

Hierarchical Structure

To investigate to what extent the contextualized representations of probe words acquired by the RNN are hierarchically organized, I conducted two analyses. In the first set of analyses, I visually inspected the learned representational landscape of all 700 probe words at the end of training. A 2-dimensional approximation of the 512-dimensional contextualized semantic vector space was obtained using t-SNE (Van der Maaten & Hinton, 2008), and was reproduced in the left panel of Figure 3.2. The low-dimensional reconstruction revealed a principled separation of probe words that belong to different but related semantic categories. To investigate this pattern more quantitatively, the pairwise representational similarities of all 700 probe words were computed, aggregated by semantic category membership, and then hierarchically clustered. The resulting dendrogram heatmap is shown in the right panel of Figure 3.2. I found that probe words that are both members of the same category tended to be more similar than probe words not belonging to the same category. Moreover, representations of probe words belonging to related categories (e.g. MAMMAL and BIRD; FRUIT and VEGETABLE) were more similar than probe words that did not belong to related categories (e.g VEHICLE and MEAT). This demonstrated that the RNN was able to differentiate probe word pairs from the same, related, and different categories in a graded fashion. Because relatedness was found to be, on average, higher between probe words in the same category compared to related categories, I conclude that the RNN learned semantic relationships at two distinct levels in a semantic hierarchy. This hierarchical organization emerged automatically, and is consistent with previous work by McClelland and Rogers (2003), who showed that a feed-forward neural network, learning about concepts in terms of the



(a) A t-SNE dimensionality reduction projection, showing a 2-D representation of the relative similarities of the 700 probe-words learned by the RNN.

(b) Dendrogram heatmap diagram showing the average similarity of probe words within and between categories. Similarity was computed using the Pearson correlation of pairs of contextualized representations.

Figure 3.2: Analysis of the structure of contextualized representations of probe words learned by the RNN.

correlational structure of their shared features or propositional content (e.g. canaries ↔ are-yellow and have-wings) can be used to explain the apparent hierarchical nature of concepts. The authors argued that hierarchical structure is an emergent property of distributed representations representing the relative similarity of concepts. This contrasts previous views, namely that surface-level statistics are not sufficiently rich for hierarchically organized conceptual knowledge to emerge automatically (Carey & Spelke, 1994; Mandler & McDonough, 1993).

Second, I performed several hierarchical clustering analyses to investigate the extent to which contextualized representations of same-category probe words are also organized hierarchically. Interpretation is based on visual inspection of hierarchical clustering diagrams shown in Figure 3.3. I briefly discuss the results for three categories (FAMILY, KITCHEN, and SPACE). Beginning with FAMILY, the most closely related probe word pairs were *grandfather* and *grandmother*, and *father* and *mother*. These words were part of a branch in the hierarchical clustering which primarily included the formal terms for family members. Another branch was identified with members such as *gran*, *granddad*, *ma*, *dad*, which are their informal counterparts. Second, two distinct branches were identified for probe words belonging to the category KITCHEN. The largest two clusters appeared to be separated according to objects used to prepare food and objects which are associated with eating. Words like *microwave* and *toaster* were found lumped together and separate from a cluster that included words like *teapot*, *silverware*, and *napkin*. Third, the clustering of probe words within the category SPACE revealed a different, surprising pattern. It showed that category labels such as *world*, *planet*, and *star* were lumped on a branch that was separate from proper nouns, such as *venus* and *mars*. This provides evidence that the model learned to separate words that label concrete objects and words that label categories of objects. Combined, the above results demonstrate that the RNN has learned structured semantic knowledge at multiple levels in the semantic category hierarchy: Not only did the RNN recognize differences in probe word pairs that span across category boundaries, but the RNN also learned to distinguish probe words at a finer-grained level of detail (e.g. formality, object function, proper noun vs. category label) within a target semantic category.

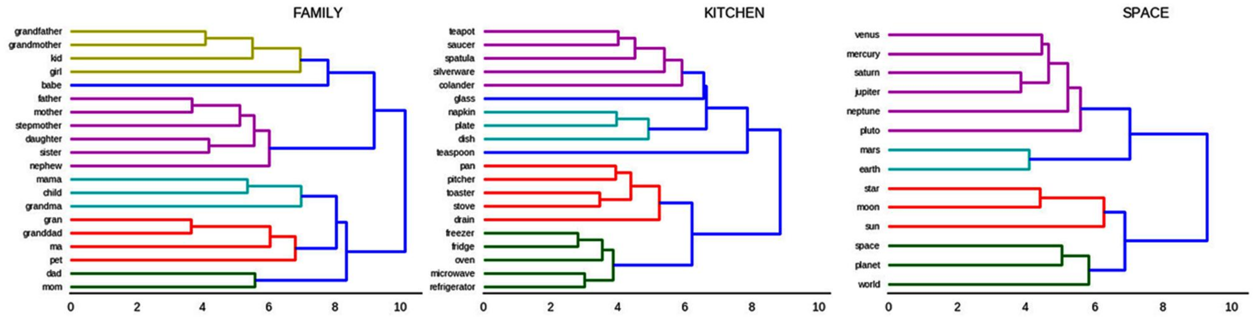


Figure 3.3: Hierarchical clustering dendrograms of probe words in the categories FAMILY, KITCHEN, and SPACE.

Semantic Categorization

As a test of whether the network's learned contextualized representations could be used to derive paradigmatic similarity estimates for the distributionally-mediated extension of category-associated features (DECAF), I conducted a semantic categorization experiment. As discussed in Chapter 1, the first step in performing DECAF is the retrieval of words that are distributionally similar to a novel word. The expectation is that the retrieved words are not only distributionally similar, but also semantically similar. That is, if the retrieved words are to prove useful for performing DECAF, they should share semantic features with each other and the target word. This minimal requirement can be straightforwardly tested with the resources at hand: Because each probe word was assigned to one of 27 semantic categories in a pilot norming study, we can examine whether the similarity relations among learned probe word representations respect human-generated semantic category boundaries. If they don't, then this would constitute evidence against the role of next-word prediction as a supplier of distributional similarity judgments for DECAF.

The quantitative procedure that underlies the semantic categorization task is detailed below, and again in Chapter 5. It should be noted that the RNN was not actually tasked with producing semantic category labels; instead, all pairwise similarities among contextualized representations of probe words were computed, and compared to discrete yes/no judgments based on human-generated semantic category membership assignments. For instance, the result of the pilot norming study is that *cat* and *dog* were both assigned members of the MAMMAL category. This means that, in order to achieve a high score in this task, the model must judge these two words to be systematically more similar than unrelated probe word pairs, such as *dog* and *bicycle*, or *cat* and *fence*. Given that probe word similarity judgments are continuous in nature (between -1.0 and +1.0, as a result of using the cosine similarity) and semantic-category membership is discrete (yes vs. no), an iterative procedure, based on signal-detection theory, was used. Essentially, large range of similarity thresholds was considered, by sweeping from -1.0 to +1.0 in small increments. At each threshold, the model's continuous similarity judgments were discretized (is the similarity above or below threshold?), and then compared to the ground-truth label (yes vs. no, depending on shared semantic category membership). At each threshold, the sensitivity ($\frac{TP}{TP+FN}$) and specificity ($\frac{TN}{TN+FP}$) was averaged to produce the balanced accuracy. The final score is the highest balanced accuracy across all similarity thresholds.

The average balanced accuracy across training and across all 10 networks is 0.679 ± 0.004 (mean ±

margin of error).[5] Notice that this is well above chance-level, 0.5. To understand which semantic categories were learned best, I derived balanced accuracies for each individual category. The results are shown in Table 3.5. Categories are ordered by classification performance, from highest to lowest. Interestingly, there is a large range in classification accuracy, indicting that some semantic categories are more difficult to cluster given the available information and/or do not manifest in the distributional statistics in input available to children. For instance, the human-generated category DAY, which consists of words such as *monday* and *tuesday*, is well approximated by a corresponding cluster in the RNN's contextualized semantic space. The semantic category with the least correspondence in the network is TIME. This category consists of a diverse set of words like *today*, *o'clock*, and *midnight*, and it is therefore likely that the RNN learned to separate these words on grammatical — as opposed to semantic — grounds, and that this impeded the formation of a tight cluster. In sum, these results clearly demonstrate that there is sufficient structure in the surface-level lexical statistics of transcribed speech to children to support the construction of form-based lexical semantic categories.

DAY	GAME	VEGETABLE	FRUIT	DRINK	VEHICLE	ROOM
0.93	0.87	0.84	0.81	0.80	0.77	0.76
FURNITURE	CLOTHING	MONTH	ELECTRONIC	MAMMAL	MUSIC	KITCHEN
0.76	0.76	0.74	0.72	0.71	0.70	0.68
SHAPE	WEATHER	BIRD	TOY	TOOL	FAMILY	DESSERT
0.68	0.68	0.67	0.67	0.66	0.65	0.68
MEAT	SPACE	PLANT	BODY	INSECT	BATH	TIME
0.65	0.64	0.63	0.63	0.63	0.57	0.52

Table 3.5: Semantic categorization accuracy (in units of balanced accuracy) for each semantic category. Chance-level performance is 0.5, and the theoretical maximum is 1.0. Given that there is no data available on people's similarity judgments on the probe word pairs used here, it is not clear where human-level performance would fall within that range. It is likely that human performance is well below 1.0.

3.4 Summary

Recent advances in our ability to study large, naturalistic datasets, combined with advanced computational modeling techniques, have allowed us to ask ever bigger and more ambitious questions about the nature and statistics of children's language environment. One of the major insights from investigating the structure of experience is that it has forced us to radically re-evaluate traditional theories of language learning and representation. Many models previously deemed insufficient (especially deep learning systems with non-linear transformations) perform qualitatively differently when faced with large amounts of data, often enabling them to solve complex tasks without being explicitly told what features to pay attention to.

[5] This number is not comparable to the balanced accuracy reported in the published paper (P. A. Huebner & Willits, 2018), because the networks that were examined in this chapter were re-trained using identical setting used elsewhere in this thesis. For instance, in contrast to the published paper, the networks trained in this thesis used a larger vocabulary, and a more sophisticated tokenization strategy. Further, the probe words used in this thesis did not include the NUMBER category. Re-training with matched settings ensured that all simulations presented in this thesis, and by extension, all the balanced accuracies, are comparable.

It has been more than 30 years now since J. L. Elman (1991) first described his pioneering work on the simple RNN. Due to limitations on computational resources available at the time, he was not able to address whether the RNN could scale up to noisy naturalistic language input. It took almost two decades until the first demonstration that the RNN could scale up to large natural language corpora (Mikolov et al., 2011). Not soon after, the RNN and its variants have revolutionized computational linguistics, replacing, for instance, language modeling techniques based on n-grams, and syntactic parsing systems based on hand-crafted rules. The work reported in this chapter continues this line of work in the domain of transcribed speech to children. The finding most relevant for the overall goal of this thesis is that the RNN grouped common nouns in a manner that reasonably approximated the semantic categorization produced by people, and that it could do so given noisy, transcribed speech to children. Therefore, it is safe to conclude that the insights of J. L. Elman (1991) do not depend on the cleanliness of the artificial dataset, and that there is sufficient structure in the input that children receive to support the distributional construction not only of grammatical, but also of fine-grained lexical semantic, categories.

In sum, the modeling results support the idea that next-word prediction, combined with distributed representations and error-based learning, are tools that a child could, in principle, employ to support his or her word learning. More precisely, the results constitute preliminary evidence that children could leverage existing prediction-based learning and processing capacities to acquire distributionally constructed estimates of lexical semantic similarity as the first step to performing the distributionally-mediated extension of category-associated features (DECAF). In the next chapter, I examine this argument in greater detail. There are still many potential issues that need to be addressed — in particular, the relation between dynamically generated contextualized representations and the need for statically available knowledge to perform DECAF.

Lastly, a disclaimer is in order: The work herein differs in important ways from previous distributional semantic modeling work. While in previous work, the knowledge learned by the RNN or a more traditional distributional semantic model is often considered to be directly incorporated into semantic memory (J. L. Elman & McRae, 2019; Landauer & Dumais, 1997; Lund & Burgess, 1996), the work herein makes no such assumption or theoretical commitment. Contrary to previous work, the goal of this thesis is not to explore how word or phrasal meanings are in part distributionally constructed; instead, the goal is to examine the feasibility of the RNN as a mere supplier of paradigmatic similarities that would enable children's distributionally mediated extension of category-associated features (DECAF). On this view, distributional statistics themselves do not necessarily contribute information to the mental representations of concepts, but, instead, support their induction — the extension of extra-linguistic semantic features from existing to newly formed meaning representations. This means that DECAF does not require that the distributionally constructed lexical representations be part of the mental lexicon or that they directly enter into computations performed during ordinary comprehension and production by adults. DECAF is a special-purpose tool that is most applicable during word learning; whether such a system continues to be of use to adults in ordinary language use is an interesting question, and beyond the scope of this thesis.

Chapter 4

Desiderata for Modeling DECAF

The previous chapter established that next-word prediction can be used to acquire contextualized representations that contain semantic category knowledge from child-directed input. While this finding is a crucial first step in testing the idea that next-word prediction might underlie children's accumulation of lexical semantic category knowledge useful for extending semantic features to novel distributionally similar words, much more work needs to be done. In this chapter, I examine theoretical concerns about the feasibility of learning *non-contextualized* lexical semantic representations in the RNN, given that it was not specifically developed for this purpose.

More broadly, I discuss several theoretical desiderata any distributional semantic model should posses in order to be taken serious as a supplier of lexical semantic representations useful for performing the distributional extension of category-associated features (DECAF). The desiderata I consider are (i) access to information about paradigmatic similarity, (ii) lexical atomicity, (iii) the ability to distinguish semantically relevant and irrelevant features, (iv) structured meaning decomposition, and (v) cognitive plausibility. Each of these is discussed in turn, below. Special attention is paid to what degree the simple RNN satisfies each of these desiderata.

4.1 Paradigmatic Similarity

In order to be maximally successful, the measure of distributional similarity that is most useful for performing DECAF is paradigmatic rather than topical similarity. Paradigmatic similarity assigns higher scores to pairs such as *dog-cat*, but not *dog-leash*. At its core, paradigmatic similarity is often considered a measure of feature overlap as opposed, to, say, associative strength.

One perspective that will be exploited in this work, is the close relationship between paradigmatic similarity and membership in the same part-of-speech (POS) class. For an illustration of this idea, consider the lexical category hierarchy depicted in Figure 4.1. At the top-level, we find broad ontological distinctions, such as between objects, properties, and actions, and these roughly map onto POS classes. On this view, POS classes can be considered top-level distinctions in a lexical semantic category hierarchy. A consequence of this framing is that linguistic substitutability can be used as a stand-in for semantic feature overlap, and, by extension, paradigmatic similarity. Not only do nouns pick out similar concepts, but their language-internal distributional properties reflect this fact. As we travel down the hierarchy, concepts share more semantic features. At the same time, their linguistic contexts should become more specific — at least, according to

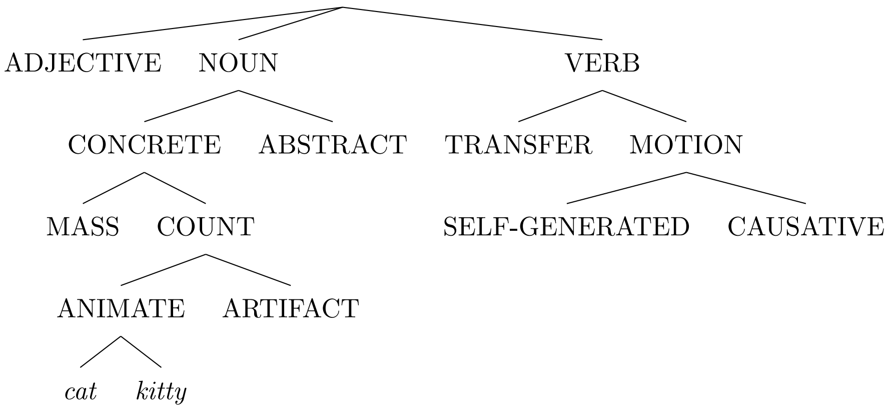


Figure 4.1: A simplistic lexical semantic category hierarchy, with POS classes as top-level distinctions, and lexical items as leaf units.

the distributional hypothesis. At the bottom-level of the lexical semantic category hierarchy are synonym clusters, such as *cat* and *kitty*, which are most substitutable in meaning and in language.

An early investigation of children's knowledge of paradigmatic similarity at the top-level was conducted by R. Brown and Berko (1960). One of the primary observations was that that children's organization of words into adult POS classes is positively correlated with amount of language exposure. More specifically, the authors suggested that the accumulation of language-internal distributional evidence (i.e. evidence about 'formal' associations) is a key factor in children's knowledge of paradigmatic similarity. This idea is adopted implicitly in this work, by training an RNN to discover paradigmatic similarities via exposure to linguistic input. It should be noted that contrasting theories have been proposed in the past, which argue that children are born with rudimentary knowledge about part-of-speech categories and that, for instance, the distinction between nouns and verbs exist in all human languages. On this latter view, the primary challenge to the learner is to discover the language-specific rules that govern their usage (Valian, 2014; Wasow, 1973; for review, see Keibel, 2005).

In the computational realm, M. Jones and Recchia (2010) argued that paradigmatic similarity is essential for integrating language-internal distributional knowledge with extra-linguistic perceptual knowledge. What psychologists would like their models of human semantic memory to do is to make inferences across these two information sources. Just like a child might exploit language-internal distributional knowledge to extend category-associated features to novel words (DECAF), M. Jones and Recchia (2010) investigated the ability of a distributional semantic model to infer perceptual properties of novel words. Borrowing an example provided by the authors, consider a child has learned from perceptual experience that sparrows have beaks. Further, consider that the same child has encountered the novel word *mockingbird*. If the child has accumulated sufficient linguistic experience to infer that *mockingbird* and *sparrow* tend to occur in similar contexts, he or she can infer that mockingbirds too have beaks. M. Jones and Recchia (2010) call this 'feature migration', and demonstrate, using computational simulations, that it works best when the distributional model constrains the migration of features to word pairs with high paradigmatic similarity. A key ingredient for learning about paradigmatic similarity is word-order; when word-order information is removed from the model, it makes many more feature migration errors relative to when word-order information is left intact. Unsurprisingly, the kinds of errors that the model makes in the absence of word-order information is to migrate perceptual semantic feature from known words to topically related words (e.g. has-beak → *tree* rather than has-beak →

mockingbird).

4.2 Lexical Atomicity

A distributional system that is used to supply lexical representations for performing DECAF should strive to approximate lexical atomicity. In short, this means that learned knowledge should be neatly divided into full-featured, standalone packages that correspond to individual words, as opposed (but, potentially, in addition) to large, chunk-level multi-word components of sentences. Specifically, I define atomicity as a property of learned lexical representations such that the semantic properties causally related to the entity denoted by the lexical item are encoded in the representation of that item as opposed to elsewhere in the system.

There are several subtleties that need to be considered when discussing lexical atomicity. Some of these subtleties need to be discussed to preempt potential preconceptions or misconceptions a reader might have about the meaning of atomicity, or expectations about the kinds of theoretical commitments that atomicity might imply.

First, atomicity does not preclude context-sensitivity. At its core, I consider a lexical semantic representation to be atomic when a system need not be in a particular (contextualized) state to fully access category-associated statistical information about the entity that the representation corresponds to. In the language-internal distributional realm, statistical associations of this kind are those that are most defining of the semantic category that a word belongs to. The basis for this distinction is discussed in greater detail in the next section. Importantly, this definition does not preclude the existence of contextualized states, context-effects on processing, or the integration of context information into lexical semantic representations — all of which are substantiated by extensive behavioral and modeling work (G. T. M. Altmann & Mirković, 2009; J. L. Elman, 2009; MacDonald, 2013; McRae et al., 2005; Seidenberg & MacDonald, 1999; Trueswell et al., 1994; Yee & Thompson-Schill, 2016). This point cannot be overstated: An atomic lexical semantic representation does not imply the absence of or disallow higher-order contextual effects. Instead, an atomic lexical representation is one that does not require supporting context to activate features that are essential for lexical semantic categorization. If supporting context is available, it should be exploited, and atomicity does nothing to prevent this.

Another subtlety related to my conception of atomicity is that atomicity is not strictly a property that a lexical semantic representation *must* have, but a heuristic that a learner may apply to make the most of his or her limited experience with language for the purpose of learning novel words. I agree with J. L. Elman (2011) that words are ‘cues to meaning’, and that words activate semantic features in a context-dependent, graded fashion. In fact, I further accept that some words may be so elusive or ambiguous that they do not activate any semantic features in-full. The interpretation of such a word is heavily dependent on the contexts in which those words occur; activation in the absence of supporting context would result in little more than the sprouting of a fuzzy, ill-defined cloud of partial feature activations. Thus, in contrast to lexical atomicity, some words may even *require* contextual support to arrive at a single disambiguated meaning. In sum, I accept that mental representations are fuzzy, graded, and often statistically derived approximations as opposed to atoms of meaning. Taking this state of affairs as a given, I can think of situations in which a child may benefit by striving for lexical atomicity. When inferring semantic features for novel words using distributional similarity as a guide (i.e. DECAF), it is critical that the knowledge that is used to perform this inference is actually about words, rather than larger chunks of language.

The atomicity desideratum does not require that all words need to be encoded atomically. Instead, whether or not atomicity should be the goal depends on the word and the target task. While some words may not activate (i.e. 'cue') any semantic features in-full, others might benefit by activating one or more semantic features in-full. By activating features 'in-full', I do not suggest that activation is all-or-nothing; instead, 'in-full' means that a semantic feature should be activated in proportion to the degree to which a given word licenses a given feature under some theory of category-relevance. For instance, the word *bicycle* licenses a graded pattern of activation over features that diagnose it as a member of the VEHICLE category regardless of what context *bicycle* has been associated with. What I am proposing is that if there are semantic features that should be activated in-full (according to some theory of category-relevance), then a distributional system tasked to supply lexical semantic representations for the purpose of DECAF, should respect and preserve such word-feature mappings by making sure they are not made reliant on some additional (potentially idiosyncratic) contextualized state. To sum up, I do not claim that all word-feature or word-word associations should be encoded atomically. Instead, atomicity is a word-specific and task-specific heuristic that emphasizes the preservation of associations that are important for diagnosing category membership for the purpose of performing DECAF. To illustrate, consider the statistical association between the word *dog* and the upcoming phrase '*plays outside*'. The latter is distributional support for membership in the ANIMATE category and should therefore be encoded in a manner that does not require contextual support. Specifically, this means that the combination of semantic units in charge of predicting '*plays outside*' should not require pre-activation. The upcoming phrase should be predicted in-full by *dog* alone, and this prediction should be made readily available and in the absence of supporting evidence. The reader should keep in mind that this does not imply an argument for or against context-independent mental representations of words or the concepts they express (for a discussion, see Fodor and Pylyshyn, 1988; Yee and Thompson-Schill, 2016). Further, I do not wish to argue against the facilitatory role that chunk-level knowledge may play in other areas of language acquisition, such as determiner-noun agreement (Arnon & Christiansen, 2017).[1]

Third, atomicity, as I will use it in this thesis, does not imply a hard on/off distinction for the activation of semantic features. While I think that a theory of category-relevance should aim at such a distinction, it is likely that a theory that cleanly divides features into yes/no sets (i.e. is category-relevant or not?) is untenable. Considering that a graded theory is inevitable, atomicity is more about the preservation of the graded nature of category-relevance than some artificial distinction between which features should be on or off when a a particular word is activated. Atomicity is not about imposing additional constraints on representation, learning or processing, but a heuristic aimed at preserving those statistical associations that are potentially most generalizable (e.g. most useful for performing DECAF). Specifically, lexical atomicity aims to preserve the semantic category-associated graded pattern of activation inside the representation of the target word, and de-emphasizes encoding the same pattern in the combined activation of the target word and the idiosyncratic linguistic contexts in which it had been previously observed.

Fourth, and last, atomicity should not be confused with context-independence. Learning in any distributional semantic model, including the RNN, is based entirely on contextual information; the representations learned by such model are therefore — by definition — not context-independent. Instead, atomicity has to do with whether or not a particular context-dependent state is required to fully activate all category-relevant semantic features of a word (i.e. predictions about likely upcoming words). Put differently, atomicity has to

[1] It is plausible that people represent linguistic expressions at multiple levels of granularity simultaneously, such that more or less atomic lexical representations and representations of multi-word chunks can co-exist in the same system — that is, they are not mutually incompatible.

do with the properties of a system that has already been trained, rather than the training procedure.

To understand lexical atomicity, consider Figure 4.2. As words are input to a hypothetical distributional semantic model one step at a time (x-axis), they activate semantic units in that system. In the y-axis, I only show the activation along units associated with semantic category-associated features for the word *gorilla*. That means that activation of these units predict upcoming words like *breathes*, *eats*, *jumps*, and other words that are in part diagnostic of membership in a semantic category (e.g. APE, MAMMAL, ANIMAL, ANIMATE). Importantly, the difference between a system that has encoded the word *gorilla* atomically vs. non-atomically is shown in the difference between the blue and red bars respectively. When atomic, the lexical semantic representation for the word *gorilla*, for instance, would contain all the knowledge previously acquired about gorillas, and would not require the pre-activation of incidentally related concepts (e.g. *jungle*) for this knowledge to become available. In contrast, when non-atomic, the word gorilla does not by itself fully activate units associated with semantic category-associated features (see difference in bar height), but requires pre-activation provided by the supporting context (e.g. *jungle*). Only when the combined activation is accumulated across time steps, do the semantic category-associated features become fully activated. In situations like these, I refer to semantic category-associated knowledge as being 'trapped' in the processor (e.g. hidden layer of a neural language model).

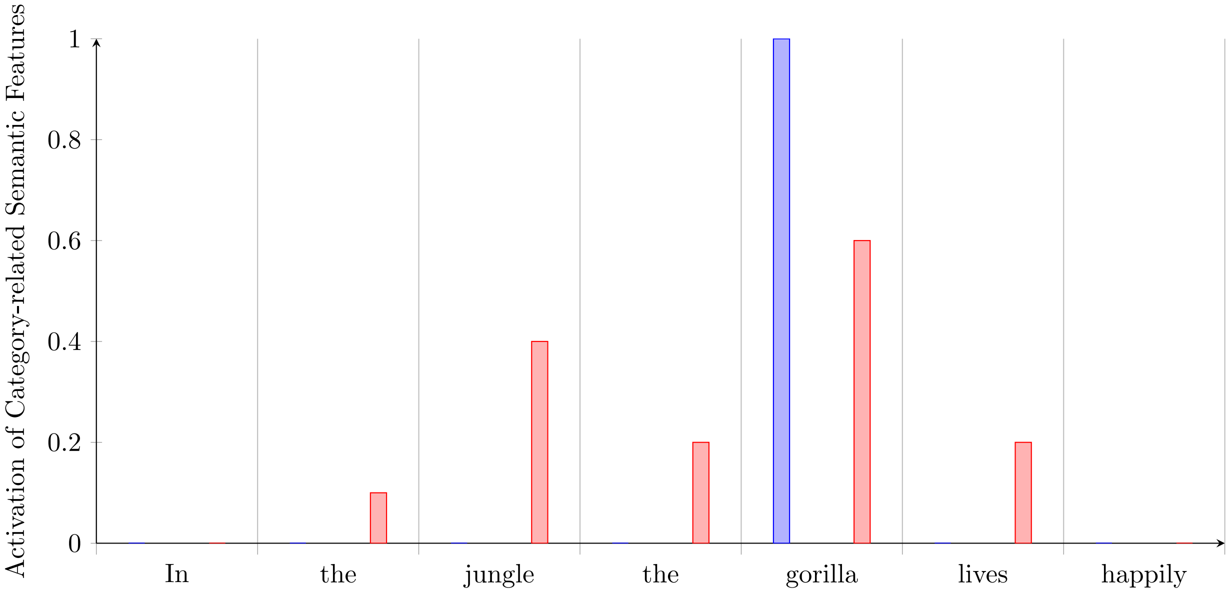


Figure 4.2: A schematic illustrating the difference between an atomic (🟦) and non-atomic (🟥) lexical semantic representation for the word *gorilla*.

Atomicity a tricky topic. To be clear, my view of the organization of the mental lexicon and how concepts are represented by people aligns with proposals by J. L. Elman (2011) and Yee and Thompson-Schill (2016) who have rejected the traditional notion of static, context-independent concepts in favor of probabilistic knowledge that is activated in a graded manner and is highly sensitive to context. As such, I do not claim that concepts or word-meaning mappings are atomic; the purpose of this thesis is not to argue for or against a particular kind of mental representation. Instead, it appears that for the specific task I am interested in, the distributionally mediated extension of semantic category-associated features, atomicity is likely desirable. Importantly, however, this does not preclude the possibility that there are other tasks, situations, or systems that draw upon non-atomic linguistic representations.

Furthermore, I do not consider atomicity as a hard constraint on the kinds of linguistic representations that are potentially useful for performing DECAF. Instead, I consider lexical atomicity as a useful starting point, which enables a young learner to get the most out of their inventory of co-occurrence statistics. Once a learner has exploited the atomicity of his or her distributionally constructed representations, it is likely that a weakening of atomicity would be useful so that a learner can take advantage of more complicated chunk-level statistical relationships when performing DECAF. Put differently, I accept that a learner should not be stuck with atomic representations; to keep benefiting from language-internal distributional knowledge, a learner should be able to make additional room for more complex statistical dependencies, such as (i) dependencies that span longer distances, (ii) dependencies between words at the chunk-level (i.e. noun phrase), and (iii) dependencies informed by more sophisticated knowledge of syntactic relations and even world knowledge.

The reason that I think that lexical atomicity is a useful starting point is not because there are no meaningful relations among words at the chunk-level or that these relations are not useful for guiding inferences during word learning. In the contrary, examples of this abound in language. For instance, the complex noun phrase '*a hungry gorilla*' is semantically more compatible with the completion '*eats bamboo*' than the bare noun phrase '*a gorilla*'. Noticing these chunk-level distributional phenomena can be enormously helpful to children. The key point is that once more complex associations of this type are encoded in systems that track arbitrary sequential dependencies, it can be difficult to decompose those statistical associations. The challenge of using prediction-based neural language models to derive lexical semantic representations is not that they fail to capture higher-order contextual phenomena — this is precisely what they were developed to do – but, instead, that they rely on higher-order contextual phenomena *at the expense of atomic associations*. Stated differently, the challenge of using language models is to get them to do something they are not particularly good at, but which would yield representations of immense use to language-learning children. Once atomic representations have been established in such models, it is straightforward to expand their statistical repertoire; however, it is much more difficult to do this in the reverse order. I discuss this idea in detail in Chapter 9 and test it in Chapter 10.

To summarize: Atomicity is useful when performing the distributional extension of category-associated features (DECAF) where it is important that all — and only — features relevant to a given lexical item and its semantic category are considered during similarity computation. Why? If the knowledge encoded in the distributionally constructed lexical semantic representations require additional context to be activated, then these representations are less likely to apply to a broad range of situations, and therefore be of less use to a young word learner. In a nutshell, the atomicity desideratum states that the degree to which lexical co-occurrence associations should be encoded atomically should be proportional to the extent to which the association is category-relevant. The more important an association is for diagnosing lexical semantic category membership, the less support from context-based pre-activation should be required in order to activate those associations in-full. Consequently, the most productive way to think about the lexical atomicity desideratum in the context of graded statistical learning systems, is that such systems should approximate atomicity as closely as possible if the representations it supplies are to be used to perform DECAF. Put differently, atomicity is not necessarily a goal that must be met in order for DECAF to succeed at all, but a target worth aiming at.

We must keep in mind the distinction between theory and practice. The difference between atomic and non-atomic lexical semantic representations is a theoretical distinction, but not necessarily one that must hold in practice. Atomicity is not necessarily something must be formally baked into a system, but something that may be approximated. The success of a distributional system therefore does not rely on meeting the lexical atomicity desideratum formally, but is proportional to the degree to which lexical atomicity is respected — regardless of whether the system was specifically developed for that purpose or not.

4.3 Distinguishing Relevant from Irrelevant Associations

Not all lexical relations are of the same type. In this thesis, I distinguish between obligatory predicate-argument relations and non-obligatory predicate-adjunct relations (Koenig et al., 2003). While the former often provide useful information about the semantic category of participants, the latter relations typically do not. For simplicity, I refer to these as category-relevant and category-irrelevant dependencies, respectively. While there is no consensus, as of yet, how to distinguish between dependencies of this sort in any principled manner, I will argue that a model tasked with supplying lexical semantic representations for performing DECAF must have some rudimentary ability to draw this distinction.

The distinction between category-relevant and category-irrelevant is not only relevant to formal semanticists, but also to psychologists interested in semantic development. For instance, F. C. Keil et al. (1998) pointed out that whether or not perceptual features such as color are integral to the definition of a concept depends on the concept: Although many washing machines are white, whiteness is not critical to the concept ‘washing machine’. In contrast, many polar bears are white, and whiteness is in fact integral to the concept ‘polar bear’. How do people know that the former correlation is not relevant, whereas the latter is? This question is not only applicable when identifying individual concepts, but extends to semantic categories. Whereas motion is a critical feature of vehicles, the presence or absence of observed motion alone is not a basis for categorizing an objects as a vehicle. A moving object need not be a vehicle, and a stationary object may be a parked vehicle. The reason these questions are important to consider here, is because it matters what is encoded in the lexical representations that are constructed by a distributional semantic model. If the learned representations do not capture more category-relevant than category-irrelevant information, then distributional similarities based on pairwise comparisons of these representations will not be useful for modeling children’s distributional extension of category-associated features (DECAF). Under many linguistic theories, category-relevant semantic dependents are encoded in the lexical entry of a word, whereas category-irrelevant dependents are not (Koenig et al., 2003). Following this insight, it would appear that a model in charge of producing lexical similarity judgments based on distributional overlap should also be tuned to this distinction.

I do not claim that a distributional semantic model like the RNN language model must arrive at a perfect understanding of what features — in this case, co-occurrence features — define a given semantic category. It is not even clear whether this distinction is tenable in theory, let alone in practice (Murphy, 2004). Furthermore, a distributional semantic model is provided only access to language-internal relations and therefore cannot distinguish between correlated and causally related co-occurrences. To ask such a system to arrive at a causal decomposition of its input would be naive, and indicate a misunderstanding of the goals of the connectionist/emergentist enterprise. Instead, the best a distributional semantic model like the RNN can do is to separate category-relevant (i.e. category-defining) from category-irrelevant relations statistically when there is sufficient information in the data. The challenge for such models is to get the most out of the existing data. That said, this is an area in which corpus-based connectionist and distributional semantic models still struggle, and the simple RNN is no exception.

A similar debate is ongoing in the field of concept acquisition. For instance, some scholars think that infants use internal knowledge to tease apart similarity-based perceptual features that are conceptually relevant for categorization from those that are not (Carey & Spelke, 1994; Gopnik & Meltzoff, 1987; Mandler & McDonough, 1993). Similarly, some think that the identification of causal relationships is an important pre-requisite for being able to tease apart spurious from category-diagnostic associations (Gopnik & Meltzoff, 1987). While associative learning does not provide a clear mechanisms for how to distinguish cues that are informative (i.e. predictive) but nonetheless category-irrelevant, proponents of associationism have argued that the correlational structure — in particular the multivariate correlational structure — among features belonging to the same category can go a long way to explaining phenomena in concept acquisition (Rogers & McClelland, 2004). Others have argued that teasing apart category-defining from other features is not necessary, and does not accurately characterize people's conceptual knowledge and responses in behavioral tasks (Goldstone, 1994).

Let us return to language-internal distributional learning. To what extent to distributional semantic models capture the distinction between category-relevant and category-irrelevant lexical associations? While there is not a lot of work in this area, preliminary findings suggest that traditional distributional semantic models are able to pick up on statistical information that correlates with this distinction. In particular, models based on co-occurrence counting can, on average, distinguish between noun-verb pairs where the noun is an obligatory vs. non-obligatory argument (J. A. Willits et al., 2007). This maps onto the relevant vs. irrelevant distinction because obligatory arguments are considered to be semantically more closely connected with their verb than non-obligatory arguments. Therefore, this result demonstrates that traditional (i.e. count-based) distributional models can satisfy this desideratum.[2]

The ability to distinguish relevant features from irrelevant features that happen to be present in the learning environment is a pre-requisite to 'predication'. Predication is the ability to form predicates, which represent abstract relations between its arguments. Put simply, predicates are functions that perform (linguistic, conceptual) operations on arbitrary inputs without modifying the content of the arguments that are input to the function. In addition, the relation that a predicate defines does not depend on the content of its arguments — the predicate fulfills the same role for any combination of legal arguments. The ability to perform predication is indicative of a system that is able to separate abstract relations that hold between sets or categories of items from the content of those items. In the cognitive sciences, a long-standing debate revolves around whether predication is necessary to fully characterize human cognition (including language abilities), and whether sub-symbolic and distributed representations can converge on or approximate formal behaviors of systems with predication built-in (L. Doumas & Hummel, 2005; Hummel & Holyoak, 2003).

How exactly does a distributional semantic model distinguish relevant from irrelevant statistical associations? This depends on the precise definition of what is meant by 'relevant'. In this work 'relevant' means 'relevant to semantic categorization'. Separating distributional cues along these lines, at first blush, appears to require additional extra-linguistic experiential knowledge such as cause-effect relations and rich, potentially theory-driven, world knowledge (Carey & Spelke, 1994; Gopnik & Meltzoff, 1987; Mandler & McDonough, 1993). It is not entirely clear yet just how far one can go with distributional information alone. It is likely that a system that prioritizes lexical atomicity would be more successful if it is integrated with the conceptual system in a bi-directional manner. For simplicity, in this work, only a uni-directional interaction is assumed, such that the burden of teasing apart category-relevant from category-irrelevant associations rests entirely on

[2] It should be noted that not all models that were tested by J. A. Willits et al. (2007) succeeded; to do so, a model must track co-occurrences within a window of a particular size (8-word window).

the distributional semantic model.

4.4 Structured Meaning Decomposition

A system that processes words sequentially, and which is tasked to produce lexical semantic representations in a format useful for DECAF, should be able to transfer its learned lexical semantic information from the processor to lower levels of the system where it can be statically available and readily accessible. If not, semantic information would be trapped inside the processor, and would therefore not be accessible to a child who might wish to consult his or her knowledge about which known words are distributionally similar to a novel word. In this work, I refer to the ability to derive lexical-level information from sequence-level input (i.e. sentences) as 'structured meaning decomposition'. I will argue it is another important desideratum that a distributional semantic model should be able to do if it is to account for children's distributionally-mediated extension of category-associated features (DECAF). Structured meaning decomposition comes in many flavors. On one end of the spectrum, formal semanticists and proponents of compositionality have argued that humans comprehend novel utterances by decomposing chunk-level and sentential meaning into standalone units that each contribute to the overall meaning according to grammatical rules (Chomsky, 1957; Fodor & Pylyshyn, 1988). The smallest units of meaning are typically called 'lexical items', and, in a compositional system, they can be combined to form larger units while maintaining the representational content of the smaller units. There are many advantages of thinking about linguistic structures in this way, and numerous scholars in the language sciences have built on this idea to evaluate existing and develop new models of language processing and acquisition (Abend et al., 2017; L. A. A. Doumas et al., 2008; Gershman & Tenenbaum, 2015; Gordon et al., 2019; Martin, 2020; Martin & Doumas, 2017; Noelle & Zimdars, 1999). On the other end of the spectrum, structured meaning decomposition need not involve an explicit grammatical formalism (e.g. phrase structure grammar, combinatorial categorial grammar, lambda calculus); instead, temporal structures are considered to emerge gradually via experience-dependent learning as in connectionist networks (Christiansen & Chater, 1999a; Gulordava et al., 2018; John & McClelland, 1990; Pannitto & Herbelot, 2022; Rabovsky & McClelland, 2020; Tomasello, 2005). When I speak of 'structured meaning decomposition', I align my definition with the latter camp, wherein domain-general statistical abilities are the primary driver of the discovery of latent linguistic structure. That said, it is possible that in order to acquire fully atomic lexical semantic representations, a stronger version of structured decomposition will be necessary.

Broadly, structured meaning decomposition is related to a long-standing challenges in machine learning, namely learning (i) disentangled, and (ii) hierarchically structured representations Bengio et al. (2013). First, learning disentangled representation requires that single computational units or components become sensitive to changes in single factors of variations in the data, while being relatively invariant to changes in other factors. For example, a computer vision model trained on a dataset of images might learn factors such as object class, position, scale, lighting, or colour. In the domain of language, these factors could be word meaning and syntactic relations between words. Learning disentangled representation of language should separate these two (Kádár et al., 2017; Li et al., 2016; Martin, 2020). Second, neural network models are often criticized for not representing language or concepts in a hierarchical way that is necessary for language (Fodor & Pylyshyn, 1988; Gershman & Tenenbaum, 2015; Marcus, 1998; Pinker & Prince, 1988). But it is useful to distinguish between what a neural network can represent, and what a neural network can learn to represent. Any structured, hierarchical representation can be encoded in a vector representation, and can be represented in a network's weights. Neural networks with hidden layers are, after all, universal function approximators (Scarselli & Tsoi, 1998; Schäfer & Zimmermann, 2006; Siegelmann & Sontag, 1992). Thus, there is nothing about neural networks that is incompatible with a theory that says that language must be represented as a system of discrete, hierarchically-organized symbols. The question is whether any particular neural network model can learn the correct structured representation of the language from the input.

A hallmark ability of systems that perform structured meaning decomposition is to identify how changes in individual words (e.g. replacement of one word with another) changes the overall meaning of a sentence. Often, this ability requires knowledge of the syntactic relations among words in a sentence, and how syntactic transformations can impact interpretation of meaning. For example, consider the sentence '*The hungry mouse that followed me chased the cat*'. Despite being well formed, it describes a highly implausible event — a mouse chasing a cat. An English-speaking adult would have little difficulty identifying the minimal change needed to make the sentence more plausible. For instance, we might replace the subject noun *mouse* with the word *dog*. However, a distributional system that does not perform structured meaning decomposition, or does this only poorly, might instead replace larger chunks of the original sentence, including, for instance the adjectival modifier and relative clause alongside the subject noun. Such a model might replace the chunk '*hungry mouse that followed me*' with a more plausible noun to produce '*The dog chased the cat*'. This would be evidence that the model has not learned how adjectival modifiers and relative clauses relate to the overall meaning of sentences — in this case, the meanings expressed by each did not contribute to the implausibility of the event described by the sentence. In fact, neural language models are known to struggle precisely in this way; their tendency to chunk predictable multi-word sequences, while useful for next-word prediction, is a handicap for discovering how smaller lexical units relate to each other within those chunks. This idea is supported by mounting evidence from computational studies. For instance, Gershman and Tenenbaum (2015) examined the ability of a variety of neural network based distributional models to rank sentences according to the amount of meaning overlap. Given the base sentence '*A young woman in front of an old man*', the authors manipulated sentence components to produce sentences that differ in the amount of semantic congruence: 'An old man behind a young woman.' (meaning preservation). '*A young man in front of an old woman*' (noun change). 'An old woman in front of a young man' (adjective change). 'A young woman behind an old man' (preposition change). Surprisingly, neural network based models produced rankings that differed widely from human subjects; while humans tended to rate meaning-preserving sentences as highly similar to the base sentence, the neural network models investigated tended to score these as most dissimilar. The authors concluded that structured meaning decomposition does not emerge automatically in their sample of neural network based distributional models, and that more sophisticated procedures in addition to tracking co-occurrence associations are needed to account for people's knowledge about the role of individual words in semantic composition (i.e. construction of phrase and sentence-level meaning).

An early example of a neural-network based model developed to perform structured meaning decomposition is the Sentence Gestalt (SG) model by John and McClelland (1990). The (SG) model was used to simulate sentence comprehension by jointly modeling syntactic, semantic and thematic constraints. Importantly, this structured knowledge, the authors claimed, can be used to determine how individual words in a sentence contribute to the overall meaning of the sentence. The model consists of an update and a query network: The first is used to update the sentence gestalt - an evolving representation of the sentence as a whole - on a word-by-word basis. Each word modifies the sentence gestalt produced at the previous time step, updating information about the event described by the sentence. To train the model to develop a useful sentence gestalt, the model is probed at every time step, using the query network. This network processes both the sentence gestalt and a query to produce a prediction concerning one aspect about the event described by the sentence. For example, a query might involve a question such as "Who is the agent of the event?" or "What is the action of the event?". Crucially, the model is able to infer aspects of an event not explicitly described by the sentence provided in the input. For instance, the network might output *steak* if asked about the patient of the sentence '*The busdriver ate the food*', given that bus drivers often eat steak in the training data. The sentence gestalt is essentially a hidden representation of the sentence. It uses a distributed code shaped by the thematic role information used to probe the hidden representation during training. The thematic roles used to train the network consist of both semantic and syntactic information, which, once encoded into the hidden representation, are difficult to tease apart. In fact, the aim of the model is to develop representations of multiple constraints and their interactions, rather than a principled distinction between syntactic and syntactic factors on sentence comprehension. This can be considered both a weakness and a strength. The advantage of this approach is that interactions between semantics and syntactic factors need not be specified in advance, and, instead, emerge automatically during training. On the flipside, because the lexical knowledge of the network is implicit in the sentence gestalt representation, and requires a separate query system to be retrieved, it would be difficult to combine the lexical knowledge it has acquired with other systems and tasks (e.g. DECAF) in a modular way. For a more recent discussion of connectionist systems and structured decomposition, see (Rabovsky & McClelland, 2020).[3]

4.5 Cognitive Plausibility

Few distributional semantic models are intended as mechanistic accounts of children's lexical semantic development. Historically, distributional semantic models have focused on outlining the computational problem, with little consideration of limitations on memory or other computational resources of language-learning children. Furthermore, some models have been created, and used solely as tools for understanding just how much and what kind of information there is in corpora for learning about word meanings. However, the applicability of a system for modeling DECAF would be bolstered by considering the plausibility that a computational solution can be implemented in the still-developing brains of children. Plausibility may be evaluated at the cognitive and/or neural level; here, I focus on the former.

A useful way to illustrate what cognitive plausibility entails is to give an example where cognitive plausibility falls short: For instance, Word2Vec has been proposed as an efficient engineering solution to capture aspects of word meaning in Natural Language Processing (NLP) applications (Mikolov, Chen, et al., 2013). The introduction of Word2Vec has revolutionized computational semantics, due to the model's ability to efficiently process large amounts of natural language data, and has quickly become a the go-to tool for learning word representations from text. The representations learned by Word2Vec perform reasonably well in a number of semantic tasks, including analogical reasoning (Mikolov, Chen, et al., 2013). However, Word2Vec raises some concerns with regards to being taken seriously as cognitively plausible models of semantic development. For example, it contains a number of optimizations to speed training on large corpora, but some of these optimizations seem unlikely to be the way that children learn. One requirement for training Word2Vec is knowing beforehand the frequency of words in the corpus. This is needed to so that relatively frequent words can be down-sampled — knowledge that is inaccessible in online learning circumstances faced by children. Another concern is Skip-gram's negative sampling procedure (Mikolov, Sutskever, et al.,

[3] Rabovsky and McClelland (2020) argue that statistical decomposition of language input can yield quasi-compositional representations that may be sufficient to account for the productivity of human language.

2013), where for each prediction, only a subset of possible words are sampled from the vocabulary, including the correct next word, and others drawn from a distribution that does not include the correct word. This procedure requires knowing the correct prediction before the outcome of the prediction is computed. While this speeds training and increases performance in a machine learning context, there is no evidence for such a complex memory-based process in online human learning. A number of other optimizations (such as using the current word to 'postdict' previous words in the stream) have no current basis in theories of human language processing, though this of course does not mean that such processes are cognitively impossible. There are many other potential shortcomings with regards cognitive plausibility that are not specific to Word2vec. For instance, many distributional semantic models use stop-word lists to exclude words during training (Bullinaria & Levy, 2012). These lists are language-specific and include frequent words with little semantic content. While it is straightforward for researchers to compile such lists, children are likely not born with this knowledge, and instead must learn to compile their own 'stop-word lists'. Lastly, many distributional semantic models perform heavy pre-processing of corpus data, such as stemming (i.e. morphological parsing) that splits or removes affixes from morphological complex words. This simplifies the learning problem by collapsing statistics across grammatically and morphological distinct contexts; however, children must learn to do this on their own.

There are other models, which are more closely aligned with the psychological literature, such as BEAGLE (M. N. Jones & Mewhort, 2007) and COALS (D. L. Rohde et al., 2006), which do not perform the kinds of cognitively implausible optimizations Word2Vec is known for. In fact, I consider some of these models to be as cognitively plausible as the RNN. BEAGLE, in particular, is a strong competitor, because it does not require massive memory resources to store and update a large co-occurrence matrix.

4.6 Component vs. Interaction-dominant Dynamics

The questions posed in this chapter relate to a long-standing debate in the cognitive sciences about the importance of component-dominant versus interaction-dominant dynamics for modeling the organization of human mental representations. Proponents of component-dominance argue that the semantic content that a lexical item contributes to a phrasal interpretation is context-invariant (Fodor & Pylyshyn, 1988; R. Jackendoff & Jackendoff, 2002; Pinker, 1987). This position is typically associated with the requirement that there be an inventory of discrete units (e.g. lexical items stored in the lexicon) and a central executive that performs rule-based operations on these units. On the other hand, proponents of interaction-dominant dynamics argue for a less distinct — fuzzy —- separation between units and operations, in favor of context-dependence, continuous activation values, and distributed representations (J. L. Elman, 2009, 2011; McRae et al., 1998; Spivey, 2008; Trueswell et al., 1994).

While connectionist and dynamical systems with interaction-dominant dynamics address some shortcomings of component-dominant approaches (e.g staticness, context-freeness, no account of induction), such systems have shortcomings of their own. For instance, recurrence — or any other kind of interactivity that can result in attractor dynamics — is known to impede combinatorial generalization (O'Reilly, 2001; Servan-Schreiber et al., 1991). In interactive systems, like RNNs, input features are allowed to interact arbitrarily with all other input features, and this typically results in trained models with input features that do not contribute information independently (combinatorially) to the internal representations of the network. The result is that the RNN tends to learn distinct codes for entire sequences, hampering combinatorial generalization.

What would combinatorial generalization in the RNN look like? To illustrate this, consider a hypothetical

RNN that has learned different state space trajectories (attractors) for different sentence *types* rather than for each unique sentence. Regardless of the identity of, say, the subject noun, the RNN will continue along its trajectory through its hidden state space, after having encoded the subject, as it would for any other subject. The choice of subject is temporarily saved in memory such that it does not influence future processing (the state space trajectory dictated by the sentence type). This would require the emergence of units in the hidden layer of the RNN that are exclusively reserved for memorizing semantic information, and which do not overlap with other units. Such a principled distinction between units dedicated for memory and others for processing is possible in principle, but rarely achieved with error-based learning (Baroni, 2020; Linzen and Baroni, 2021; O'Reilly, 2001; but see Tabor, 2002).

It should be noted that this thesis is not an attempt to adjudicate between interaction-dominant or component-dominant theories of human mental representations; rather, my aim is to empirically evaluate the RNN to better understand *under what conditions* interaction-dominance vs. component-dominance is most useful. In addition, I will argue that this distinction is also useful for thinking about the *development* of lexical knowledge in the RNN. Because a randomly initialized network does not have specialized units, all of the network's unit will work in tandem to produce the pattern of activation at its output. As a consequence, the initial behavior of a randomly initialized RNN will be extremely interaction-dominant. As stated previously, I think this type of processing is not inherently worse or better than component-dominant dynamics; but, importantly, I will argue that interaction-dominance is especially harmful at the earliest phase of training, in which it is important to establish the skeletal structure of the target task. For the purpose of this thesis, the target task is the distributionally-mediated extension of category-associated features (DECAF). The skeletal structure of a task can be understood as a small set of the most basic building blocks that can explain a large amount of variation in the data. Here, these building blocks are the features diagnostic of semantic category membership. To be useful to the network (i.e. to scaffold subsequent learning), these building blocks must be mapped to individual or groups of units (components) in the network such that they might act somewhat independently of each other. If building blocks are not mapped to separable components in the network, they cannot contribute information combinatorially. The overall goal of this work is to provide support for this idea in the context of learning lexical semantic representations in the RNN. The point is that if semantic category membership is encoded by simultaneously recruiting large numbers of units rather than separable components, this information would be more vulnerable to distribution shifts in the data, such as those identified in Chapter 2. The purpose of encoding information into separable components is to make them more robust against variance in the data, and to learn more permanent, generalizable, stand-alone representations that can be used outside of the system in which they first emerged.

4.7 Summary

In this chapter, I motivated several criteria that a distributional semantic model should strive to approximate if it is to be used as a model of children's construction of form-based lexical semantic representations on which the extension of category-associated features are based (DECAF). While the simple RNN satisfies the need for paradigmatic similarity, cognitive plausibility, and a basic ability to statistically tease apart category-relevant from category-irrelevant distributional semantic features, it is less clear to what extent the RNN is able to perform the kind of structured meaning decomposition needed to preserve atomic lexical relations. That is the topic of the next chapter.

Chapter 5

Basic Findings

In this chapter, I demonstrate basic findings concerning the construction of form-based lexical semantic representations in the RNN. Using handcrafted artificial datasets, and analyses of learned representations, my goal is to develop a simple and principled understanding of how lexical semantic category knowledge is actually encoded in the RNN, and how the format in which it is encoded is sensitive to the statistical structure of the input.

A key departure of the research presented in this chapter from previous work is the analysis of the input-to-hidden weights learned by the RNN, rather than the average activation patterns at its hidden layer. Previous scholars have typically examined patterns of activation at the hidden layer, where semantic and grammatical information is accumulated over the course of sequential processing. The non-contextualized lexical representations learned by the RNN at the input weight matrix have been conceptualized as mere "instructions" for how to transition between hidden layer states (J. L. Elman, 1990, 2009), but not as sources of static knowledge about individual words. On this view, the lexical representations are to be interpreted by the dynamics at the hidden layer, but not by an external system, or researcher. Being stripped from the dynamics of the rest of the system (i.e. the language processor), the de-contextualized knowledge stored at the input-to-hidden weights may not be useful in downstream tasks or in external systems (e.g. DECAF). However, this assumption has, to my knowledge, never been tested (or reported). Here, I examine this assumption empirically: Just how useful are the static representations learned at the input-to-hidden weights compared to the knowledge accumulated at the processor (i.e. hidden layer)? Under what conditions, are they most useful? Finally, how do these conditions compare to those faced by language-learning children faced with noisy, naturalistic input? If the statically available representations at the input-to-hidden weights — henceforth lexical semantic representation — acquire standalone knowledge about semantic category membership, then it could be argued that, children, too, could leverage next-word prediction to construct form-based semantic representations useful for performing DECAF.

First, I discuss the structure of the artificial languages, and the corpora that were generated from them. Second, I explain the procedure used to evaluate the lexical semantic knowledge acquired by the RNN over the course of training on these corpora. Finally, I report the results of two experiments: In the first, I examined how the RNN encodes lexical semantic information when this information is provided by neighboring items that either occur to the left *or* right of a target word, or when this information is available in both the left *and* right neighbors. In the second experiment, I consider additional probabilistic languages to zoom in on how gradations in informativity (i.e. predictability) can further influence how lexical distributional statistics are encoded in the RNN.

5.1 Artificial Languages

For the first experiment, I constructed five simple artificial languages, shown in Table 5.1. Each language generates 4-item sequences, which (i) must abide by the sequential structure defined by the language, and (ii) must contain one semantic cue in one (or two) position(s) that constrains the set of items that can occur in another position — referred to as the target position — in the same sequence. All languages use the same symbols, and differ only in regards to which position in the sequence the semantic cue occurs.

Each language is based on language N, which generates sequences of the basic structure 'A X B .', where X is always in target position. Upper-cased letters refer to categories (i.e. sets) of lexical items and the period is a punctuation symbol. For simplicity, I refer to these categories as A-words, X-words, and B-words. Because they always occur in this order, their sequential order defines the syntactic structure of language N. All languages inherit this property of language N, and differ only in their semantic category structure. The semantic category structure is determined by the relationship between X-words and neighboring words (A-words and/or B-words). Because I am primarily interested in lexical *semantic* knowledge, neighboring items only constrain the set of words that can occur in target word position, but provide no additional information about the ordering of items in a given sequence. As such, I consider knowledge of the possible sub-set of items X that may occur in target position as semantic, and not syntactic, knowledge.

The different languages differ in the way in which the predictive semantic relationship between X-words and their neighbors is realized. When A-words (left-context), or B-words (right-context) are semantically informative about X-words, I refer to them as Y-words. Whereas A-words and B-words are sampled such that they are orthogonal to the semantic category structure of X-words, Y-words are perfectly predictive of the semantic category structure. More specifically, knowing the subset that an observed Y-word belongs to, perfectly determines the subset of X-words that can occur in the same sequence. Although the relationship between the two is symmetric, I will talk about Y-words as providing a semantic cue about X-words. In language N, there is **N**o semantic cue. In language L, the semantic cue, Y, occurs in place of A, to the **L**eft of X-words. Conversely, in language R, the semantic cue, Y, occurs in place of B, to the **R**ight of X. In language O, the semantic cue Y is present in A **O**r B, but not both at the same time. Finally, in language A, the semantic cue Y always occurs both to the left of X **A**nd to the right of X. The location of the symbol Y in Table 5.1 determines the position in the sequence where the semantic cue is inserted in each language. The table also shows that there are 30 semantic categories, that subdivide X-words and Y-words into 30 disjoint subsets. Thus there are Y_1, Y_2, Y_3, ... Y_{30}, each of which is semantically associated with a corresponding subset of X-words, X_1, X_2, X_3, ... X_{30}. The semantic category structure is the same in all languages.[1] Note there are 700 items in each category; that is, the sizes of sets A, X, B, and by extension Y, are each 700.

I will refer to the relationship between X-words and Y-words as category-relevant, because it is the only relationship that determines the subset (category) of X-words and Y-words that are legal. All other relationships — between X-words and A-words, and between X-words and B-words —- are purely syntactic. This distinction can, alternatively, be expressed from a formal linguistic perspective: The category-relevant relationship between X-words and Y-words, is akin to obligatory semantic selectional constraints on nouns and verbs. Other items in a sentence are less constrained: For instance, adjuncts are optional constituents

[1] This does not apply to Language N which does not have semantic category structure.

that modify existing constituents — but, importantly, adjuncts are not selected by other constituents in the sentence. I consider A-words and B-words, but not Y-words, to be adjuncts in this sense.

For each language, I generated a corpus of 50,000 sequences by randomly sampling from the total population of sequences that are legal given the rules of a language.

Sequential Structure	
Language	**Rules**
N	S → A X B .
L	S → Y X B .
R	S → A X Y .
O	S → A X Y .
	S → Y X B .
A	S → Y X Y .
Semantic Category Structure	
	if $y_i \in Y_1$ then X → X_1
	if $y_i \in Y_2$ then X → X_2
	...
	if $y_i \in Y_{30}$ then X → X_{30}

Table 5.1: Re-write rules that determine the sequential and semantic structure of sequences in the artificial language corpora. The last four rules determine the semantic category structure and apply to each language. S denotes a sequence and the symbol → means 'is replaced by'. Upper-cased letters refer to categories of items, while lower-cased letters refer to individual lexical items. Subscripts on upper-cased letters refer to subsets (semantic categories), and subscripts on lower-case letters refer to unique lexical items. Language names are derived from the first letter of the following words: **N**o, **L**eft, **R**ight, **O**r, **A**nd. These words describe how the semantic relationship between X and Y is realized.

5.2 Methods

Next, I discuss the methods used to train, and tune the RNN on the artificial languages, and the evaluation procedure used to examine the atomicity of learned lexical semantic representations.

5.2.1 Hyper-parameters and RNN Training

Ten simple RNNs were trained on each corpus in a standard language modeling task. Learning takes place via back-propagation of the next-word prediction error, the cross-entropy between the predicted and actual next item. I identified the best hyper-parameter configuration empirically by optimizing the network's performance on the semantic categorization task discussed below. The resultant hyper-parameters are shown in Table 5.2

5.2.2 Lexical Categorization for Evaluating Lexical Representations

To examine the quality of learned lexical representations after training, I isolated the lexical representations from the input-to-hidden weights of the RNN, and used them as input to the same semantic categorization task described in Chapter 3. Crucially, because only lexical representations are used, the RNN cannot rely on knowledge stored in its recurrent or hidden-to-output weights, or knowledge that would result from combining

window size	7
hidden layers	1
hidden size	512
embedding size	512
embeddings initialization	uniform $\pm\sqrt{\frac{1}{512}}$
learning rate	0.4
optimizer	AdaGrad
batch size	64
steps	70K
non-linearity	*tanh*

Table 5.2: Hyper-parameters used to train the RNN on artificial language corpora and AO-CHILDES. I use the term ‘embeddings’ as a shorthand for lexical representations.

multiple lexical representations at its hidden layer. Without the possibility for interactions between lexical representations, perfect performance on this task indicates that all category-relevant knowledge is encapsulated within a given lexical representation, rather than spread across separate locations in the network. When semantic category knowledge is fully present in the organization of lexical — as opposed to contextualized —-representations, I consider learned lexical semantic representations to be ‘atomic’.

Lexical atomicity is discussed at a high level in Chapter 4. In this chapter, I use the term ‘atomic’ to describe a specific situation in which all knowledge relevant to the target semantic category structure (i.e. the category-relevant dependency between X-words and Y-words) is available, and can be extracted from, the RNN’s lexical semantic representations of X-words at its input-to-hidden weights. On the flipside, when category-relevant knowledge is not encapsulated in the representations of X-words, the representations of X-words can be considered ‘leaky across time steps’. Leaky representations are non-atomic because full retrieval of category-relevant information requires the combination of more than one lexical item (atom) at the hidden layer.

I made several predictions prior to training. All else being equal, atomicity, and therefore lexical semantic categorization accuracy, should be highest when there is exactly one way to encode the target semantic category structure in the network. The position of the semantic cue, Y, relative to X in the sequence, determines the number of ways it is possible to encode the semantic category structure of each language in the RNN. Because Y occurs before X in language L, there is no benefit to encode the semantic cue in the representations of X-words — doing so does not yield any benefits for the network with regards to better predicting subsequent items. On this view, I predicted that the network encodes the target semantic category structure in the representations of Y-words, where it is of use to the network’s prediction apparatus. While this constitutes high atomicity, only Y-word representations can be considered to be atomic, whereas X-words representations remain semantically vacuous. In contrast, because the semantic cue Y, in language R, is predicted by X-words, I hypothesized that the model readily encodes the the target semantic category structure in the lexical representation of X-words. Following the same logic, I predicted that a network trained on language O, where the semantic cue provided by Y is useful for prediction only half of the time, will encode the semantic category structure in the lexical representations of X-words half as fast as networks trained on language R. Regarding language A, a naive observer might suggest that the semantic category structure should be encoded in the representations of X-words fastest, because *both* neighbors are semantically related to X. With twice as much information about the target semantic category structure, the model should learn to categorize X-words twice as fast. But, in fact, there is not twice as much information; instead the *same* information is presented twice: The second presentation is redundant with the first, which should produce chunk-level representations and promote leakiness across time steps. Therefore, I predicted that learning from language A will result in lower semantic categorization performance of X-words (indicating low atomicity) compared to language R where there is no such redundancy.

5.2.3 Computation of the Balanced Accuracy

I operationalized atomicity as the amount of lexical semantic category information encoded in the lexical representations of X-words. In turn, the amount of semantic category information can be quantified using a lexical semantic categorization task. In this task, judgements about category membership are based on a similarity matrix S obtained by computing all pairwise similarities[2] between X-word representations. To obtain the model's learned Lexical representations for one set of words, I retrieved the vector that connects the input unit corresponding to a single word with the hidden layer. Each similarity in matrix S was used to make a 'same vs. different' judgment within a signal detection framework, tested at multiple similarity thresholds (r = 0.0 to 1.0 with step size 0.001) to determine the threshold for maximum accuracy. If two words with indices i and j belong to the same category, and if $S_{ij} > r$, a hit is recorded, whereas if $S_{ij} < r$, a miss is recorded. On the other hand, if the two words do not belong to the same category, either a correct rejection or false alarm is recorded, depending on whether $S_{ij} < r$ or $S_{ij} > r$. At each threshold, I computed the balanced accuracy by taking the average of sensitivity and specificity. The measure of interest is the balanced accuracy at the similarity threshold which yielded the highest value. I used this process to compute a balanced accuracy score for each model, at each evaluation time point. Chance-level performance on this task would produce a balanced accuracy of 0.5.

The balanced accuracy is appropriate because it eliminates bias due to the unbalanced distribution of correct 'same' and 'different' judgements - the vast majority of X-word pairs do not belong to the same category. Because the balanced accuracy is the average between the sensitivity and the specificity, it measures the average accuracy obtained from both the minority and majority classes. This quantity reduces to the traditional accuracy if a classification accuracy is identical for either classes. But, if the high value of the traditional accuracy is due to taking advantage of the distribution of the majority class, then the balanced accuracy will decrease compared to the traditional accuracy.

It is worth noting that the categorization task does not involve the prediction of category labels. There are neither category labels in the training data, nor is the RNN trained to words to a category label. In order to successfully reconstruct the target category structure, the RNN must acquire lexical representations such that their similarity is higher for same-category members than members that belong to different categories. In this sense, the categorization judgement is entirely similarity-based. Importantly, the similarity structure of the learned internal representations is an indicator of the overall organisation of the representational landscape learned by the model.

The categorization task was performed twice: Once with lexical representations of X-words as input, and a second time using contextualized representations of X-words as input. The former evaluates the atomicity of individual lexical semantic representations, while the latter evaluates how well the semantic category structure is represented in the contextualized (i.e hidden layer) representations of X-words. To obtain contextualized

[2] As a measure of similarity between two vectors, I used the cosine of the angle between them.

representations for one X-word, I computed the hidden states given all sequences in the training corpus that end with the X-word, and then averaged the resulting hidden state vectors. These vectors not only capture the lexical representation of X-words, but how they interact with the items that have occurred before them in the corpus. Importantly, all networks should achieve high performance when classifying contextualized representations, but only models with atomic lexical semantic representations can achieve a high balanced accuracy when the representations of X-words are obtained at the input weights (i.e. not contextualized).

Finally, it should be noted that the procedure to calculate the balanced accuracy (which I term the 'semantic categorization task') should not be mistaken for an actual task that a person might perform. I use the term 'task' more generally to refer to a diagnostic procedure, rather than a real-life situation faced by a language learner. Put differently, this task is nothing more than a procedure used by the researcher to quantify the amount of lexical semantic knowledge stored in a particular location in the network. When applied to lexical representations, this procedure allows quantification of lexical atomicity; when applied to contextualized representations, the same procedure allows quantification of how much lexical semantic knowledge has been accumulated at the hidden layer. That said, I use the resulting performance in the semantic categorization task as a proxy for the performance of the distributionally-mediated extension of category-associated features (DECAF). If semantic categorization is poor, then, it goes without saying that the same lexical representations will likely not be of much use for performing DECAF. In sum, the categorization accuracy is an indirect measure of the success that children would have if using the lexical semantic representations learned by the RNN for extending learned meanings to novel words.

5.3 Results

The results of the semantic categorization analyses of networks trained on languages L, R, O, and A are shown in Figure 5.1. The left panel shows how the balanced accuracy changes with training for lexical representations of X-words, and the right panel illustrates how the balanced accuracy changes with training when contextualized representations of X-words are used. To organize the discussion, I will focus on two distinct questions: First, I discuss what the results reveal about encoding semantic cues that occur either to the left or to the right of a target word (language L vs. R). Second, how does redundant semantic information (language O vs. A) influence learning of lexical semantic information?

5.3.1 Left vs. Right Asymmetry

The left panel of Figure 5.1 shows a stark difference in semantic categorization performance between models trained on language L versus language R. When trained on language R, in which X-words are useful for predicting the set of items that can occur next, performance reaches ceiling-level after approximately 30,000 training steps. In contrast, the performance of models trained on language L, where X-words are not useful for predicting *upcoming* words, remains at chance for the entire duration of training. Does that mean the networks trained on language L do not encode the semantic category structure? No; the semantic category structure is simply not encoded in the lexical representations of X-words. The right panel of Figure 5.1 shows that when contextualized representations of X-words are used, both models reach near-perfect categorization accuracy after approximately 1,000 training steps. This illustrates that, models trained on language L encode the target semantic category elsewhere in the network. Presumably, the semantic category structure is encoded in the representations of Y-words, which precede X-words in language L.

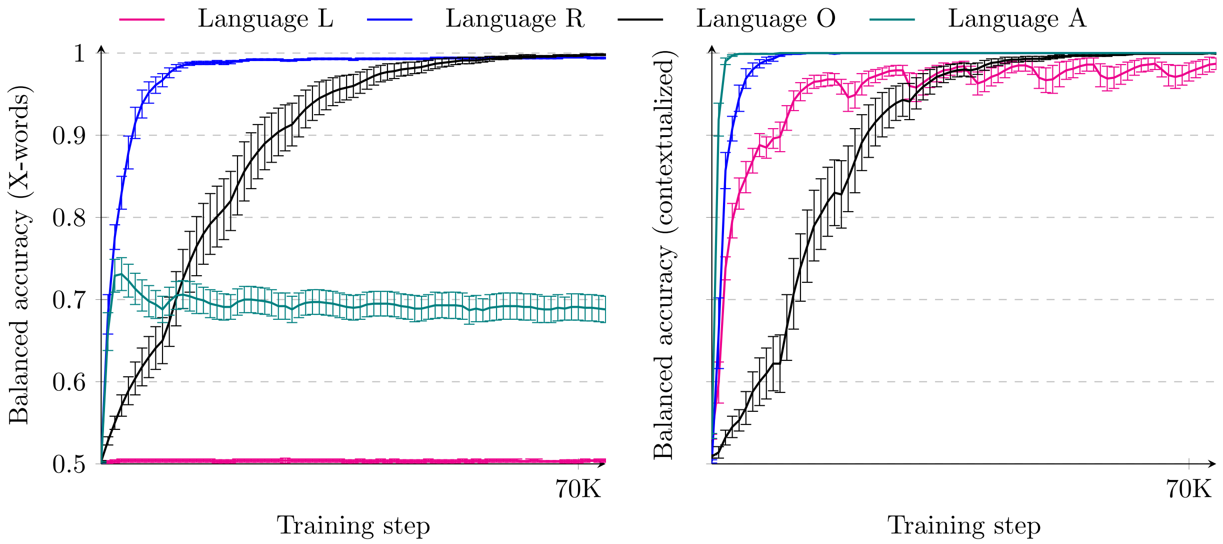


Figure 5.1: Lexical semantic categorization performance as quantified by the balanced accuracy, across training. Each line represents the average performance across 10 RNN simulations, error bars indicate 95% confidence intervals. The balanced accuracy measures the degree to which the internal organization of the RNN corresponds to the externally-defined semantic category structure. The two panels differ only in whether lexical (left panel) or contextualized (right panel) representations are input to the categorization task.

5.3.2 Across-Target Redundancy

To understand the effect of redundancy on learning of lexical semantic knowledge, I compared the balanced accuracy of models trained on language O and A. Because sequences in language O never contain more than one semantic cue Y, I refer to it as the no-redundancy condition. In contrast, because the semantic cue Y is duplicated – occurs before (A)nd after an X-word — in language A, I refer to it as the full-redundancy condition.

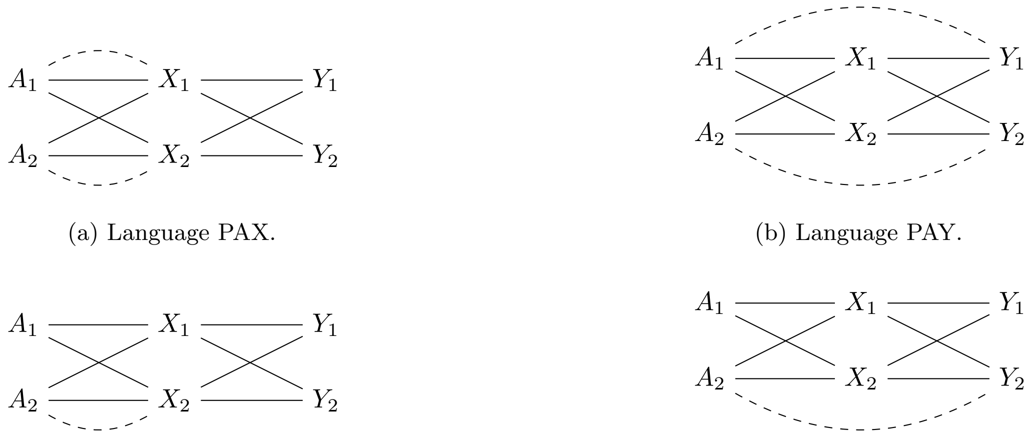
At the end of training, the balanced accuracy for X-words in the no-redundancy condition (language O) is at ceiling. Perfect categorisation means that X-word representations are organized into clusters that perfectly corresponds to their target categories. As a consequence, I concluded that RNNs in this condition have acquired atomic lexical semantic representations of X-words. However, in the full-redundancy condition (language A), in which items in the left and right contexts of X-words provide fully redundant information about the target semantic category structure, semantic categorization accuracy asymptotes at no more than 0.7 at the end of training — despite the fact that the semantic relationship between X and Y-words was identical. For RNNs trained in this condition, I concluded that the target semantic category structure was only partially captured in the organization of lexical representations of X-words. In the presence of neighboring items that co-predict the target semantic relationship, information that would otherwise have been represented atomically, is instead spread (i.e leaked) from the representation of X-words to adjacent items that contributed the redundant information.

5.4 Additional Languages

Given that RNNs trained on language A learned semantically impoverished lexical representations, I explored additional input conditions that might yield similarly poor results. I hypothesized that the amount of redundancy in the input surrounding X-words is negatively associated with atomicity of learned lexical semantic representations of X-words. Using language A as a starting point, I created four additional languages. Each new language preserves the category-relevant semantic relationship between X-words and Y-words, which, in principle, preserves the ability of the RNN to achieve perfect performance on the downstream semantic categorization task. However, each differs from language A in that each A-word in each language no longer predicts a subsets of Y-words (i.e. the semantic category), but predicts a distinct X-word or Y-word. In other words, the relationship between A-words and X-words or Y-words is one-to-one, rather than one-to-many. Additionally, I created 4 conditions, by varying two factors in a 2-by-2 design. This design is shown in Figure 5.2, which illustrates tiny snapshots of the structure of each new language. The left panels illustrate languages PAX and DAX, where the letters 'AX' indicate that A-words predict X-words, and the right panels illustrate languages PAY and DAY, where the letters 'AY' indicate that A-words predict Y-words. The top and bottom panels separate languages by whether this relationship is **P**robabilistic or **D**eterministic; the relationship between A-words and upcoming items in languages PAX and PAY (top panels) is probabilistic, and in languages DAX and DAY (bottom panels) it is deterministic.

The final difference between languages PAX, PAY, DAX, and PAY, and language A is that the redundant information provided by A-words, X-words, and Y-words is not all-or-nothing but graded. To vary the amount of redundancy, I introduce the parameter α. This parameter functions differently depending on whether the language is probabilistic or deterministic: In all languages, the number of A-words corresponds to the number of X-words, and the number of Y-words. This means it is possible to assign each A-word to exactly one X-word or Y-word. In the probabilistic languages PAX and PAY, α determines the probability that an A-word is chosen (during corpus creation) based on this assignment. When α is 0.5, an A-word is chosen based on its assignment to an X-word (PAX) or Y-word (PAY) half of the time, and the other half of the time, an A-word is chosen randomly from the full set of A-words. When α is 1.0, every A-word perfectly predicts exactly one upcoming X-word (PAX) or Y-word (PAY). In the deterministic languages DAX and DAY, α determines the proportion of A-words that are assigned a distinct X-word (DAX) or Y-word (DAY). Once assigned, an A-word must always co-occur with its assignee in the corpus — this is what makes the relationship deterministic.

The purpose of training and evaluating the RNN on these additional languages is threefold: First, by setting α to values below 1.0, I could examine how the RNN responds to intermediate levels of redundancy, as opposed to the all-or-nothing comparison between languages O and A reported above. A further advantage of setting α to an intermediate value between 0.0 and 1.0, is that the additional languages better approximate natural languages, where lexical items are almost never perfectly redundant with other items in the same sentence. Second, by separately examining conditions where A-words are either probabilistically or deterministically related to upcoming items, I can better understand how the presence of counterfactual examples (the same A-word does not always predict its assignee in probabilistic languages PAX and PAY) influences the course of learning lexical semantic representations. Third, by separately examining languages where A-words predict X-words as opposed to Y-words, I was able to examine how sequential structure interacts with redundancy to impact learning. On the one hand, when A-words predict X-words, the items participating in the redundancy are adjacent (A predicts X, which in turn predicts Y); on the other hand, when A-words predict Y-words, the items participating in the redundancy are non-adjacent, because the lexical relationship between A and Y-words skips the intervening X-word slot. I will refer to the latter type of dependencies as 'across-target' dependencies. I predicted that the presence of across-target dependencies would impede the formation of atomic lexical semantic representations of the intervening items, X-words.



(a) Language PAX.

(b) Language PAY.

(c) Language DAX.

(d) Language DAY.

Figure 5.2: Schematics illustrating the differences between languages PAX, PAY, DAX, and DAY. Solid lines indicate category-relevant relationships between X-words and Y-words. Dashed lines indicate category-irrelevant relationships between A-words and X-words, and A-words and Y-words. For brevity, I only consider a minuscule selection of all possible relationships per language: 2 A-words, 2 X-word from the same semantic category, and 2 Y-words from the same semantic category. The statistical properties illustrated here for only one small fraction of one semantic category apply to all other semantic categories in the same language. In language PAX and PAY, α determines the probability that an A-word is chosen that perfectly predicts an upcoming X-word or Y-word. In language DAX and DAY, α determines the proportion of X-words or Y-words that are perfectly predictable given the A-word in the same sequence. In the bottom panels, α is 0.5 because half of the X-words or Y-words have a category-irrelevant relationship with one A-word.

Consistent with my previous observations, all models trained on languages PAX, DAX, DAX, and DAY, eventually achieve a perfect balanced accuracy when the input to the semantic categorization task consists of contextualized representations. For brevity, I therefore only report the results of inputting non-contextualized lexical representations of X-words (i.e target words) to the downstream categorization task. The balanced accuracies for all four languages are shown in Figure 5.3.

5.4.1 Languages PAX and PAY

The results of training RNNs on languages PAX an PAY are shown in the top panels of Figure 5.3. I trained 10 RNNs on 6 versions of each (α is either 0.0, 0.2, 0.4, 0.6, 0.8, or 1.0), where each version varies in the probability that an A-word is perfectly predictive of an upcoming X (top left panel) or Y-word (top right panel). I observed that for all languages with α below 1.0, ceiling-level categorization tends to be achieved at about the same time during training (5-10K steps). This was somewhat surprising, given that I predicted that semantic categorization accuracy would be inversely proportional to the amount of redundancy in the data. On the contrary, the RNNs exhibited almost no drop in atomicity in face of less-than-perfect levels of redundancy (i.e. α is lower than 1.0). This is especially true for language PAX, where networks trained with α of 0.8 are nearly indistinguishable from networks trained on input with much lower levels of α. For language PAY, however, networks trained with α of 0.8 require approximately double the number of training steps to reach ceiling performance. What is striking, is that a qualitative shift in the learning trajectory occurs when α is exactly 1.0: When redundancy between neighboring items is perfect, networks learn an extremely impoverished lexical semantic organization, as evidenced by the observation that even after 70K training steps, the balanced accuracy has converged far from ceiling-level.

5.4.2 Languages DAX and DAY

The results of training RNNs on languages DAX an DAY are shown in the bottom panels of Figure 5.3. I trained 10 RNNs on 6 versions of each (α is either 0.0, 0.2, 0.4, 0.6, 0.8, or 1.0), where each version varies in the proportion of X or Y-words in the corpus that perfectly predict a specific A-word in the same sequence. For instance, when α is 1.0, all A-words in the corpus are perfectly predictive of an upcoming X (bottom left panel) or Y-word (bottom right panel). In contrast, when α is 0.5, only half of all A-words perfectly predict an upcoming item. What distinguishes the two deterministic languages (DAX and DAY) from their probabilistic counterparts is that each occurrence of an A-word is either perfectly predictive of or not predictive about an upcoming item — with no gradation between these two extremes. The results closely resemble those obtained with languages PAX and PAY. However, language DAY appears to be more difficult for the RNN when redundancy is close to maximal: When α is 0.8, semantic categorization accuracy still has not converged after about 35K steps, by which time RNNs in all other conditions have reached ceiling-level performance.

5.5 Summary

Given the results presented in this chapter, I draw the following conclusions: First, the uni-directional left-to-right processing of input strings by the simple RNN severely constrains how semantic information is encoded in the lexical representation at its input-to-hidden weights. When a semantic cue is provided by items that precede a target word in the linear ordering of items in the input, then the semantic information will be encoded in the lexical representations of the preceding items as opposed to the target word. This constraint is likely maladaptive in situations where, say, target words are nouns and their left neighbors are adjectives that provide semantic information about the nouns they co-occur with. A simple RNN which processes such strings left-to-right will encode the semantic relationship between adjectives and nouns in the lexical representations of *adjectives* and not nouns. Consequently, when the network's lexical knowledge about nouns is probed, much of the relevant semantic information is simply not available.

Second, not only does the relative position of the semantic cue influence learning of lexical representations, but the number of locations where the semantic cue occurs: When the semantic cue Y only occurs once per sequence, but alternates between the left and right position (language O), the RNN eventually encodes all semantic category-relevant information in the lexical representations of the target words. Given the left-to-right processing asymmetry of the simple RNN, it accomplishes this, presumably, by taking advantage of the semantic cue when it occurs in the right but not the left position. However, when the same semantic information is made available in both the left and right position (language A), then the RNN fails to encode the target semantic category structure in the representation of the target words. Presumably, the RNN recruits additional locations in the network, such as the recurrent weights and other lexical representations, to fully capture the target semantic category structure. I argue that this kind of dynamic is maladaptive because information about semantic category membership is not correctly aligned with the lexical items which are causally related to that information (in the symbolic program used to generate the language).

Third, the evaluation of RNNs trained on languages PAX, PAY, DAX, and DAY with partial redundancy between A, X, and Y-words demonstrates that the RNN is surprisingly good at 'ignoring' category-irrelevant co-occurrence relationships when they are not the most reliable source of information about the underlying semantic category structure. Even under conditions where left neighbors of X-words predict an upcoming

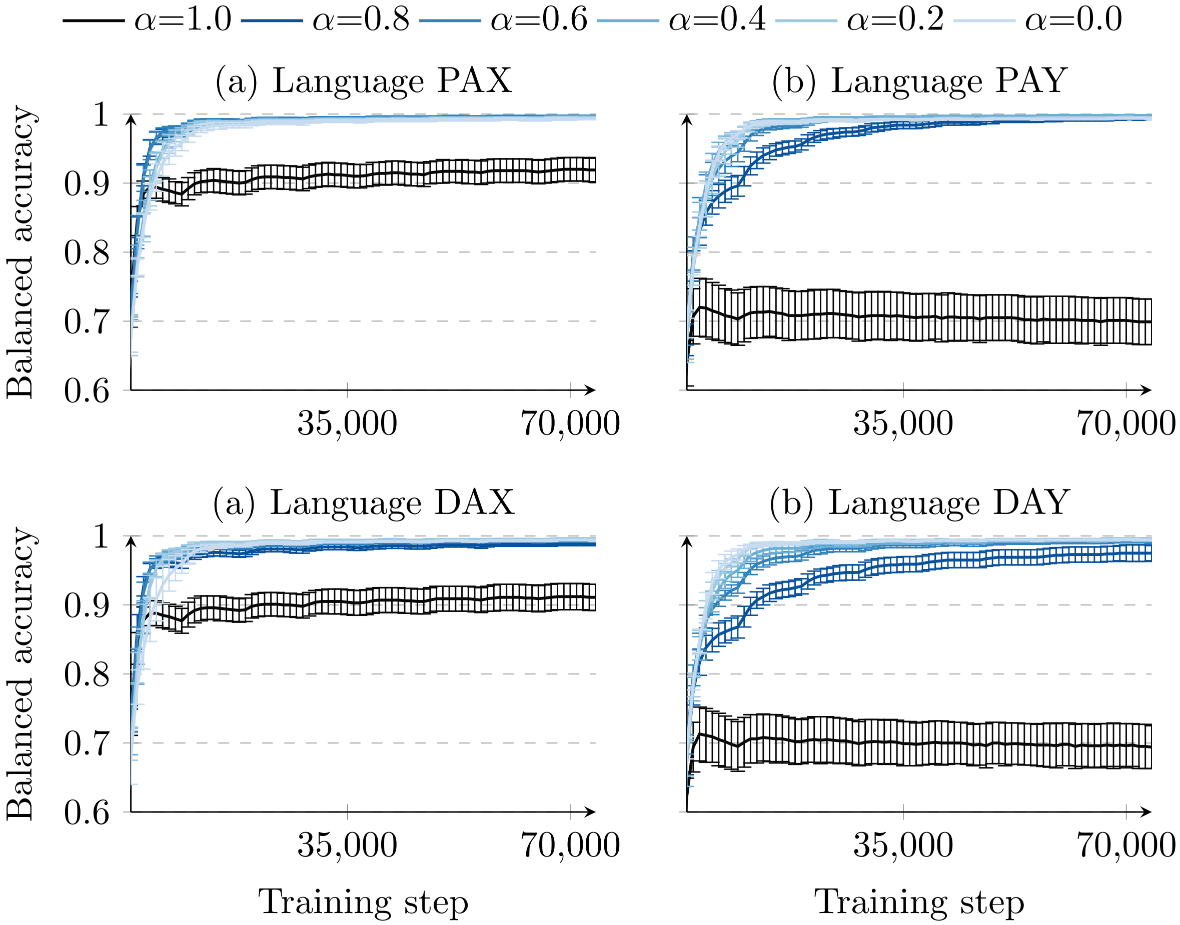


Figure 5.3: Lexical semantic categorization performance as quantified by the balanced accuracy (y-axis), across training (x-axis). Models were either trained on language PAX (a), PAY (b), DAX (c), or DAY (d). Each line represents the average performance across 10 RNN simulations, error bars indicate 95% confidence intervals. The balanced accuracy measures the degree to which the internal organization of the RNN corresponds to the externally-defined semantic category structure. I used only lexical (non-contextualized) representations as input the categorization task.

item 80% of the time, the RNN is able to disregard such regularities in favor of the more reliable — and in this case, category-relevant — information source. It can be said that, under these conditions, the RNN follows the optimal encoding strategy: Exclusively encode the most reliable source of information and ignore any other less reliable (possibly redundant) source. I might further interpret this behavior of the RNN as learning lexical representations of X-words that are invariant under permutation of A-words. The ability to learn representations that are selectively invariant to irrelevant features of the input is key requirement for systems to discover abstract structural relations and to predicate those relations (L. A. A. Doumas et al., 2008).

Fourth, the experiments with languages PAX, PAY, DAX, and DAY in the full-redundancy condition revealed that when two sources of information about the same underlying semantic category structure are equally reliable, the RNN behaves qualitatively different compared to a network trained on input where one source is more reliable than another. In the full-redundancy conditions, when A-words perfectly predict an upcoming item, X-words do not contribute any additional information, and therefore have no value to the language modeling objective used to train the RNN. Because A-words occur *before* X-words, the simplest encoding strategy is to encode the information provided by both A and X-words in the lexical representations of A-words alone. Because A-words always occur first in every sequence, the RNN can access all the information it needs for predicting upcoming items right away, rather than waiting for part of that information to become available later when processing an X-word. Put differently, there is no incentive for the RNN to differentiate the lexical representations of X-words, as they do not contribute any information that has not already become available when processing an A-word. This would explain the poor balanced accuracy that results when inputting representations of X-words into the lexical semantic categorization task. When these results are combined with insight from previous studies (D. L. Rohde & Plaut, 1999; Servan-Schreiber et al., 1991), a clear picture of RNN learning dynamics emerges: The contribution of lexical representations to the pattern of activations at the hidden and output layer is to further differentiate the set of predicted next-words *relative to information already available at the hidden layer.*

The key takeaway message from this chapter is the following: The RNN does not learn atomic lexical semantic representations when trained on data where neighboring items provide redundant semantic information about the target word. This occurs whenever providers of redundant information occur before the target word in a sequence. Additionally, I have identified two failure modes: Either providers of redundant information (i) always provide exactly the same information, or (ii) always provide more information than that which is made available by a target word concerning the prediction of upcoming category-relevant items. In both cases, the RNN learns a shortcut that effectively excludes target words from being semantically differentiated.

Broadly, the demonstrations in this chapter point to an important conclusion about how artificial neural networks learn: Although the underlying structure of the data (e.g the semantic category structure) remained identical across all simulations, the RNN proved sensitive to how this structure is statistically realized. It simply does not suffice that some target structure be available in the data; subtle details can potentially influence how this structure is encoded. In sum, this means that the degree to which lexical atomicity is achieved does not depend only on *whether* the target semantic category structure is made available, but, more importantly, on *how* it is made available. My hope is that this work will ultimately open the door to comparisons of systems governing learning in machines and human brains at the algorithmic level.

Finally, it should be noted that all of the experiments reported in this chapter were repeated with a more advanced version of the simple RNN, namely the Long Short Term Memory (LSTM), which uses multiplicative gates to control the flow of information across time steps (Hochreiter & Schmidhuber, 1997). I used the same hyper-parameters when training the LSTM that were identified when tuning the simple RNN, and found that the LSTM behaves qualitatively identical to the simple RNN on all languages and conditions. This suggests the issues demonstrated in this chapter are broadly applicable to the class of RNN models, rather than a particular architecture.

Chapter 6

Semantic Property Inheritance (SPIN) Theory

The purpose of this chapter is to compile and formalize the basic findings presented in the previous chapter, culminating in a simple set of language-agnostic principles, which I will refer to as 'Semantic Property Inheritance' (SPIN) theory. The basic motivation behind SPIN theory is to identify the conditions in which next-word prediction can construct atomic lexical semantic representations. In addition, I examine the lexical atomicity of the RNN trained on child-directed input, and use SPIN theory to derive a training strategy that can promote the formation of more atomic lexical semantic representations.

6.1 The Credit Assignment Problem

Any system that learns from data by updating its parameters based on some error signal suffers from the so-called 'credit assignment problem': Which of the — potentially hundreds or thousands of – parameters should be updated given a scalar error-signal produced at the output layer? The RNN is special in this regard because it suffers from an addition 'temporal' credit assignment problem: Which parameters *at which time step* are responsible for the observed error signal? This problem is not addressed by back-propagation, but is instead cleverly avoided by it. The error signal is simply back-propagated as far back in time as is possible given the capability of the system and the modeling parameters. Put differently, back-propagation, simply spreads the error top-down to all units that were involved in generating an output. Credit is simply assigned to all lexical representations that participated in the feed-forward pass used to output a next-word prediction.

While back-propagation-through-time (Rumelhart et al., 1986; Werbos, 1990) is an immensely successful algorithm, and used widely for a large variety of sequence-based prediction tasks, it is not necessarily ideally suited for linguistic data, where causal relations tend to involve only a small subset of all possible relations in the input. For instance, although the number of category-relevant relations per sentence is a small fraction of the possible relations, backpropagation-through-time establishes a relationship between every item that has occurred in the temporal window considered by the backward pass. While the LSTM is slightly more advanced in this respect, in that it is able to better direct the error flow across time steps, the LSTM must learn how to do this given the data which it is given. Needless to say, the data itself does not necessarily reveal how credit should be assigned (especially in the perfect-redundancy conditions reported in the previous chapter). This is supported by the results of the artificial language simulations: Regardless of the architecture that was used, the simple RNN and the LSTM both failed to learn atomic lexical semantic representations in all perfect-redundancy conditions. This suggest that the results are not due to the quirks of a particular architecture, but due to the under-determination inherent in the data and the absence of linguistic biases in the architecture that would assign credit where it is due.

Given the lack of linguistic constraints on the backpropagation procedure in the RNN, is it possible that the network can learn to distinguish a category-relevant from a category-irrelevant cue when each cue provides perfectly redundant information? The simple answer is no. In the full-redundancy conditions (when α is 1.0), this distinction is under-determined, which means that it requires additional information such as knowledge about the causal structure of the language generator (which is inaccessible). Needless to say, the full-redundancy condition is an extreme situation, and which is not likely to be encountered by language-learning children. As children are exposed to gradually more language data, it becomes increasingly likely that two previously competing — and perfectly redundant — cues will eventually be teased apart. For example, during early acquisition, a child might happen to hear *doll* exclusively in the context of *your*, and never hear *doll* in any other context such as '*my doll*' or '*that doll*'. A purely statistical analysis of their input would result in a representation of *doll* that is inextricably linked with the determiner *your*: Distributional features of *doll* will become partially active when the determiner *your* is heard or otherwise activated by the learner. This may be adaptive for navigating the idiosyncratic language environment of our hypothetical child, but may pose problems down the road when attempting to develop a more mature understanding of the structure of their language. While they will eventually be able to tease apart these two words, a full re-organization of their initial representational landscape may be more difficult when starting with highly entangled representations.

The (temporal) credit assignment problem is a well known and difficult problem, with no straightforward solution. Rather than tackling this problem by innovating novel architectures or algorithms, the remainder of the paper explores the possibility that we can take advantage of knowledge about the statistical structure of natural language to bias a network towards learning more atomic lexical semantic representations. To do so, we need a clear understanding of how the statistical properties of language input influence the lexical semantic organization of the RNN.

6.2 SPIN Theory

Given the results of the artificial language learning simulations and the temporal credit assignment problem, is it possible to learn atomic lexical semantic representations when training the RNN on natural language? In other words, could the RNN be used to supply lexical semantic representations to an external system for performing DECAF? To answer this question, I propose a theory that specifies the conditions in which the learning of atomic form-based lexical representations is possible.

> To guarantee the formation of atomic lexical semantic representations of a set of target words in a unidirectional RNN that processes input left-to-right, the training data
>
> (1) must have sufficient information about the target semantic category structure present in the right contexts of target words, and
>
> (2) must not have information present in the left contexts of target words that (partially or perfectly) predicts the semantic category membership of target words.

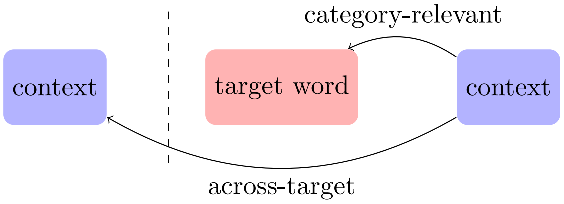


Figure 6.1: A schematic illustrating requirement (2). Category-relevant information provided by a right-context must not be predictable prior to the occurrence of a target word.

I term this ‘Semantic Property Inheritance’ (SPIN) theory, to highlight the idea that distributional semantic properties may be inherited by non-target words that are irrelevant to the semantic category structure, but which may nonetheless be correlated with it. Requirement (1) is not difficult to fulfill, as there are plenty of semantic category relevant cues in natural language data. I showed this extensively in the results of the hidden layer representations learned by the RNN trained on AO-CHILDES in Chapter 3. I am not the only one in this regard — plenty other researchers have obtained similar evidence (Asr et al., 2016; Riordan, 2007; Unger et al., 2020). The focus of this paper is, instead, on requirement (2). This requirement is very strict, and likely excludes a majority of natural language corpora. The potential for incidental — as well as category-relevant — statistical relationships between pre-target (left-context) and post-target (right-context) words is simply too large given the immense vocabularies and complexity of natural language. Even if the window over which an RNN is allowed to track co-occurrence were restricted to the size of a sentence, it would still encounter numerous sentences with numerous across-target relationships (i.e. relationships between the right and left contexts of a target word).

To illustrate requirement (2), consider Figure 6.1. By disallowing across-target relationships, the error signal that is back-propagated across time steps cannot stick to the lexical representation of a left-context word. Requirement (2) essentially imposes a ‘bottleneck’ at the target-word, which eliminates any statistical association between words in the left and right context. In the schematic, the bottleneck is illustrated by the dashed vertical line. With this bottleneck in place, a left-context can no longer inherit the semantic properties of a target word.

For stylistic reasons, I will often refer to ‘across-target dependencies’ simply as ‘redundancy’. Unless otherwise noted, ‘redundancy’ means ‘across-target redundancy’, the statistical association that occurs when both a left-context and a target word reliably predict a category-relevant word in the right-context of the target word.

6.2.1 Fragmentation

To simplify the discussion, I will refer to a violation of requirement (2) as ‘fragmentation’ of the target semantic category structure.[1] In the statistical sciences, the term fragmentation is used to refer to situations in which information that is causally attributable to one item is represented as an interaction over multiple

[1] The target semantic category structure is a set of assignments mapping a content word (i.e. noun, verb, or adjective) to exactly one semantic category. Certainly, there exist more complicated category structures, such as hierarchical and non-disjoint structures. The theory developed here generalizes to hierarchical semantic category assignments, but non-disjoint category assignments are out of the scope of my proposal. In a non-disjoint structure, a subset of members of category A and B may also be members of category C. In the target category structures used in this work, each target word belongs to one and only one semantic category.

items, some of which are not causally - but, nonetheless, statistically - related (Jakulin & Bratko, 2003). I adopt this term here to describe co-occurrence data with multiple layers of sub-category distinctions that make it difficult to recover a more general, superordinate-level category structure. As a corpus-analytical tool, fragmentation produces a numeric value between zero and one, with values close to zero indicating little to no fragmentation, and values closer to one indicating high fragmentation. An in-depth discussion of fragmentation, how to compute it, and the amount of fragmentation of nouns in child-directed input is provided in the next chapter, Chapter 7. I will use the term fragmentation to talk about individual categories, collections of categories (i.e. a category structure), as well as the contexts that cue a category. I will use the expressions 'a category is fragmented' and 'the contexts that cue a category are fragmented' interchangeably.

In general, fragmentation of left-contexts is problematic for the following reason: When the fragmentation of the left-contexts in which members of some lexical category (e.g. nouns) occur is high in a given corpus, those same contexts are at greater risk of participating in maladaptive across-target dependencies. That is, when fragmentation is high, redundancy, too, tends to be high. As stated above, when redundancy is high, the left-contexts inherit the semantic properties of the target word — in effect, robbing the representation of the target word semantic information about itself. This is a classic failure of temporal credit assignment, at least from the perspective of learning atomic lexical semantic representations.

Fragmentation is not specific to left-contexts; both the left and right contexts of target words can be fragmented with respect to some target semantic category structure. Fragmentation of left-contexts is a violation of requirement (2), but fragmentation of right-contexts is one possible violation of requirement (1). While I am primarily concerned with requirement (2), it should be noted that fragmentation of right-contexts is problematic for a reason unrelated to temporal credit assignment: At the limit, fragmentation of right-contexts eliminates category-diagnostic distributional signals by replacing regularities at the target category level with sub-regularities within the target category. This is consistent with findings from Chapter 4; in order to learn atomic lexical semantic representations, right-contexts must provide distributional information about a target word's category membership, and left-contexts must not provide any information already provided by right-contexts about category membership. These observations can now be restated more generally: Fragmentation of right-contexts impedes learning of semantic category knowledge, and fragmentation of left-contexts impedes the atomicity of learned lexical semantic representations. The former, a violation of requirement (1) is intuitive, while the latter, a violation of requirement (2), is far from obvious, and, to the best of my knowledge, has not been documented by previous studies. For this reason, all subsequent discussion about fragmentation concerns left-contexts, unless otherwise noted.

When trained on data without fragmentation of a given target category structure, the RNN would trivially acquire atomic lexical semantic representations of target words because there would be no ambiguity about which distributional signal is relevant for making a category distinction — there is exactly one signal. However, when fragmentation is high, then there are additional distributional signals causally unrelated to the target category structure but correlated with it, and, as a consequence, lexical atomicity is no longer guaranteed. As will be discussed in an upcoming section, such correlations are likely extremely common in natural language. In the presence of high fragmentation, the corpus is rife with additional distributional signals due to lexical or sub-category regularities within one ore more target categories. For instance, many lexical items can be classified into more than one semantic category (e.g. APE, MAMMAL, ANIMAL, LIVING-THING), and different distributional regularities may be associated with each level in the category hierarchy. When a corpus contains these sub-regularities, the RNN cannot know which category a given distributional signal is causally related to. The reason is that the language modeling objective promotes a carving up of the RNN's representational landscape into the smallest sub-category distinctions licensed by the data: If *gorilla* and *elephant* are the only two words that occur after *big* in some hypothetical training corpus, they are encoded as a distinct sub-category within MAMMAL separate from other members of MAMMAL. In this case, the RNN has learned to represent *gorilla* and *elephant* differently than the rest of the mammal words. The goal of a language model is to differentiate as many lexical items as possible — at the expense of category-based similarities — in order to predict next-words as accurately as possible. When present, the RNN language model will exploit such fragmentation at the expense of atomicity.

In a nutshell, fragmentation in the co-occurrence data provides the RNN with shortcuts (e.g. across-target dependencies) for predicting upcoming words. This foregoes the bottleneck at the target word needed to capture category-relevant statistics in the representation of the target word. If the left-context provides information useful to next-word prediction *prior to the observation of the target*, the target word is at risk of becoming disposable (i.e. semantically vacuous). Without a bottleneck in place at the target word (i.e. requirement (2)), the learned lexical semantic representations may be of little use for performing the distributionally-mediated extension of category-associated features (DECAF).

6.2.2 The Counterbalancing Requirement

What can be done to promote learning of atomic lexical representations when faced with highly fragmented left-contexts? Requirement (2) is essentially a restatement of the problem, rather than a solution. It simply states that in order to acquire atomic lexical representations in the RNN, left-contexts that appear in the training data must not be fragmented to begin with. To resolve the ambiguity about which of multiple redundant cues are causally related to the target semantic category structure, the RNN must either be provided with more (e.g. extra-linguistic) information to disambiguate their causal origin, or not be provided redundant cues in the first place. The latter solution is conceptually equivalent to requirement (2). By training an RNN on data where left-contexts are minimally fragmented, the RNN cannot use any information other than what must be encoded in the lexical representation of a target word for next-word prediction. Doing so would necessarily concentrate semantic category information at the target word. If, on the other hand, a left-context were to provide such information, the RNN would be able to (partially) ignore the target word and instead (partially) rely on the left-context. To avoid the latter situation, an RNN must be trained on data where the left-contexts in which same-category targets occur are drawn from the same distribution. In other words, the distribution of left-contexts must be 'counterbalanced' across same-category targets. For this reason, I will refer to requirement (2) as the 'counterbalancing requirement'.

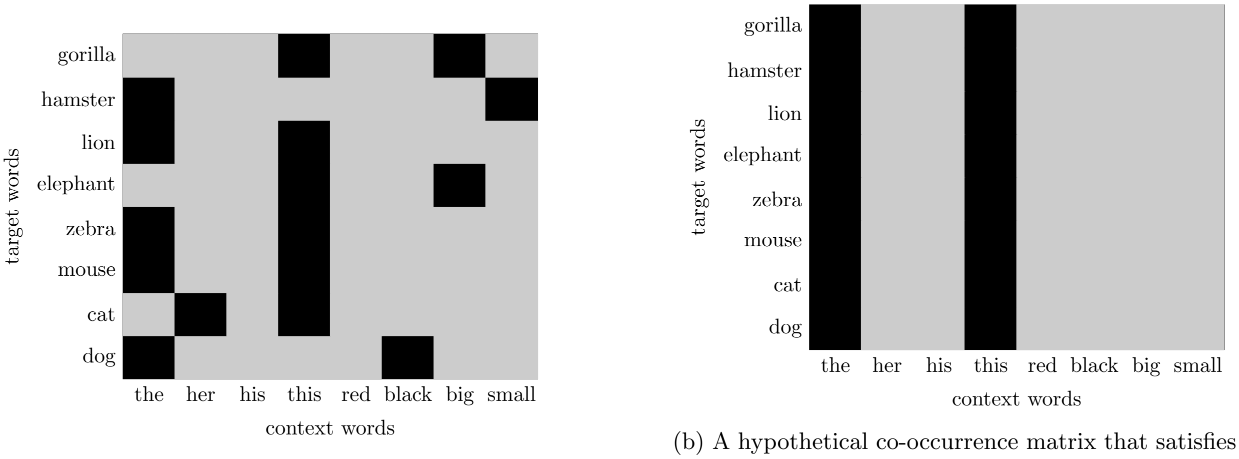
When the input is not perfectly counterbalanced in the way proposed above, an RNN language model will exploit distributional signals not relevant to the semantic category structure to minimize prediction error. For instance, if the data used to train an RNN is organized like that in the left panel of Figure 6.2, the lexical representation for *gorilla* will not resemble other representations of words in the same category (MAMMAL), such as *hamster* and *lion*. In this hypothetical data, the word *gorilla* more closely resembles the word *elephant*, than other members of MAMMAL by virtue of occurring after *big*. The co-occurrence pattern shared by *gorilla* and *elephant* but not by other members of MAMMAL will likely result in the learning of a sub-category of MAMMAL. Similarly, the RNN will learn to represent *hamster* distinctly from all other MAMMAL words because it is the only one that occurs after *small*. On the one hand, these sub-category signals enable the RNN to make more accurate next-word predictions; on the other hand, they promote the formation of fragmented, non-atomic lexical semantic representations by concealing the superordinate semantic category structure. In many cases, learning finer-grained categories is not necessarily undesirable, but fragmentation may divide target semantic categories into conceptually incoherent, idiosyncratic groupings. Fragmentation is just as likely to highlight plausible groupings such as SMALL-MAMMAL vs. LARGE-MAMMAL, as it is to distinguish *dog* and *cat* which form a highly coherent sub-category (i.e. PET). In other words, there is no guarantee that distributional statistics respect semantic category boundaries in any principled manner. A distributional learning system, therefore must be somewhat robust against idiosyncrasy of this variety.

Fragmentation and category formation are intimately linked. When fragmentation of left-contexts is high, a distributional learner can be highly certain about which target word follows a given left-context. When this is the case, there is little need for clustering target words into categories — their left-contexts are doing that work already. This idea is supported by Cassani et al. (2018) who found that effective lexical categorization depends on some uncertainty regarding which contexts a target word occurs in. Because, in the RNN language model, lexical semantic categories emerge in the service of the next-word prediction task, there would be no need for category formation if target words were highly predictable. When maximally predictable, each target word would be encoded as if it were its own category. If so, the similarity of target lexical representations between semantic categories would be nearly as high relative to within semantic categories. Under such extreme conditions, no semantic category structure would emerge. Without distributionally-mediated clustering, it would be nearly impossible to perform the distributionally-mediated extension of category-associated features (DECAF). The purpose of the counterbalancing requirement is to prevent this from happening. The counterbalancing of left-contexts preserves prediction uncertainty for as long as possible during processing of an input sequence, until a target word occurs. When the target word occurs, it is the target word itself that guides next-word prediction, rather than the words that occurred before it.

We can now restate the ‘counterbalancing requirement’ more succinctly: When the left-contexts in which target words occur are fragmented, the RNN cannot acquire fully atomic lexical semantic representations. To promote atomicity, we require a corpus with minimally fragmenting left-contexts. Such a corpus would be organized like the hypothetical co-occurrence matrix shown in the right panel of Figure 6.2. Target words belonging to the same semantic category (e.g. MAMMAL) each occur after the same left-contexts, *the* and *this* (and with equal probability). By ensuring that left-contexts provide no information about the semantic category of an upcoming target word, the RNN is prevented from encoding semantic information about the target word in the lexical representation of a preceding word, and is, instead, forced to encode this information in the lexical representation of the target word itself.

6.3 Incidental Redundancy in Natural Language

The artificial language corpora used in the previous chapter are extremely unrepresentative of the structure of natural languages, where multiple items providing perfectly redundant information about the semantic category of a target word probably almost never occur in the same sentence. After all, natural languages are shaped by communicative pressures on speakers that, among other functions, reduce the amount of redundant information that is conveyed. While it is certainly possible to use natural language to construct sentences with high level of redundancy among words, it is reasonable to assume that speakers avoid them in favor of more compact utterances. By creating artificial languages with perfectly redundant items, I do not mean to suggest that they accurately represent human language environments. In fact, the artificial language experiments are not meant to simulate how humans might learn from real world input. Instead, the artificial corpora are scientific tools that allow us to better understand the learning dynamics of the RNN under theoretically interesting conditions. In many cases, theoretically relevant conditions do not necessarily



(a) A hypothetical co-occurrence matrix that does not satisfy the counterbalancing requirement. There is considerable fragmentation in this data due to numerous predictive relationships between left-contexts and individual target words.

(b) A hypothetical co-occurrence matrix that satisfies the counterbalancing requirement. Each MAMMAL word (y-axis) co-occurs with the same left-context words. Fragmentation is entirely absent in this data as there are no left-contexts that can be used to predict which target word will occur next.

Figure 6.2: An illustration of two hypothetical co-occurrence matrices for target words that belong to the semantic category MAMMAL. While the data in (a) does not fulfill the counterbalancing requirement, the data in (b) does.

correspond to conditions that are frequent — or even attainable — in the natural world. Computational modelers may opt to explore models under extreme conditions, that isolate and emphasize a particular aspect of the world that is of theoretical interest. What, then, is the theoretical interest in redundancy if natural languages likely do not generate many sequences of semantically redundant words? While natural language sequences may avoid high levels of redundancy, there are nonetheless rife with *incidental redundancy.*

Incidental redundancy is an inevitable statistical consequence of exposure to a small sample of all possible stimuli. Because language is considered an unbounded generator of linguistic stimuli, any finite sample — that is not perfectly counterbalanced in the way proposed above — must contain spurious correlations among sequentially occurring words. Consequently, there is a high likelihood that the left-contexts in which same-category members, such as *dog* and *cat*, occur are not identically distributed in the sample. There is no principled reason why *black* should not occur before *cat* but not before *dog*, as both combinations are grammatical and semantically plausible. However, the idiosyncratic nature of language use may result in a learner being exposed to a sample where '*black dog*' occurs but not '*black cat*'. Such a leaner might mistakenly infer that cats cannot be black, or that being black is a typical or category-diagnostic feature of dogs. Encoding incidental redundancy in learned lexical representations is therefore maladaptive, and should be avoided if the representations are to be used for performing DECAF.

6.4 Testing SPIN Theory using Child-directed Input

Given the strictness of the counterbalancing requirement and the ubiquity of incidental redundancy in natural language data, it appears that using the RNN to model the construction of atomic form-based lexical semantic representations from natural language input may not be tenable. That said, it is likely that the counterbalancing requirement is too strict to accurately describe the success that next-word prediction might

simple RNN		LSTM	
contextualized	lexical	contextualized	lexical
0.679 ± 0.004	0.651 ± 0.004	0.698 ± 0.003	0.648 ± 0.003

Table 6.1: A comparison of balanced accuracy at correctly distinguishing same-category from different-category probe word pairs when representations are computed dynamically at the hidden layer or statically accessed at the input-to-hidden weights. Results for both the simple RNN and LSTM converge on the same conclusion: While more information about semantic category membership is captured by the dynamic procedure used to generate contextualized representations at the processor, a majority of that information is also made statically available in the input-to-hidden weights where it can be more readily accessed in a cognitively plausible manner. Values shown are mean ± margin of error with $\alpha = 0.05$ and n = 10.

have when constructing lexical semantic category knowledge from natural language input. It is likely that, in practice, the RNN can tolerate a great deal of redundancy without sacrificing lexical atomicity (as shown in the simulations using language PAX, DAX, PAY, and DAY in Chapter 5). So, just how robust is the RNN when provided language directed to children as input?

In particular, I tested the hypothesis that the semantic categorization performance of learned non-contextualized lexical representations should be lower relative to learned contextualized representations at the hidden layer. This follows from the intuition that natural language, including input to children, is rife with incidental redundancy[2], and that this redundancy impedes the static availability of learned semantic category knowledge at the input-to-hidden weights. If semantic category knowledge is indeed trapped inside the processor[3], then the balanced accuracy of lexical semantic representations should be lower. Just how much lower determines how usable the learned representations are for performing the distributionally-mediated extension of category-associated features (DECAF), critical for guiding children's inductive inferences about novel word meanings in the absence of referential information.

I trained 10 simple RNNs and 10 LSTMs on AO-CHILDES, using the training methodology described in Chapter 3. In addition to evaluating contextualized representations, which requires external memory only accessible to the experimenter to be collected, I also evaluated the non-contextualized lexical representations, which are statically accessible, and therefore a much more plausible source of information available to children. The results, reported in Table 6.1, show that some information is indeed trapped in the dynamically constructed contextualized representations at the hidden layer, but, importantly, not much. For the simple RNN, the difference in balanced accuracy between contextualized and non-contextualized representations is only 0.679 - 0.651 = 0.028. A two-tailed t-test with $\alpha = 0.05$ shows that this difference is statistically significant.

[2] I confirmed that left and right contexts of probe words in AO-CHILDES redundantly encode semantic category membership by training a classifier (Linear Discriminant Analysis implemented in the Python package *sklearn*) to predict the semantic category of each probe word based on accumulated frequencies of co-occurrence with words that immediately precede or follow the probe word. By training on one of two halves of the corpus and tuning accuracy on the held-out half, the best accuracy on held-out data was well above chance (24% and 18% of all probe words were correctly classified, using left and right context word statistics, respectively).

[3] By 'processor', I mean the dynamics that give rise to the contextualized representations at the hidden layer.

6.5 Relaxing the Counterbalancing Requirement

The findings presented above have implications for language acquisition research, namely that some distributional information useful for performing DECAF may not be readily available in a format that children can use — at least, if their distributional learning system is based on next-word prediction. Is there anything that can be done to help semantic information that might otherwise be trapped at the processor to settle at the static input-to-hidden weights? Yes, there are many potential research directions that can be explored. I will focus on one direction that I think is particularly relevant to language acquisition research.

I have identified one potential relaxation of the counterbalancing requirement, that, when met, may improve the atomicity of lexical semantic representations constructed from natural language input. The relaxed requirement is inspired by longitudinal analyses of the statistical properties of child-directed input: Considering nouns as target words and the noun category as a single target category, I found that both left and right contexts of nouns are less fragmented in language to younger compared to older children (1-3 years vs 3-6 years). The reduced fragmentation of pre-nominal contexts in input to younger children is a natural approximation of the counterbalancing requirement. A detailed discussion of the steps used to quantify fragmentation is provided in Chapter 7.

This finding raises two possibilities: First, according to SPIN theory, an RNN trained on input to younger compared to older children should develop more atomic lexical semantic representations of nouns. Second, once atomicity for nouns has been established by training on input to younger children, it might be possible to preserve the atomicity that has already been established while training on input to older children. Whereas SPIN theory proscribes a strict counterbalancing of left-contexts of an entire corpus, it is possible that such a counterbalancing can be relaxed once maximal atomicity has been established in the network during a pre-training phase on minimally fragmented input. In other words, is it possible that atomicity, once established, is long-lasting, even when faced with input with a high level of fragmentation? Based on insights into neural network learning dynamics (Achille et al., 2018; Roark et al., 2022; Saphra & Lopez, 2020; Servan-Schreiber et al., 1991), I hypothesize that this is possible, and empirically test this prediction in Chapter 9.

6.5.1 Staged Training Regime

In particular, I propose that a staged training strategy, that emphasizes the coherence of the noun category during early training, places the RNN in a better position to acquire more atomic lexical semantic representations of nouns. By emphasizing the coherence of nouns as a single category, noun representations should be more receptive to semantic cues, and less likely to participate in chunk-level phenomena. Put another way, by first recognizing the *similarity* between nouns (membership in the same grammatical category), the RNN should be in a better position to discover semantic *dis-similarities* between nouns. In Chapter 9, I provide more detail about how I have arrived at these assumptions.

The staged training regime that I have in mind requires that (i) the RNN be trained incrementally, and (ii) that its training data be staged. I distinguish between incremental training and staged learning as follows: The basic idea behind incremental training is that a model does not have full access to the data at every point in training, by, say, being exposed to stimuli that are randomly sampled from the whole dataset. Staged learning builds on this idea, and additionally requires the identification of periods during learning that differ qualitatively in some aspect, such as what a model is learning, or what kind of data is being presented to it. In order to test the prediction that an RNN will learn more atomic lexical semantic representations when trained incrementally on age-ordered input to children, both an incremental training regime, and age-ordered data are required. I will refer to the combination of these two elements as a ‘staged training’ regime or strategy.

6.5.2 Developmental Plausibility

Contrary to standard practice in neural network training, in which a network is exposed to stimuli sampled in random order from the dataset, incremental training exposes a network to ordered partitions of the data, such that previously seen partitions are never revisited during subsequent training. For more details regarding my proposed staged training regime, I refer the reader to Chapter 10, where it is experimentally validated. Briefly, a corpus is partitioned into equally sized chunks, and instead of iterating over the entire corpus multiple times, one corpus partition is iterated over at a time. This kind of training regime typically yields worse learning outcomes compared to iterating globally over the entire dataset[4], but is essential for modeling language acquisition because the statistical learning system of children probably does not integrate stimuli across time spans of months and years. It is more likely that children process language input as soon as possible, without returning to inputs they have processed many months ago. Also, by randomly sampling from children’s language environment, one disregards the non-stationary nature of the data distribution; as shown in Chapter 2, surface-level statistical properties of language directed to children undergoes large age-related changes in lexical and combinatorial diversity, among others. Respecting the non-stationary aspect of the training data may allow researchers to capture developmental phenomena in acquisition, or induce desirable properties in the network such as combinatorial generalization to novel stimuli (Vankov & Bowers, 2020).

6.5.3 Challenges: Perfect-Redundancy Pockets

There is a reason why incremental training is not standard procedure in neural network science. Iterating over smaller partitions of a dataset results in a temporary over-fitting on a given partition, and iterating over multiple partitions can lead to forgetting of knowledge acquired during training on past partitions. In addition, local over-fitting due to incremental training can compound the negative effects of incidental redundancy on atomicity. To illustrate this, consider that the expression ‘*big gorilla*’ occurs once in a hypothetical partition of some corpus, and that this is the only time the word *gorilla* appears in this partition. Further, assume the target semantic category to be learned is MAMMAL, and that its members contain both large and small animals. This means that whether a gorilla is large or small is not diagnostic of its membership in the category. This setup is problematic because iterating multiple times over the same partition, results in ‘*big gorilla*’ being observed multiple times during training, which reinforces the spurious association between *big* and *gorilla*. This is equivalent to the artificial language simulations reported in the previous chapter where left-contexts provide information that is perfectly redundant with the target word (i.e. when $\alpha = 1.0$). Based on these results, I predict that the RNN will learn a representation of *gorilla* that is heavily entangled with *big*, rather than a representation of *gorilla* that is semantically encapsulated (i.e. atomic). I will refer to corpus partitions with this property as ‘perfect-redundancy pockets’ because an RNN trained on such a partition will encode the pre-nominal word and the noun in expressions like ‘*big gorilla*’ as providing perfectly

[4] I have verified this by comparing semantic categorization accuracy between RNNs trained on AO-CHILDES using either a single partition vs. 8 partitions.

redundant information about upcoming words. Essentially, the training strategy teaches the RNN to treat the chunk '*big gorilla*' as a compound cue, rather than as consisting of two separate lexical units, each possessing its own predictive properties.

It follows that any corpus partition with highly fragmenting contexts is more likely to be a perfect-redundancy pocket. With more idiosyncratic left-context + noun combinations, the likelihood that there is one such combination that is also the only time a target word occurs in the same partition is increased. Iterating over such a partition multiple times would strengthen the idiosyncratic relationship between the left-context and the noun, and as a consequence would make it more difficult to learn atomic lexical semantic representations. Perfect-redundancy pockets are inevitable with highly fragmented corpus data and when iterating over small partitions of the data. The question is not how to avoid them, but *when during training* to avoid them. My hypothesis is that it is most important to avoid perfect-redundancy pockets during the earliest stages of training, when the potential for organizing the network unfavorably is largest. The remainder of thesis examines the motivation behind this hypothesis, and provides empirical support.

6.6 Summary

The results of the artificial language learning simulations in the previous chapter demonstrated that a left-to-right predicting RNN learns sub-optimal lexical semantic representations when information about semantic category membership provided by an item in the left context of a target word is redundant with that provided by an item in the right context. To generalize these observations and make them more useful to other researchers, I have proposed a language-agnostic set of principles (SPIN theory) to understand and predict under what conditions a left-to-right predicting RNN language model can expect to learn fully atomic lexical representations. A key component of SPIN theory is the counterbalancing requirement, which states that left-contexts must not predict category-relevant semantic information provided by right contexts of target words. Next, I noted that this requirement is violated by many natural languages such as English where semantic information about nouns can occur both pre- and post-nominally. For instance, pre-nominal adjectives convey additional semantic information which may be used to predict post-nominal items, and therefore may provide similar (partially redundant) or identical (perfectly redundant) information about the semantic category of the intervening noun. This is to be avoided to learn atomic lexical semantic representations of nouns — a desideratum for performing the distributionally-mediated extension of category-associated features (DECAF).

Next, I hypothesized that one way to promote atomicity when faced with natural language input of this kind is to train the RNN incrementally on ordered input. In this weaker version of the counterbalancing requirement, only the first few samples that the RNN is trained on must be counterbalanced. Specifically, the left-contexts of target words should provide minimal redundant information about upcoming semantic cues, in order to avoid the encoding of maladaptive across-target dependencies. Furthermore, I noted that input to children that has been age-ordered (across developmental time) approximates this weakened counterbalancing requirement. As a result, the proposal developed in this chapter predicts that an RNN trained on input to younger children first learns more atomic lexical semantic representations of nouns than an RNN trained on input to older children first.

Before discussing the results of RNN simulations with child-directed input, in the next chapter I examine whether there is indeed less fragmentation of the left-contexts of nouns in input to younger compared to older children. This is necessary to confirm that the premise underlying my hypothesis holds.

Chapter 7

Age-related Increase in Fragmentation of the Noun Category

In this chapter I examine the degree to which language input to children conforms to the counterbalancing criterion, a key component of SPIN theory that was presented in the previous chapter. In particular, I focus on the noun category in input to English-learning children between the ages 1 and 6 years. Quantifying the amount of counterbalancing is akin to quantifying the amount of 'fragmentation', which I have defined as the conceptual inverse of counterbalancing. Putting a number on the amount of fragmentation in children's language input is useful for making predictions about the atomicity of lexical semantic representations of target words that would emerge in a system that learns via next-word prediction. Much of the work presented in this chapter has been made previously made available in a peer-reviewed journal (P. A. Huebner & Willits, 2021b).

First, I provide an in-depth explanation of fragmentation, its utility for studying distributional statistics of language corpora, and how to quantify it. Next, I describe the steps I have taken to extract co-occurrence matrices from AO-CHILDES. I extracted two matrices, that represent co-occurrence counts of nouns and their neighbors in language to younger children or language to older children. I compute fragmentation separately on each co-occurrence matrix to determine whether there is any age-related difference. In the remainder of the chapter, I discuss additional follow-up analyses that zoom in on different kinds of syntactic phenomena that may be responsible for the observed age-related difference in fragmentation.

7.1 Fragmentation

I borrow the term 'fragmentation' from the statistics and machine learning literature, in which it is defined as a learning trap that arises from assuming an interaction between inputs that are in fact independent (Jakulin & Bratko, 2003). While I adopt the term 'fragmentation' from the learning literature verbatim, I will use it slightly differently to refer to a property of the *data*, and not assumptions made by a *learner* — even when such assumptions were prompted by the data in the first place. It is important to keep this distinction in mind when reading the current work, because the discovery of fragmentation in the data, does not require that a learner will actually be influenced by it.

To illustrate how fragmentation might make it more difficult for distributional learners to arrive at useful representations, I return to the noun category, a primary topic of this thesis. Because many nouns often occur after determiners, and often precede verbs, it is relatively straightforward for distributional analysis to recover the set of words that linguists have termed nouns. However, while they are members of the same category, nouns nonetheless differ in meaning, and individually enter into distinct lexical relationships that are not shared by all members of the category. If this were not so, there would be no need for different words that refer to different objects. Moreover, nouns are not a homogeneous category, but can be divided into virtually infinite numbers of smaller subordinate categories, to distinguish, say, animate from inanimate objects, or fast food from vegan food. Even these smaller, and often semantic, subcategories leave behind unique distributional signatures that can be detected via distributional analysis of large corpora. Importantly, each noun is characterized by a mixture of distributional information, which identify its membership not only in the noun category, but also in numerous subcategories. By definition, distributional information that helps to distinguish between subcategories of nouns, interferes with the ability to learn that those subcategories all belong to a single superordinate category. Because the discovery of a noun category is contingent on exposure of nouns that occur in similar contexts, difficulties may arise when a learner is initially exposed to nouns in lexically or subcategory-specific contexts that obscure the presence of the superordinate category.

It can be useful to think of fragmentation as the opposite of anchoring, which is the stable marking of linguistic categories by highly repetitive, lexically-specific frames such as *In X*, *What do X*, *Are you X*, *It's X*, *Let's X*, *Look X*, *I think X*, *If X* (Cameron-Faulkner et al., 2003). Morphological markers such as *-ing* are also highly frequent units that are known to anchor part-of-speech categories, by highlighting a stable difference between nominal and verbal contexts (J. A. Willits et al., 2014). Cameron-Faulkner et al. (2003) suggested anchor points are starting points from which children enter into the more complex and formal aspects of language acquisition. Whereas input with anchor points prioritizes distinctions between broad categories, fragmentation prioritizes finer-grained distinctions by flooding the learner with lexically specific information that obscures the presence of the broader category. Invariably, the hierarchically organized sub-category structure of language data contains statistical regularities of each kind; while some co-occurrence information is more anchor-like, other regularities highlight structure at the subordinate or lexical level. These conflicting cues produce a tug-of-war between encoding regularities at super- vs. subordinate category levels —- the stronger a learner encodes the distinctions that exist within a category, the weaker the representation of the category, and vice versa. Therefore, the ideal situation for the incremental learner is to be exposed as early as possible to anchor points that can reduce the effects of fragmentation. After the broadest distinctions have been acquired, a learner will be less influenced by fragmentation, and therefore will be less reliant on anchor points to combat fragmentation.

Fragmentation, as I have defined it, exists on a theoretical continuum, with both ends representing idealized scenarios. On one end of the spectrum, fragmentation is nonexistent because all lexical distributions associated with all members of a particular category are identical — they are maximally similar. In essence, all words within the category are indistinguishable from one another in terms of their distributions. On the other end of the spectrum, fragmentation is maximal, and this occurs when the lexical distributions of words belonging to the same category are maximally different, and have no overlap. In such situations, there is no basis for hypothesizing that the words belong to a common category. The precise degree of fragmentation of a particular category can vary dramatically, depending on the corpus under study, the amount of data available, and the complexity and lexical diversity of the corpus. Because the surface-levels statistical properties of language to children differs dramatically from that between adults, it is reasonable to assume that fragmentation changes over developmental time, and in this way implicitly contributes to the shaping of linguistic hypotheses.

7.1.1 A Visual Demonstration of Fragmentation

To illustrate fragmentation, I constructed and visualized three hypothetical co-occurrence matrices that differ in fragmentation, shown in Figure 7.1. Before discussing the demonstration, some housekeeping. First, rows correspond to nouns, and columns correspond to the contexts in which they occur. Second, the black and white elements in the three co-occurrence matrices shown in Figure 7.1 indicate that either a noun co-occurs with a context (black) or that it does not (white). Third, my use of binary matrices is for simplicity of demonstration, and the arguments I present in this chapter hold for co-occurrence values of any range.

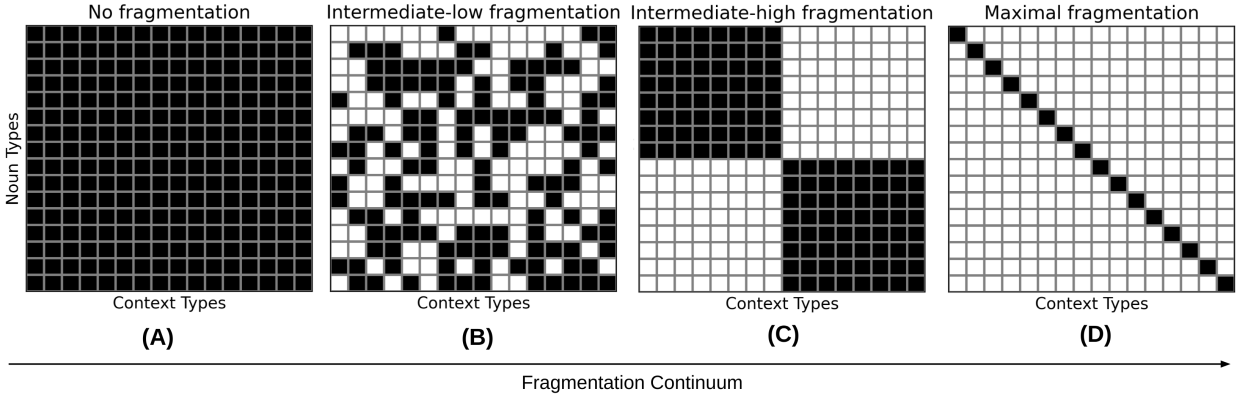


Figure 7.1: A visual demonstration of the fragmentation continuum, with hypothetical co-occurrence matrices (nouns in rows, contexts in columns) that exhibit either no fragmentation (A), intermediate-low fragmentation (B), intermediate-high fragmentation (C), or maximal fragmentation (D).

In (A), each of the 16 nouns occur with each of the 16 contexts and in exactly the same pattern (uniform here, but can be any other pattern). This co-occurrence pattern is extremely improbable in natural language, and therefore should be considered no more than an idealization of nouns. Nonetheless, the idealisation of nouns represented by the co-occurrence matrix in (A) is a useful construct for explaining fragmentation, because it exemplifies the most extreme left end of the fragmentation continuum, where fragmentation is totally absent. Any departure from this idealized pattern must results in fragmentation because the rows will no longer be identical to each other.

One way in which the co-occurrence pattern may depart from (A) is shown in (B). Here, nouns do not occur with each possible context systematically, but with a smaller number of contexts, and with no apparent sub-category pattern — indeed, the matrix was generated by randomly populating each element with either a 0 or 1. As such, (B) is more fragmented than (A) — the ability to learn that each word belongs to the same category is impaired.

The co-occurrence matrix (C) is even more fragmented than (B). Just like (B), the words in (C) are not identical in terms of their co-occurrence patterns, but unlike (B), the ways in which the words vary is systematic, forming two distinct subcategories. In keeping with the noun example, this situation can occur when there are two highly coherent subcategories of nouns (like animate and inanimate nouns), and where the distributional contexts of the nouns perfectly predicted this difference. For instance, we can think of lexical contexts as picking out specific categories of nouns: Contexts such as '*happy X*' or '*grumpy X*' are much more likely to be used in combination with animate than inanimate nouns. This strong sub-category division makes it even more difficult — if not impossible — to learn that members of both sub-categories also belong to the same larger category.

Lastly, consider the co-occurrence matrix (D), in which every noun co-occurs with exactly one context. Like (A), the simulated co-occurrence matrix in (D) is extremely unlikely to realize in natural language corpora, and as such is useful only as a theoretical construct. But (D) is useful as a demonstration of maximal fragmentation, in that there is no distributional overlap between any of the words. As in (C), there are no cues that group the words together into a single noun category. However, (D) is more fragmented than (C), because overlap between nouns is entirely absent, precluding any grouping into intermediate subcategories present in (C).

Fragmentation of lexical co-occurrence data can be either advantageous or disadvantageous, depending on the learner's goal. If, for instance, the goal of distributional analysis is to discover broader categories, such as the grammatical categories like nouns and verbs, a learner would benefit the most by being exposed to data that is minimally fragmented (as close as possible to (A) for each part-of-speech category). However, such a learner would not have access to distributional evidence of subcategory structure. Consequently, the fragmentation continuum reveals a fundamental trade-off between superordinate and subordinate category cues in lexical distributional data: If the co-occurrence patterns at a subordinate or lexical level (idiosyncrasies associated with usage of individual words) are stronger than those at a superordinate level, the discovery of the larger category is impaired. Conversely, a corpus with more formulaic constructions and/or limited lexical diversity, can produce strong distributional regularities at a superordinate level that can obscure the presence of structures below. It is precisely this aspect of caregiver language which prompted the question motivating the work in this chapter: Does the formulaic nature of language input to younger children provide distributional patterns that are better suited for the discovery of the noun category than input to older children? In other words, are noun less fragmented in input to younger compared to older children? Next, I discuss the steps taken to answer this question.

7.2 Methods

7.2.1 Quantifying Fragmentation

As discussed before, the ideal data for the distributional learner tasked with discovering the noun category is sequences where the left-contexts of nouns are perfectly counterbalanced. I refer to any departure from this idealization as fragmentation. It follows that the presence of any lexically specific pattern (applying to sub-groups of or to individual nouns) obscure the target hypothesis, which is that "all nouns are identical in terms of their co-occurrence patterns". How can this informal intuition be formally quantified?

Because fragmentation measures the degree of lexically-specific relationships that are unique to one or more nouns, but are not shared across all nouns, we cannot use bivariate similarity metrics (e.g. Pearson correlation coefficient). Bivariate metrics do not take into account overlap across multiple vectors (e.g. row vectors for all words in a category); instead, they measure overlap between pairs only, and this completely ignores the global pattern — multivariate (higher order) correlations — exhibited by a larger set of vectors. Even via pairwise aggregation, the use of bivariate correlations is a poor choice for identifying multivariate correlations, because there is no guarantee that all higher-order correlations are captured. Because my goal is to quantify the degree to which a set of co-occurrence vectors instantiate a single, shared co-occurrence pattern, I require a multivariate tool. For example, multiple correlation is the correlation between one variable's observations (i.e. is the word a noun?)[1] and the best predictions that can be computed linearly from a set of predictive variables (i.e. co-occurrence frequency). In order to account for the highest possible variance, the best linear transformation must identify the co-occurrence pattern that is shared by all observations (i.e. nouns). It follows that the variance *not* explained by the best linear transformation can be used as an operational definition of fragmentation. We can think of the best linear transformation as the prototype co-occurrence pattern for a particular category, or as the category's baseline co-occurrence frequency pattern that is hidden beneath a myriad of sub-category and lexically-specific patterns.

7.2.2 Singular Value Decomposition

To identify the best linear transformation, I opted for singular value decomposition of the noun co-occurrence matrix.[2] For simplicity, I will refer to the best linear transformation as the prototype co-occurrence pattern, or just prototype vector. This prototype vector is considered 'best' because it maximizes, roughly speaking, its overlap with all row vectors in the co-occurrence matrix. When the rows of a co-occurrence matrix correspond to distributional patterns of nouns, we can think of the prototype vector as the prototypical noun pattern. There's only one such vector, and it is equivalent to the first singular vector — the basis vector which accounts for the highest amount of variance in the data. Singular value decomposition (SVD) enables us to compute this singular vector straightforwardly. SVD is a tool for decomposing a matrix into left and right singular vectors, which separate the variance in the rows and columns into orthogonal dimensions (also referred to as singular dimensions or basis vectors). Additionally, for each singular vector, SVD provides a corresponding singular value which is proportional to the amount of variance explained by it. Because I am interested in the amount of variance explained by the first singular dimension, I can obtain my desired results directly using SVD, as opposed to a two-step procedure consisting of (i) the computation of the prototype vector, and (ii) the computation of how much variance it accounts for using multiple correlation.

This proposed method is consistent with work on basis vectors of large lexical co-occurrence matrices by Lee (2015) who concluded that the first basis vector is the "defining" vector that encodes the most general information about a category, and that all other subsequent basis vectors encode more "specific" information pertaining to subsets of or individual words. Because the singular dimensions identified by SVD are ordered by the amount of variance explained, the sub-spaces spanned by each subsequent singular dimension can be considered prototype vectors for distinguishing between sub-categories within a larger category.

Because the first singular value quantifies the extent to which a co-occurrence matrix can be explained in terms of a single dimension (e.g. noun-ness), I subtracted the first singular value from the sum of all singular values to compute a measure of fragmentation — the amount of variance *not* explained by the prototype vector. Let fragmentation = $Frag$, then

$$Frag = \frac{(\sum_{i=1} s_i) - s_1}{\sum_{i=1} s_i} = 1 - \frac{s_1}{\sum_{i=1} s_i} \tag{7.1}$$

where s_i is the i-th singular value.

[1] In this case, the independent variable would simply be a 1 every time, indicating that each word in the noun co-occurrence matrix is a noun.

[2] Multiple regression is an alternative option, but it explicitly models error and intercept terms which is not required here.

7.2.3 Selection of Nouns and Non-nouns

To decide which co-occurrence counts to include in the co-occurrence matrices, I created a list of frequent nouns in AO-CHILDES. This list was created as follows: First, I part-of-speech tagged AO-CHILDES (using the Python package *spacy* v2.3.7), and collected all words that were tagged as a noun at least once. Next, I manually inspected the resulting list by removing words that cannot or are extremely unlikely to be used as nouns in child-directed speech. I also excluded plural nouns, proper nouns, interjections, number words, and gerunds. I further excluded words which did not occur at least 10 times in AO-CHILDES. The resulting lists contains 707 singular noun types, which overlap to a large degree with the probe words used in Chapters 3 and 6.

When collecting co-occurrence data, I did not simply collect data for any occurrence of a word that is in the noun list. Instead, the tagger must have first assigned a word as a noun in the sentence in which it occurs, before it is checked against the noun list. This 2-step procedure has the advantage of (i) using only words that are tagged as nouns in the sentences in which they actually occur, and (ii) reducing false positives produced by the tagger.

Additionally, I created a list of non-nouns, which I used as control words. I used these words in the same way as nouns to examine if any age-related trend observed for nouns are also true of non-nouns, which would indicate a more global shift in the distribution of the data, rather than a noun-specific effect. I did so by pairing each noun with a randomly selected word from AO-CHILDES that is approximately matched in frequency.

7.2.4 Partitioning AO-CHILDES by Age

To examine potential differences in fragmentation by age of the target child, I split AO-CHILDES into two sub-corpora. The steps taken are reproduced below.

First, it is important to consider that the number of transcripts in AO-CHILDES are not uniformly distributed across age. For example, there is an order of magnitude more data for children 800-1000 days old compared to children 200-400 days old. That is, AO-CHILDES is extremely biased towards 2-year olds. This is not surprising, as many studies used to populate the CHILDES database recruited children when they were right around 2 years of age. The non-uniformity of the age distribution prevented me from splitting the corpus to produce two equally-sized sub corpora representing similarly-sized age ranges. Splitting the corpus in half based on number of words would have resulted in one sub-corpus with primarily speech to 1 and 2 year olds, and another with speech to 2-6 year olds. This unequal representation of age in the two halves of AO-CHILDES prompted me to explicitly split by age, rather than by the number of tokens. To do so, I searched for two equally sized age ranges (in days) that produced two approximately equally sized sub-corpora. This resulted in a first sub-corpus that contains 1635 transcripts (2.7M tokens), and a second that contains 1665 transcripts (2.5M tokens). The resulting two sub-corpora contain speech to children between the age of 90-1090 days and 1140-2140 days, consisting of 1639, and 1665 transcripts, respectively. I will refer to them as sub-corpus 1 and 2, and the age groups they represent as age group 1 and 2, respectively.

7.2.5 Collecting Co-occurrences

The input to the computation of fragmentation requires a co-occurrence matrix. I obtained two co-occurrence matrices, one that characterizes noun co-occurrence counts in input to younger children, and another for older children. First, I collected all sliding windows of size 3 for both sub-corpora. A window size of 3 considers the target word and only its left and right neighbor. Next, I separated the windows based on whether the target word met the criteria for noun or non-noun membership: If the word was in the non-noun control word list, it was used for the construction of a non-noun co-occurrence matrix. On the other hand, if the target word was a noun, it was used to construct the noun co-occurrence matrix. In both cases, the target word (noun or non-noun) labeled the rows of the co-occurrence matrix, and contexts labeled the columns. For both nouns and non-nouns, I created two co-occurrence matrices, one in which the context is defined as the word preceding the target word (backward direction), and another in which the context is defined as the word following the target word (forward direction). I did not include a combined condition, because the presence (or absence) of fragmentation in one or the other condition necessitates its presence (or absence) in the combined condition[3]. An additional reason not to examine a combined condition is the finding by Freudenthal et al. (2013) who showed that independent contexts can classify items with a higher degree of accuracy than combined contexts. Next, because of the unequal size of the two sub-corpora and the unequal number of nouns in each (noun density is higher in sub-corpus 1), I stopped collecting co-occurrences when their number reached a threshold. This threshold was determined based on the number of nouns (non-nouns) in the sub corpus with the fewest nouns (non-nouns). Because there are far more nouns in sub-corpus 1, I had to drop approximately 30,000 noun occurrences to equate the number of nouns in sub-corpus 1 with the maximum number of nouns in cob-corpus 2 (77,677). Similarly, I dropped about 3,000 non-noun occurrences in sub-corpus 2 to equate the number of non-nouns across the two age groups (104,394). I did so to remove any confound of frequency when comparing sub-corpora (i.e. age groups).

I investigated the influence of many variables, treating each as a factor in my experimental design. The full list of factors and factor levels are shown in Table 7.1. For example, I varied whether original words or their lemmatized forms were collected. Lemmatization involved a rule-based removal of all inflectional morphemes. I included this factor because children can pool evidence across morphological inflections when making linguistic generalizations. Doing so also reduces biases due to the fact that all of the measures depend on the counts of many rare, discrete events. Additionally, I investigated the influence of including versus excluding punctuation, which can be considered as textual markers of prosodic and temporal boundaries in fluent speech. One consequence of punctuation removal, is that when collecting co-occurrences between a target word and its right neighbor, it was possible that a word's right neighbor would be the first word in the subsequent sentence as opposed to the punctuation symbol which would have been collected had punctuation not been removed. I also tested for any influence of using the raw vs. normalized co-occurrence matrix (each element divided by its column sum), given that normalization is a routine procedure in computational linguistics and a pre-processing step before multivariate analysis. Normalizing by the column-sum scales the variance in each column such that its proportion of the total variance is the same as every other column, and this can reduce the influence of columns which prior to normalization accounted for disproportionately more variance than other columns. Lastly, for each condition, I also collected a control co-occurrence matrix for randomly-selected non-nouns matched in type and token frequency to the noun lists. This was done to test whether any age-related trends observed for nouns are also more generally true of other words in the corpus. If so, i cannot conclude that age-related trends in fragmentation of noun contexts is specific to nouns.

[3] To make this point clear, in a combined condition, the co-occurrence matrix would simply be a horizontal concatenation of the matrices collected in the forward and backward directions, and therefore, fragmentation in one or the other (or both) would persist in the combined condition.

age(days)	noun	context	lemma	punctuation	normalization
90-1090	True	forward	True	intact	none
1140-2140	False	backward	False	removed	divide by column sum

Table 7.1: Factors influencing the construction of co-occurrence matrices.

7.3 Age-related Increase in Fragmentation of the Noun Category

Given prior work which has shown that speech to younger children is more repetitive, less lexically diverse (Foushee et al., 2016; Hayes & Ahrens, 1988; Kirchhoff & Schimmel, 2005), and more template-like (Cameron-Faulkner et al., 2003), I predicted that the noun category would be less fragmented in speech to younger children. Furthermore, given the privileged status of nouns in children's early learning, I predicted that the pattern of increasing fragmentation with age is specific to nouns, and therefore would not extend to the control word list (non-nouns).

The results of the fragmentation experiment is shown in Figure 7.2. Each panel contrasts two conditions: the effect of age group (90-1090 days in blue, 1140-2040 days in orange), and the effect of using the lexical contexts of nouns vs. non-nouns. The different normalization conditions (raw frequency vs. normalization by column sum) are separated horizontally, and the remaining three combinations of conditions (context direction, lemmatization, and punctuation) are separated vertically across the different panels. I am primarily interested in a potential effect of age group, and whether such an effect is influenced by any of the other manipulations.

I begin by focusing on the 16 comparisons shown in the left panels only — that is, for co-occurrence matrices that were not normalized. As predicted, I found reduced fragmentation of nouns in speech to younger children (blue bars tend to be lower than orange bars), and this held true in all but two conditions. The average absolute difference in fragmentation across these experimental conditions was 0.38 ± 0.021 (mean ± std). Loosely speaking, this means that for age group 1, 3-4% more of the total variance in the co-occurrence patterns of non-nouns is explained by the first singular dimension, the prototype noun co-occurrence pattern. Moreover, the age-related increase in fragmentation appears to be noun-specific: In each control condition, non-nouns were *more* fragmented in speech to younger compared to older children (see pairs of bars marked 'non-nouns' on the x-axis). The average absolute difference in fragmentation across these control conditions was 0.35+/- 0.018 (std). Zooming in, I found that fragmentation was reduced with and without lemmatization, and regardless of whether co-occurrences are collected in the forward or backwards direction. The only conditions in which I did not find a noticeable difference in fragmentation between age groups are those in which only forward co-occurrences are counted, and punctuation had been removed. This interaction between right-contexts and punctuation suggests that utterance boundary markers, such as periods, exclamation marks, and question marks are not only frequent right neighbors of nouns, but also play an important role in helping to group nouns together into a category, and thereby protecting them from fragmentation in input to younger children. In fact, the absence of a difference in fragmentation in the conditions in which punctuation was removed, suggests that punctuation symbols are the primary, if not the only, reason for reduced fragmentation at age group 1.

Next, I focus on the 16 comparisons shown in the right panels of Figure 7.2, in which fragmentation was computed on co-occurrence matrices where the co-occurrence counts were normalized by their column sums (i.e. the total number of co-occurrences for that context). Normalization, such as dividing by the sum of an element's row and/or column sum (or more sophisticated methods like point-wise mutual information)

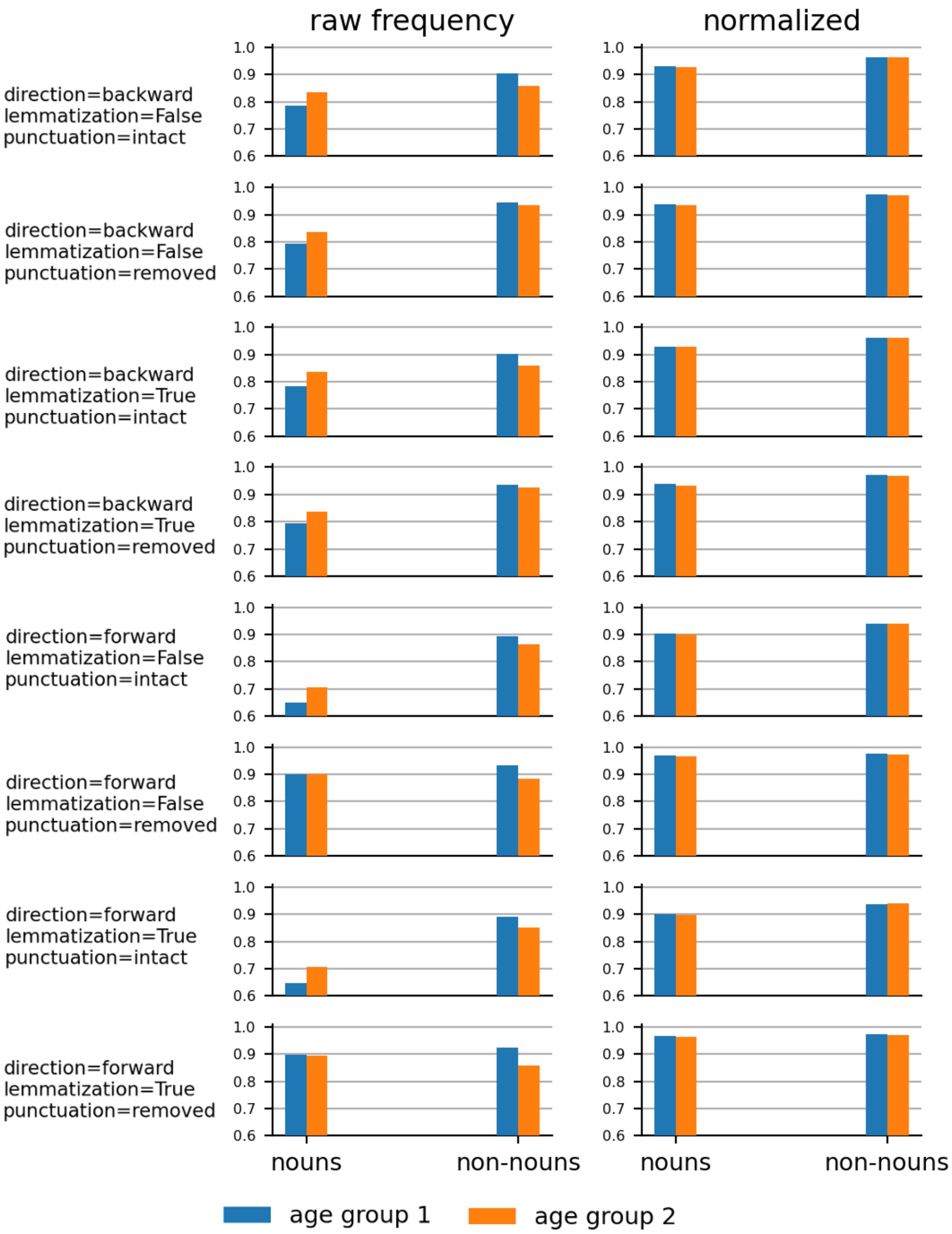


Figure 7.2: Fragmentation of the noun category (left panels), and a frequency-matched set of non-nouns (right panels) for each condition. The factors context direction (forward vs. backward), lemmatization, and punctuation vary top-to-bottom, and age, word list, and normalization vary left-to-right. Fragmentation is the proportion of variance explained by all but the first singular dimension of the co-occurrence matrix, and therefore is bounded between 0 and 1.

is standard practice in computational models of language and tends to improve distributional semantic models (Bullinaria & Levy, 2007; Turney & Pantel, 2010), and for this reason I included normalization as a factor. Notably, whereas I observed marked differences in fragmentation across age groups when no normalization was applied, these differences were completely abolished by normalization. Rather than casting aside this finding as a null-effect, I think it reveals an important clue related to the differential importance of different contexts for diagnosing noun category membership. Essentially, normalization removes information about total frequency differences between context words. This has consequences on the computation of fragmentation because each context word is treated similarly in terms of its importance for diagnosing membership in the noun category. It goes without saying that this assumption is unwarranted, as it is likely that there are only a few context that are highly diagnostic of noun category membership (e.g. punctuation, determiners) while many others are much less important. I discuss this distinction in more detail in the next chapter, where I introduce the notion of 'entropy-maximizing' contexts (i.e. anchor points), which are most useful for diagnosing category membership.

7.4 Follow-up Analyses

The results above demonstrate that both the left and right contexts of nouns are less fragmented in English input to younger compared to older children. This supports the proposal in the previous chapter, namely that by training an RNN language model on input to younger children first may promote the formation of more atomic lexical semantic representations of nouns. In the remainder of the chapter, I attempt to understand what linguistic factors are correlated with the increase in fragmentation as a function of age. I conducted three follow-up analyses, focusing on left-contexts: In the first, I explored which left-contexts are most and least fragmenting in each age group, Next, I turned my attention to two specific grammatical phenomena, pre-nominal adjectival modification, and noun compounding.

7.4.1 Most and Least Fragmenting Left-Contexts

The goal of my first follow-up analysis is to find those left-contexts which are the most and least fragmenting in language to children. As a starting point, I used a subset of the co-occurrence matrices used above to compute fragmentation (context direction = backward, lemmatization = False, punctuation = True).[4] For each age group, I computed the projection of the co-occurrence matrix on the first singular dimension, to isolate the co-occurrence pattern that is most diagnostic of the noun category. The first singular dimension captures the pattern of co-occurrence between nouns and contexts that is most shared across nouns, and the variance accounted for by this dimension is therefore used to quantify fragmentation. By summing across rows, and rank-ordering the resulting loadings from low to high, I obtained a list of left-contexts ordered from most to least fragmenting.

The results for age group 1 and 2 are shown in the left and right portions of Table 7.2, respectively. Each contains only the top-5 least and most fragmenting left-contexts, their total loading on the first singular dimension, and their frequency in the sub-corpus. I begin by focusing on age group 1. Unsurprisingly, I found that the least fragmenting left-contexts tend to be high-frequency words, that can occur with a broad range of nouns regardless of their semantic category, like the determiners *the* and *a*. Similarly, utterance

[4] In these analyses, I restricted nouns to the set of probe words used in Chapter 3.

boundaries marked in AO-CHILDES with period, exclamation, and question symbols, are all in the top-10 least fragmenting left-contexts. Of these three utterance boundary markers, the period symbol is least fragmenting, as shown in the table. I call such contexts ‘entropy-maximizing’ given that they provide high uncertainty (i.e. the least amount of predictive power) regarding which specific noun is likely to occur next. The results for least-fragmenting left-contexts are similar in age group 2, except that the period symbol is no longer in the top-5, *that* is switched with *this*, and, surprisingly, the adjective *little* appeared in the top-5. Inspection of occurrences of *little* in AO-CHILDES revealed that it is used with a large variety of nouns, primarily from the MAMMAL (e.g. *bird*, *sheep*) and FAMILY (e.g. *baby*, *sister*) categories, but also with non-animate objects (e.g *doll*, *tummy*, *shirt*).

I observed an interesting trend in what kind of left-contexts are most fragmenting. In the top-10 most fragmenting words in age group 1, I found multiple members of the category SPACE which primarily includes nouns referring to planets in our solar system. By inspecting occurrences in AO-CHILDES, I found that planet words tend to be produced in predicable order, tracking their distance from the sun, akin to enumeration of letters in the alphabet or months in the calendar to promote rote memorization. The consequence of such production patterns is fragmentation of the category such words belong to. For example, *mars* precedes *jupiter* much more often than any other planet, and *jupiter* precedes *saturn* much more often than any other planet. As a consequence of this, each planet is highly predicable given its left-context, which fragments the SPACE category. This, however, does not explain why, say, *cash* is highly fragmenting in AO-CHILDES. When I inspected the corpus, I found that *cash* consistently precedes the word *register* to form *cash register*, but almost never precedes other member of the same target semantic category (i.e. ELECTRONIC). This indicates that left-contexts which are components of compound nouns are another way in which the noun category is fragmented in AO-CHILDES.

Loading	Left-context	Frequency	Loading	Left-context	Frequency
21091.3	the	19328	19021.4	the	17447
9131.1	a	11295	9911.9	a	11538
3172.8	your	6193	3883.1	your	6537
2669.7	.	4512	2202.7	this	1816
2205.6	that	1860	2054.1	little	1972
0.000012	ostrich	1	0.000137	pouched	1
0.000008	mars	1	0.000308	january	1
0.000005	porky	1	0.000805	struck	1
0.000005	mercury	1	0.001131	among	1
0.000000	cash	11	0.001131	senor	1

Table 7.2: Top-5 least and most fragmenting left-contexts in sub-corpus 1 (left) and sub-corpus 2 (right). Sub-corpus 1 and 2 correspond to age group 1 and 2, respectively.

7.4.2 Pre-nominal adjectives

In the next analyses, I further zoom in on the left contexts of nouns in the two AO-CHILDES sub-corpora. Although the density of adjectives does not increase with age (as shown in Chapter 2), it is possible that the pattern of adjective usage in input to children changes with age in a way that does not manifest as a change in relative frequency, but how they pattern with other words. For instance, I predicted that the number pre-nominal adjective modifiers increases with age (i.e. either higher number of tokens or types, or both). To keep overall adjective density constant across corpus partitions, we can imagine a change in which adjectives become increasingly less frequent with copulas (e.g. *is orange*, *are good*) and instead become increasingly frequent in pre-nominal modifier positions.

To answer this question, I first dependency-parsed AO-CHILDES using the Python library *spacy* v2.3.7. Then, for each of 64 corpus partitions, I extracted only those spans which included (i) a probe word and (ii) one or more words that are linked to the probe word under the 'amod' dependency and precedes the probe word. The dependency 'amod' is part of the Universal Dependency (UD) framework. I collected spans with one or more 'amod' dependencies. Examples of the latter are *big black horse* and *exasperated wily cat*, and the top-10 most frequent pre-nominal adjectival modifiers of nouns in AO-CHILDES are *little*, *big*, *good*, *more*, *other*, *red*, *new*, *blue*, *last*, and *many*. For each partition, I counted the number of span tokens, the number of span types, and the type-token ratio (number of span tokens divided by the number of span types).

The results are shown in Figure 7.3. Counter to my prediction, I did not find a reliable increase in pre-nominal adjective usage with age; instead, the results suggest the opposite is true (left panel). Interestingly, there is a temporary spike in the token frequency of identified spans, which demonstrates that the same spans are repeated more often in input to the youngest children in AO-CHILDES. Language directed to younger children, including infants and toddlers, is known to include many repetitions of the same words, phrases, and utterances (Ambridge et al., 2015; Lester et al., 2021), and this, combined with the fact that nouns are more densely represented in input to younger children, is likely the reason for the higher token frequency in the first few partitions of AO-CHILDES. Whether repetition impacts fragmentation cannot be determined without further analyses. Greater repetition may either (i) increase fragmentation if a small set of spans are repeated more often than others, or not affect fragmentation if all spans were equally often repeated, consistent with the counterbalancing requirement.

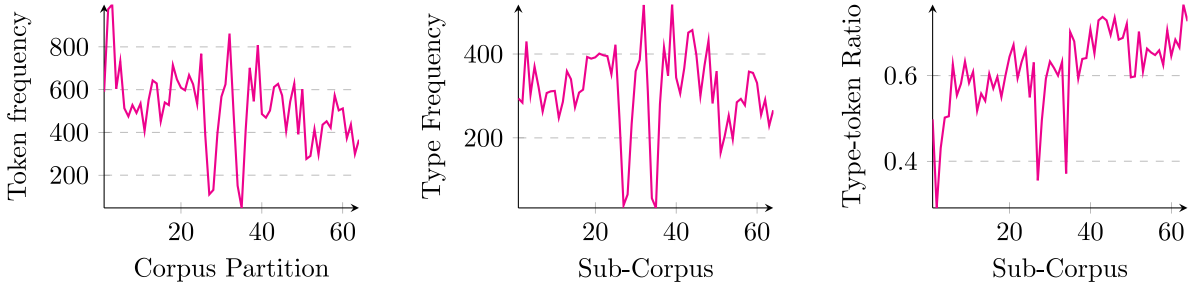


Figure 7.3: Frequency statistics, for each of 64 age-ordered AO-CHILDES partitions, of spans that include a probe preceded by one or more words linked to the probe under the 'amod' dependency (part of the Universal Dependency (UD) framework). The left panel shows the total number of spans, the center panels shows the number of unique spans, and the right panel shows the percentage of spans that are unique (i.e. type-token ratio).

What, then, do these longitudinal trends in pre-nominal adjective usage reveal? They paint a complicated picture: The age-related increase in fragmentation does not appear to be as simple as an increase in pre-nominal adjectival modification. Further, they point to the difficulty in attempting to interpret changes in fragmentation using linguistic constructs and/or uni-variate measures. Uni-variate lexical statistics are insufficient to explain differences in fragmentation, because fragmentation is a multivariate phenomenon. While fragmentation is concerned with the the co-occurrence pattern between observations of two variables (e.g. pre-nominal context and noun), word-level and span-level statistics like frequency and diversity are inherently uni-variate. That said, lexical diversity can be used to approximate fragmentation under certain circumstances: When the type frequency of identified spans (adjective-noun phrase) is extremely low, this almost certainly has a fragmenting effect on the target category, because only a subset of all category members are predicted by a unique left-context. However, when the type-frequency increases, this can be due to one of three reasons: Either, (i) the number of pre-nominal types increases, (ii) the number of target word types increases, or (iii) both types increase simultaneously, and each may affect fragmentation differently. Because it is impossible tell these apart without further analysis, interpretability is limited. To lay this question to rest, I directly computed the fragmentation of pre-nominal adjectives in sub-corpus 1 and 2. To do so, I collected only co-occurrences between probe words and left-contexts that were tagged as adjectival modifiers, and constructed two co-occurrence matrices, one for each age group. Consistent with the frequency-based results, fragmentation did not increase with age; instead fragmentation remained almost perfectly constant (0.793 vs. 0.792).

To zoom in on adjectival modification of probe words, I extracted pre-nominal adjectives that are more frequent in sub-corpus 2 relative to sub-corpus 1. The results shown in Table 7.3. None of the identified words occur in sub-corpus 1. Many of the identified words are highly predictive of individual probe words. For instance, *dutch* precedes *girl* 21 times in sub-corpus 2, and *pickled* precedes *pepper* 19 times in sub-corpus 2. This is in agreement with the hypothesis that adjectives play a role in the fragmentation of nouns in input to older children. However, other words that were identified are examples of minimally fragmenting left-contexts: For instance, the word *live* is shared by many nouns, among which are *zebra*, *tiger*, *child*, *dog*, *bird*, and *tree*. Similarly, *giant* co-occurs with a variety of nouns, including *airplane*, *plant*, *caterpillar*, *spider*, and others.

Left-context	Age Group 2 Frequency
dutch	21
giant	19
friendly	17
pickled	16
pussy	16
live	13
mr	12
smallest	9
angler	9
brave	9

Table 7.3: Left-contexts of probe words that are linked with probe words under the ‘amod’ dependency. Shown are left-contexts with largest increase in relative frequency from sub-corpus 1 to sub-corpus 2 of AO-CHILDES. Because the top-10 words that were identified do not occur in sub-corpus 1, only frequency in sub-corpus 2 is shown. Sub-corpus 1 and 2 correspond to age group 1 and 2, respectively.

7.4.3 Compound Nouns

Given the results of the previous analysis, it appears pre-nominal adjectives do not noticeably contribute to the age-related increase in fragmentation of the noun category in language input to children. What other grammatical phenomena might be involved? One possibility is noun compounding, because knowing the first component of a noun compound makes the second component much easier to predict. For instance, the left-context *cash*, identified previously as belonging to the compound ‘*cash register*’, fragments the category ELECTRONIC by isolating *register* from other same-category members (e.g. *phone*, *radio*). It is possible that compound nouns are less frequent in input to younger children, given that they are longer, and may refer to items that are less available in young children's referential contexts. That said, there are a wide variety of compound types, and one kind in particular frequently shows up in language directed to children: complex proper nouns. Consider, for example, '*woody woodpecker*', '*gingerbread girl*', '*mister potato*', and '*pooh bear*'. This is just a small selection of the most frequent compound nouns used to refer to individuals. Other frequent compounds in AO-CHILDES are common items such as '*scotch tape*' and '*volley ball*'. All of these are likely to fragment the noun category, including the target semantic category structure.

Compound nouns are different in an important way from adjective-noun combinations: While pre-nominal adjectives often do not contribute category-relevant meaning information about the head noun, the initial noun in a noun compound usually modifies the overall meaning of the compound in a way that cannot be discarded as category-irrelevant information. Nouns that participate in a compound form a unit of meaning that changes the overall meaning of the unit; the overall meaning is often very different compared to the meaning of the final noun only. For instance '*teddy bear*' does not refer to a mammal, nor does '*sweetie pie*' refer to a type of food. In contrast, the adjective-noun phrase '*furry bear*' does refer to a mammal, and '*yummy pie*' does refer to a type of food. Of course, the degree to which compound nouns preserve semantic properties of the final noun varies widely. For instance, '*teddy bear*' inherits some semantic properties of living bears, such as shape and texture, but not size or self-propelled motion. Using language-internal statistics to separate these subtle aspects of meaning is challenging, especially with limited language exposure.

To test if noun compounding increases with age in AO-CHILDES, I split the corpus into 64 partitions and counted the number of spans that include (i) a probe word (appendix A) and (ii) one or more words that are linked to the probe word under the 'compound' dependency and precedes the probe word. The results are shown in Figure 7.4. In contrast to my prediction, I did not observe an age-related increase in token (left panel) or type frequency (middle panel). Interestingly the type-token ratio does increase gradually across the age-ordered partitions of AO-CHILDES. However, this must be interpreted with caution: This increase is primarily due to the decrease in the number of tokens. As mentioned previously, the initial spike in token frequency is likely due to repetition which characterizes language to very young children, and the overall greater density of nouns in language to younger children. A decrease in noun density and/or repetition of noun compounds cannot drive fragmentation. I verified this by directly computing the fragmentation of left-contexts of probe words that participate in noun compounds: Consistent with the frequency-based results, fragmentation did not increase with age; instead fragmentation decreased slightly (0.94 vs. 0.90). The evidence at hand suggests we look elsewhere for what drives the age-related increase in fragmentation.

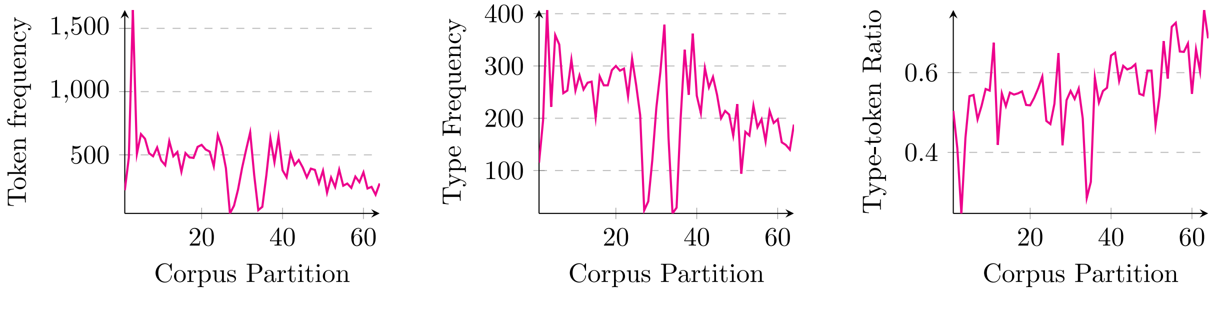


Figure 7.4: Frequency statistics, for each of 64 age-ordered AO-CHILDES partitions, of spans that include a probe preceded by one or more words linked to the probe under the 'compound' dependency (part of the Universal Dependency (UD) framework). The left panel shows the total number of spans, the center panels shows the number of unique spans, and the right panel shows the percentage of spans that are unique (i.e. type-token ratio).

To zoom in on compound noun usage, I extracted those left-contexts that participate in probe noun compounds that are most frequent in sub-corpus 2 relative to sub-corpus 1. Results are shown in Table 7.4. In what follows, I provide some examples of how these words are used in AO-CHILDES. The word *loo* exclusively precedes the word *cuckoo* in sub-corpus 2 where it presumably functions as a playful rhyme, and which is repeated relatively frequently (25 times) in different variations: '*cuckoo loo cuckoo coo loo cuckoo*', '*cuckoo cuckoo loo cuckoo coo loo cuckoo*', '*cuckoo cuckoo cuckoo loo cuckoo coo loo cuckoo*', among others. This may not be a compound as traditionally defined, but was identified regardless by the UD dependency parser included in *spacy* v2.3.7. The word *care* occurs 25 times in the compound '*care bear*'; the word *toe* occurs in '*toe cream*', '*toe shoes*', *and toe bones*'; the word shark exclusively participates in the compounds '*shark flounder*', '*shark meat*', '*shark teeth*', '*shark nose*; *wars* is part of the compounds '*star wars book*', '*star wars video*', and '*star wars toy*'.

Left-context	Age Group 2 Frequency
loo	25
care	19
toe	16
gold	15
trees	14
pj	12
shark	12
wars	11
cheer	1
loop	11

Table 7.4: Left-contexts of probe words that are linked with probe words under the 'compound' dependency. Shown are left-contexts with largest increase in relative frequency from sub-corpus 1 to sub-corpus 2. Because the top-10 words that were identified do not occur in sub-corpus 1, only frequency in sub-corpus 2 is shown. Sub-corpus 1 and 2 correspond to age group 1 and 2, respectively.

7.4.4 Possessive Marking

What other linguistic explanation could be tested? One possibility is that greater use of possessive noun phrases where the subject is a proper noun could protect nouns from fragmentation in input to younger children. Due to Byte-Level Byte-Pair Encoding (BPE), the singular possessive marker *'s* is frequently split and treated as a separate vocabulary item. This means that transcribed text such as '*John's book*' is actually represented by two separate units – *John* and *'s* — in the RNN. Because there is only one singular possessive marker in English, it does not provide a whole lot of semantic information that could be useful for predicting specific nouns or sub-sets of nouns; it is much more likely that the possessive marker groups together nouns indiscriminately, if used frequently.[5] Analyses of two-item spans in AO-CHILDES that (i) end in a probe word, and (ii) start with the possessive marker *'s* revealed that this is indeed the case: There are more than twice as many such spans (2,704 vs. 1,228), and the possessive marker occurs with a wider range of probe words (276 vs. 220) in partition 1 compared to partition 2. Probe words that frequently follow the possessive marker include *foot*, *hair*, *room*, *nose*, *microphone*, *mouth*, *sister*, *shoe*, *lap*, *book*, *shirt*, *clothes*, *camera*, and

[5] A χ^2 feature selection analysis showed that the possessive marker is the fourth most noun-diagnostic left-context, behind the class of determiners, possessive pronouns, and adjectives. In this analysis, only part-of-speech tags were used as features.

many others. Moreover, the column in the co-occurrence matrix that contains co-occurrence counts for probe words and the possessive marker is more entropic (2.32 vs. 1.80) when constructed using partition 1 compared to partition 2 of AO-CHILDES. Further, two of the 5 words with the steepest increase in density across age-ordered partitions of AO-CHILDES are the possessive pronouns *her* and *my*, which suggests that the unitary possessive marker is gradually replaced by a larger set of possessive pronouns as children grow older.

7.5 Summary

In sum, the corpus analyses reported above demonstrate clear evidence for an age-related increase in fragmentation of nouns. Both pre- and post-nominal contexts are affected. The former observation, in particular, provides substance to the claim that an RNN trained on language input to children first would learn more atomic lexical semantic representations of nouns than an RNN trained on input to older children first. This prompted the question: Which grammatical phenomenon is behind the age-related increase in fragmentation of pre-nominal contexts? I predicted that pre-nominal adjectival modification and/or noun compounding may increase with age, but did not find evidence for either. Instead, I found that the possessive marker '*'s*' is more frequent in input to younger children, occurs pre-nominally with a wider range of nouns, and is less predictive of individual nouns — and therefore less fragmenting — in input to younger compared to older children. This observation is in line with intuition, namely that when talking to younger children, caregivers likely refer to individuals more often by name (i.e. using proper nouns), and eschewing possessive pronouns in favor of possessive proper noun constructions (e.g. *John's* instead of *his*).

The analyses presented in this chapter are in agreement with findings by Jiang et al. (2020a) who also studied age-related changes in the distributional statistics of child-directed language. Similar to my own analyses, the authors used a more sophisticated analysis in which the vector space of a Word2Vec model was investigated at consecutive intervals during training on age-ordered child-directed input. Their results showed that many words occur in a richer set of lexical contexts across developmental time and that this impacted the organization of the vector space that was learned by Word2Vec.

One limitation of my corpus analyses is that I focused on contexts immediately adjacent to nouns. However, an RNN can have access to non-adjacent noun contexts, by predicting next-words across long-distances, on the order of tens to hundreds of items. In this thesis, the longest distance available to the RNN is 7. While my analysis of fragmentation is primarily focused on adjacent relationships, fragmentation may be exacerbated by non-adjacent left-contexts. Importantly, fragmentation may only decrease, but never decrease, when expanding the window of analysis further backward.

Another limitation concerns cross-linguistic generalization. For instance, in English, gender and number are not marked on pre-nominals (e.g. determiners, adjectives), and this enormously simplifies the distributional discovery of the noun category by reducing fragmentation. The more diverse the inflectional marking on nominal contexts, the more fragmented the noun category becomes in languages where such marking is more prevalent. It is possible that in such languages, the relationship between noun category fragmentation and age may be null, or reversed. Interestingly, however, languages tend to be biased towards suffixes rather than prefixes (St. Clair et al., 2009), which would have benefits for the discovery of lexical categories via next-word prediction.

A final limitation concerns my exclusion of context words at non-adjacent distances. When using only information about adjacent contexts, we cannot distinguish between situations in which a signal that is diagnostic of superordinate category membership is truly absent, and situations in which the same signal is present but at a greater distance from the target word. For instance, consider the distinction between count vs. mass nouns, the latter of which is signaled by the presence of the context '*some X*'. Consider also sentences (a) and (b) below, which both provide the diagnostic signal '*some X*' for the mass nouns *juice* and *coffee*.

(a) Can I have some juice please?

(b) I want some more of that coffee!

The difference is that the diagnostic signal in (b) is not adjacent to the target word as in (a). Should this be counted as an example of fragmentation of the mass noun category? One could argue that it should not. But analyses of contexts restricted to adjacent words are blind to the possibility that category diagnostic signals — while absent nearby — may be present elsewhere in the sentence.

In the next chapter, I further examine fragmentation of noun contexts, from the perspective of information theory. In particular, I examine how fragmentation relates to redundancy in the sequential statistics of natural language data.

Chapter 8

Entropy Maximization

This chapter explores an information-theoretic perspective on age-related fragmentation in children's language input. The focus is on understanding how statistical shifts in language data, such as lexical and combinatorial diversity, can bring about changes in the predictability of linguistic units, and especially nouns. Many parts of this chapter have been previously published in a peer-reviewed journal (P. A. Huebner & Willits, 2021b).

8.1 Factors that Influence Fragmentation

Why is the noun category less fragmented in input to younger compared to older children? In particular, I focus this question on left as opposed to right contexts, given the importance of the counterbalancing requirement (SPIN theory, Chapter 6). In the previous chapter, linguistic analysis suggested that possessives (but not adjectival modification or noun compounding) better group together nouns in input to younger compared to older children, and that the decrease in possessive noun phrases with development may contribute to the fragmentation of the noun category. What other kinds of analyses might help us understand how language statistics contribute to the age-related fragmentation of nouns in child-directed input? In this section, I briefly discuss the relationship between fragmentation and combinatorial diversity (e.g. see Chapter 2 and 7), and between fragmentation and an information-theoretic quantity called conditional entropy. While each can potential provide useful insights about fragmentation, I will argue that understanding fragmentation from an information-theoretic perspective is potentially more revealing. The primary motivation to move beyond measures of diversity (i.e. number of unique words or word types) is that such measures are not sensitive to the frequency distribution of language units, nor how the shape of the distribution influences the predictability of next-words.

8.1.1 Combinatorial Diversity

To be clear, combinatorial diversity and fragmentation are distinct concepts. Strictly speaking, fragmentation is the extent to which contexts are *shared* among members of the same category, while combinatorial diversity is the number of unique context + category-member co-occurrences. To illustrate the difference, consider a hypothetical corpus where nouns occur in a large number of unique contexts, but each noun occurs with the same contexts and with equal probability. In this corpus, the combinatorial diversity of noun phrases would be high but fragmentation would be zero. Alternatively, it is possible to construct a corpus where each noun occurs with exactly one unique context, but each noun occurs with a different context. In such a corpus, combinatorial diversity would be minimal, but fragmentation of the noun category would be maximal. Thus, one could say that the relation between diversity and fragmentation is U-shaped such that both minimal and maximal diversity result in the least amount of fragmentation, and intermediate values of diversity result in the largest amount of fragmentation. With this in mind, an explanation of fragmentation in terms of combinatorial diversity assumes that we are operating in the regime in which combinatorial diversity is low to medium — the range in which fragmentation increases in proportion with an increase in combinatorial diversity, on average. It is very likely that the natural language statistics operate in this regime due to high sparsity in co-occurrence data.

One of the reasons that combinatorial diversity is not a particularly insightful explanation of fragmentation is that combinatorial diversity collapses a large variety of distinct possibilities that a researcher might wish to tease apart. I can think of at least four qualitatively distinct scenarios that can increase the combinatorial diversity of a given language sample: (i) adding novel words in a familiar syntactic role, (ii) adding novel words in a novel syntactic role, (iii) using familiar words in novel syntactic roles, and (iv) increasing the set of possible combinations. I discuss each in turn below.

Adding Novel Words in a Familiar Syntactic Position

One way to increase the combinatorial diversity of noun phrases is to expand the set of unique pre-nominals that are a member of a familiar grammatical category. In turn, this would increase the likelihood that a novel noun phrase is produced where the pre-nominal only occurs in that noun phrase (or a few others) but not many others. In particular, the introduction of highly infrequent pre-nominals would be especially effective at increasing the fragmentation of nouns: Infrequent pre-nominals may occur too infrequently for a learner to collect a large enough sample to notice that they are potentially systematically shared across nouns.

Importantly, a speaker may increase his or her inventory of pre-nominals independently of an increase in syntactic diversity by enlarging the set of words that occur in syntactically identical positions. For instance, one might expand the set of unique pronouns (*his*, *your*), adjectives (*green*, *big*), quantifiers (*many*, *some*), and numerals (*one*, *two*) provided a speaker has already used members of these syntactic categories in previous utterances.

Adding Novel Words in a Novel Syntactic Position

Second, it is possible to increase the combinatorial diversity of noun phrases by adding novel pre-nominals to the vocabulary that are not members of an existing syntactic category. For instance, a speaker may start with a basic set of pre-nominals, such as demonstratives and the definite and indefinite articles, of which there are only a small handful (*this*, *that*, *the*, *a*, *an*), and diversify this initial set of pre-nominals by adding novel pre-nominals to the vocabulary that belong to novel grammatical classes (e.g. pronouns, adjectives). In this scenario, both lexical diversity *and* syntactic diversity are increased simultaneously.

Using Familiar Words in Novel Syntactic Positions

Third, it is possible to increase combinatorial diversity by keeping vocabulary size (i.e. number of unique pre-nominals) constant, and expanding one's syntactic repertoire instead. For instance, a speaker might use familiar (i.e. previously used) nouns in a wider range of syntactic positions (e.g. subject and object position), or use familiar adjectives in a wider range of positions (e.g. not only as predicates but also as pre-nominal modifiers). One or both could drive up combinatorial diversity of noun phrases.

Increasing the Set of Possible Combinations

Fourth, it is possible to increase combinatorial diversity without adding novel words or moving familiar words to novel syntactic positions. Instead, a speaker may relax existing syntactic and semantic constraints, add novel abstractions that permit a wider set of possible lexical combinations, or use familiar words in familiar syntactic constructions in more creative ways. For instance, a speaker may have started out restricting the usage of the adjective *sweet* to talk about deserts; after some time, the same speaker may come to use sweet with a wider range of nouns (e.g. '*sweet boy*', '*sweet girl*'). In this scenario, what changed is not necessarily the size of the vocabulary or the the distribution of words over syntactic positions, but the rules or regularities responsible for generating language.[1]

Why Combinatorial Diversity is not Helpful

In sum, while combinatorial diversity is a straightforward concept and easy to quantify, it does little to help us gain insight into the age-related increase in fragmentation of nouns. As discussed above, there are many qualitatively different scenarios (i.e. reasons) that can give rise to an increase in combinatorial diversity, and the measure itself does not tease these different scenarios apart. More importantly, while fragmentation and combinatorial diversity are related, there are many ways to increase combinatorial diversity without increasing fragmentation — in some cases, fragmentation might even decrease.[2]

8.1.2 Conditional Entropy

There is another way to think about fragmentation that is particularly useful in the context of developing and testing models that learn via next-word prediction. It has to do with the ease with which nouns can be predicted given their contexts. This means we are entering the territory of information theory. The reader might ask: Why do we need information theory? The answer is simple: Fragmentation is not an intuitive concept, and cannot be straightforwardly interpreted or broken down for further analysis. Doing so using statistics like combinatorial diversity is not helpful, because combinatorial diversity (i) captures only a limited range of univariate statistical properties relevant to fragmentation, and (ii) collapses a broad range of distinct corpus-statistical phenomena. An information theoretic analysis is useful because (i) it can provide insights that are less opaque than multivariate concepts in linear algebra (e.g. fragmentation), (ii) is well known and widely used in computational linguistics, (iii) makes contact with loss functions to train modern neural network models, and (iv) is mathematically powerful enough to capture statistical phenomena that simple diversity-based measures do not capture.

In a nutshell, information theory is a toolbox for quantifying predictive uncertainty. One tool in particular will be of great use in this chapter, namely conditional entropy. A change in the statistical association between nouns and their neighbors can be quantified using conditional entropy. By itself, entropy measures the average

[1] In fact, this is not the only explanation. A speaker may use the adjective *sweet* in a novel combination because, previously, the speaker had no communicative reason to do otherwise. Put differently, the production of a novel combination may not have been triggered by an internal change in the language system, but by an external change in the socio-communicative environment that prompted the production of that phrase. It should be obvious that questions like these cannot be answered using corpus analyses.

[2] To illustrate this point, consider that almost all nouns in a corpus occur with a particular context C. It is trivial to increase the combinatorial diversity by adding novel noun phrases that 'fill in' this gap in the co-occurrence matrix. In particular, we may add combinations of nouns that have not previously occurred with context C. Whereas the combinatorial diversity would increase in this hypothetical scenario, fragmentation would decrease.

difficulty of predicting the outcomes of a random variable. In this case, we are dealing with two random variables; underlying the observations of nouns and their context are two discrete random variables which I will refer to as X and Y, respectively.[3] It is possible to compute either the entropy of contexts conditioned on nouns ($H(Y|X)$), or the entropy of nouns conditioned on contexts ($H(X|Y)$). In the former, we imagine that a noun has been observed and use this information to predict which context will occur, while in the latter, the relationship is reversed. To illustrate the difference, consider we have observed the determiner *the*. It is incredibly difficult to predict the identity of the upcoming noun given that so many nouns are equally likely to follow the determiner *the*. Because our uncertainty is high in this situation, the conditional uncertainty, $H(X|Y)$, would also be high. However, the opposite is true in the reverse direction; given we have observed a particular noun, we can be relatively certain that it was preceded by the determiner *the*. As a consequence, $H(Y|X)$ would be low. This example is restricted to left contexts; however, conditional entropy can also be used to examine the relationship between nouns and their right-contexts. Consider a learner has observed a noun and is tasked to predict the next word. Because periods frequently follow nouns in AO-CHILDES, this results in low predictive uncertainty and therefore a small value of $H(Y|X)$. On the flipside, due to the semantically uninformative nature of punctuation, it is incredibly difficult to predict in the reverse direction, and this results in a comparatively larger value of $H(X|Y)$. In other words, it is extremely difficult to predict *which* noun occurred when the only information available is that it preceded a punctuation symbol.

8.2 Entropy-maximizing Contexts

In this chapter, I will use conditional entropy to formulate a deeper understanding of why the noun category is less fragmented in language to younger children. My proposal is based on uncertainty maximization, which is essential for category formation (Chapter 1) and 3), and for combating chunk-level memorization in the RNN. At the core of this proposal is the idea that the contexts in which nouns are presented to younger children are uncertainty maximizing, or 'entropy-maximizing' to use the technical terminology.

What are entropy-maximizing contexts? They are contexts that occur indiscriminately with many members of a category. For instance, consider right-contexts such as *and*, *to*, *with*, and left-contexts such as *the*, *a*, *that*, and *my*. These contexts are all relatively semantically uninformative, and can therefore occur with all nouns with relatively similar likelihood. Due to their relatively equal probability of occurring in the context of nouns, these contexts drive up $H(X|Y)$, the difficulty of predicting which noun is likely to occur (given that we have already observed the context).

The idea behind entropy-maximizing contexts has much in common with the notion of anchor points, which are thought to facilitate the discovery and abstraction of a set of linguistically similar words (Cameron-Faulkner et al., 2003). However, I prefer the term entropy-maximization for two reasons. First, it is more descriptive of their function than the relatively abstract notion of anchoring. Second, an anchor point is a relatively discrete notion; it implies a context is either an anchor point or it is not, and provides no prediction or mechanism for how anchor points may differ quantitatively in their value as a cue to categorization. By contrast, entropy is a quantitative concept, which enables us to empirically measure — on a graded scale — how entropy-maximizing a context actually is (Matthews & Bannard, 2010).

[3] Given a co-occurrence matrix, I treat rows as samples drawn from an unknown discrete random variable X (nouns), and columns as samples drawn from a different unknown discrete random variable Y (noun contexts).

The reason I think that entropy-maximizing contexts are useful for explaining the reduced fragmentation of nouns in input to younger children are twofold: First, entropy-maximizing contexts — just like anchor points — are likely over-represented in speech to younger compared to older children (Cameron-Faulkner et al., 2003). Second, entropy-maximization allows us to explain the absence of age-related fragmentation of nouns observed when the columns of the co-occurrence matrix were normalized (Chapter 7). Because a pre-requisite for entropy-maximization is high frequency, the columns corresponding to entropy-maximizing contexts in the co-occurrence matrix must account for disproportionately large amounts of variance relative to other non-entropy-maximizing contexts. This point is explained in greater detail below. Collectively, these two lines of evidence point towards an explanation of the age-related increase in fragmentation of nouns in terms of a disproportionately larger number of entropy-maximizing nominal contexts in speech to younger children.

8.2.1 Predictions

My explanation of the age-related increase in fragmentation is that over development, there is a distribution shift from nouns occurring in entropy-maximizing contexts (e.g. anchor points like determiners and punctuation symbols that occur relatively equally often with all nouns) to fragmenting contexts (e.g. contexts that provide information specific to individual nouns or specific to noun subcategories). In this chapter, I test this idea empirically using a combination of simulation work and corpus analyses. First, I translated this informal notion into precise quantitative predictions about the pattern of conditional probabilities I expect to observe under the assumption that my hypothesis is correct. I obtained these predictions from a set of simulations in which artificially generated co-occurrence matrices, modeled after those used in the previous chapter, were compared under different assumptions about the number and importance of entropy-maximizing contexts. Once these theory-driven predictions were derived, I compared them to the actual results, obtained for AO-CHILDES data.

The outcome I simulated is the difference in the conditional entropy, $H(X|Y)$, between age group 1 and 2. To simulate this difference, I generated co-occurrence matrices for each group in an experimentally controlled fashion. As a starting point for each matrix, I adopted the average shape and sum of the co-occurrence matrices obtained for age groups 1 and 2 (Chapter 7). However, rather than using actual co-occurrence data, I filled each matrix using a random sampling procedure. I divided this procedure into two steps, one for entropy-maximizing contexts, and one for the remaining contexts. Importantly, columns corresponding to non-entropy-maximizing contexts were filled by sampling from a pseudo-Zipfian distribution. [4] In contrast, columns corresponding to entropy-maximizing contexts were filled by sampling from a random uniform distribution, given that as the number of observations increases, the random uniform distribution maximizes entropy — the pseudo-Zipfian does not. In order for entropy maximization to take place, however, one additional parameter needed to be set: The proportion of total co-occurrences that are due to entropy-maximizing contexts vs. non-entropy-maximizing contexts. I set this proportion to 0.9 (for both age groups), given that most nouns in naturalistic data occur with canonical, high-frequency neighbors (e.g. *the*, *an*, etc.). Let's call this proportion ρ_b where b stands for baseline. Given that ρ_b is above 0.5, this means that, during the generation of co-occurrence matrices, columns corresponding to entropy-maximizing contexts were filled

[4] The frequency of each pseudo-type is simply proportional to the inverse of its rank, which is arbitrarily defined as its order in the simulated vocabulary. However, the results of the simulation are invariant to the choice of distribution for non-entropy-maximizing contexts.

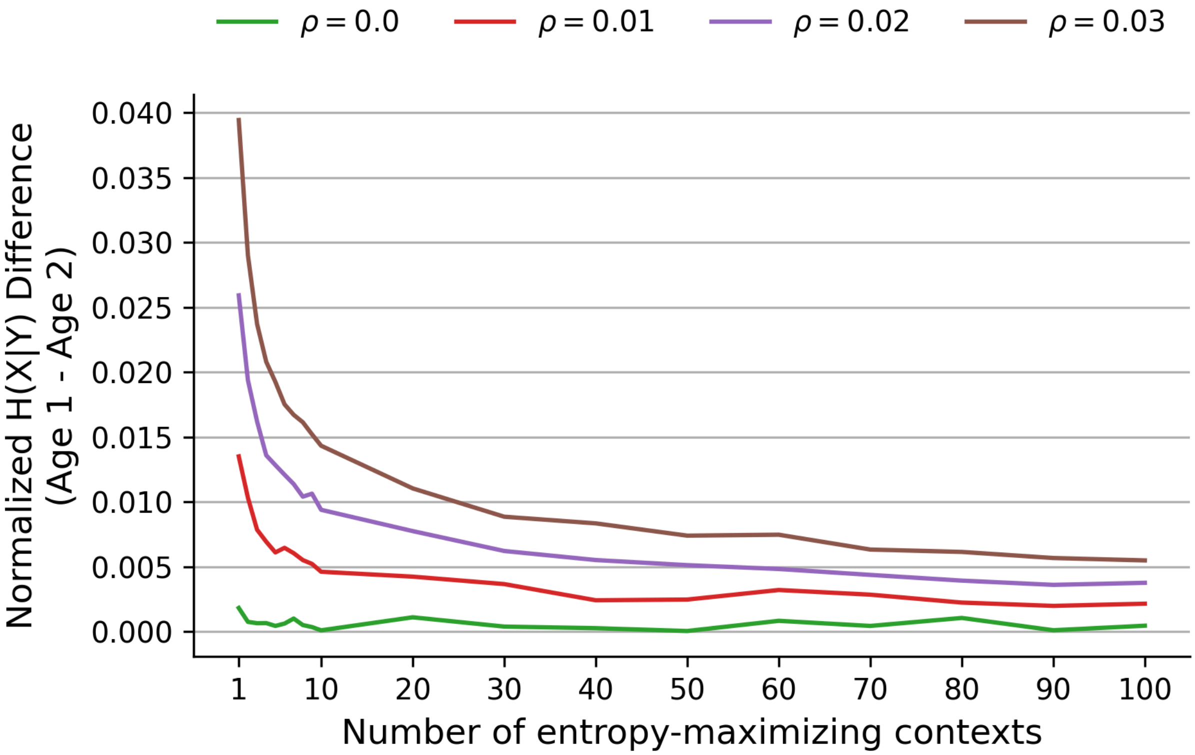


Figure 8.1: Simulated differences in $H(X|Y)$ between age group 1 and 2 based on theory. Each line shows the expected relationship between the difference in $H(X|Y)$ (age group 1 - 2) and the number of entropy-maximizing contexts, given ρ. $H(X|Y)$ is the conditional entropy of pseudo-nouns, X, given the contexts, Y, in which they occur. ρ is the additional proportion of total co-occurrences that involve an entropy-maximizing context for age group 1 relative to age group 2. For instance, when $\rho = 0.00$, the artificial co-occurrence matrices were generated identically, and only differed in shape (650 rows and 2000 columns for age group 1; 680 rows and 2600 columns for age group 2; modeled after sub-corpus 1 and 2 in AO-CHILDES). In contrast, when $\rho = 0.01$, the total number of co-occurrences that involve an entropy-maximizing context in age group 1 is 1% greater relative to age group 2. Co-occurrence matrices were generated by randomly filling one of two types of columns: Entropy-maximizing columns were filled by sampling from a random uniform distribution; all other columns were filled by sampling from a pseudo-Zipfian distribution (which does not maximize entropy as the number of observations increases).

much more frequently. This was also influenced by the number of entropy-maximizing contexts, which was varied from 1 to 100. As this number is decreased, and holding the ρ_b constant, the columns corresponding to entropy-maximizing contexts become more densely populated, and are therefore more likely to become more entropic than non-entropy-maximizing contexts.

I simulated the outcome in 4 conditions, each of which differs in ρ, the proportion of co-occurrence observations that involve entropy-maximizing contexts in addition to ρ_b for age group 1. For instance, given $\rho_b = 0.90$ and $\rho = 0.01$, the proportion of total co-occurrence observations that involve entropy-maximizing contexts in age group 1 would be 0.91 (and 0.90 in age group 2). The size of ρ therefore is of primary theoretical interest: The entropy-maximizing hypothesis predicts that when $\rho = 0.00$, there would not be a noticeable difference in $H(X|Y)$ between the two age groups; however, as ρ is increased — and therefore the importance of entropy-maximizing contexts in age group 1 — a positive difference in $H(X|Y)$ between the two age groups should become observable.

The results of the simulations are shown in Figure 8.1. First, I observed that as the number of entropy-maximizing contexts is reduced (from right to left across the x-axis), the difference in the conditional entropy $H(X|Y)$ between age group 1 and 2 increases monotonically (except when $\rho = 0.00$). This is expected, because as the number of entropy-maximizing contexts is reduced, and ρ_b is held constant, the remaining entropy-maximizing contexts are more densely populated and therefore more likely to maximize entropy[5]. Next, the rate at which the difference in $H(X|Y)$ increases depends on ρ. When ρ is small, the difference in $H(X|Y)$ rises more slowly; when ρ is larger, the difference rises more quickly. This makes sense from the perspective of entropy-maximization: As the importance of entropy-maximizing contexts is dialed up in one sub-corpus, contextual information becomes less predictive of target words compared to an otherwise identical sub-corpus where entropy-maximizing contexts are less important.

To summarize, the simulations show that as the contribution of entropy-maximizing contexts is increased in age group 1, the more likely it is that $H(X|Y)$ is larger in age group 1 compared to age group 2. Therefore, if entropy-maximizing contexts play a more prominent role in the distributional patterning of nouns in input to younger children, I expect that $H(X|Y)$ will be greater for age group 1 compared to age group 2. If, on the other hand, no difference or the opposite pattern is observed, I would have to either revise my understanding of entropy-maximizing contexts, or conclude that their role in combating fragmentation in input to younger children is unwarranted. Finally, the entropy-maximizing hypothesis does not have specific predictions for how $H(Y|X)$ should vary between age groups.

8.2.2 Results

The results of the conditional entropy analyses of AO-CHILDES are shown in Figure 8.2. I excluded conditions in which co-occurrence matrices were normalized because conditional entropy requires raw frequency as input. The analyses include both the conditional entropy computed by conditioning on contexts (labeled $H(X|Y)$ on the x-axis), or by conditioning on nouns (labeled $H(Y|X)$ on the x-axis). Y-axis units correspond to bits[6], the number of binary logical states required to encode the amount of 'uncertainty' about the outcome of a random variable. The more bits that are needed to represent the information in a random variable, the more uncertainty exists when predicting its outcomes.

First, I examined the age-related pattern of $H(X|Y)$, for which I developed predictions. These values are represented by the first group of bars in the left panels. Across all conditions in which punctuation was left intact, I observed a larger $H(X|Y)$ for age group 1 compared to 2, as predicted. The average difference was 0.016 +/- 0.004 (std) normalized bits.[7] Because this difference is non-zero and in the direction predicted by simulation, I interpret this as preliminary evidence for the entropy-maximizing contexts hypothesis. Lastly, this effect appears strongly correlated with the presence of punctuation. When punctuation is intact, the conditional entropy is higher for age group 1; conversely, when punctuation is removed, the conditional

[5] Remember that the total proportion of co-occurrences that involve an entropy-maximizing context, ρ_b, is held constant across conditions, at 0.9. Given this constant proportion, a decrease in the number of entropy-maximizing contexts, means fewer columns are filled by sampling from the random uniform distribution. This, in turn, means that the entropy of columns corresponding to entropy-maximizing contexts is closer to approaching the maximum theoretical value of the uniform distribution.

[6] Note, however, that the conditional entropy (bits) was divided by the joint entropy (bits). Strictly speaking, the y-axis units correspond to normalized bits, or percentage of joint entropy.

[7] Analyses of statistical significance are impractical if not unwarranted due to the fact that estimation of conditional entropy produces a single value for each condition, and each condition is an independent experiment.

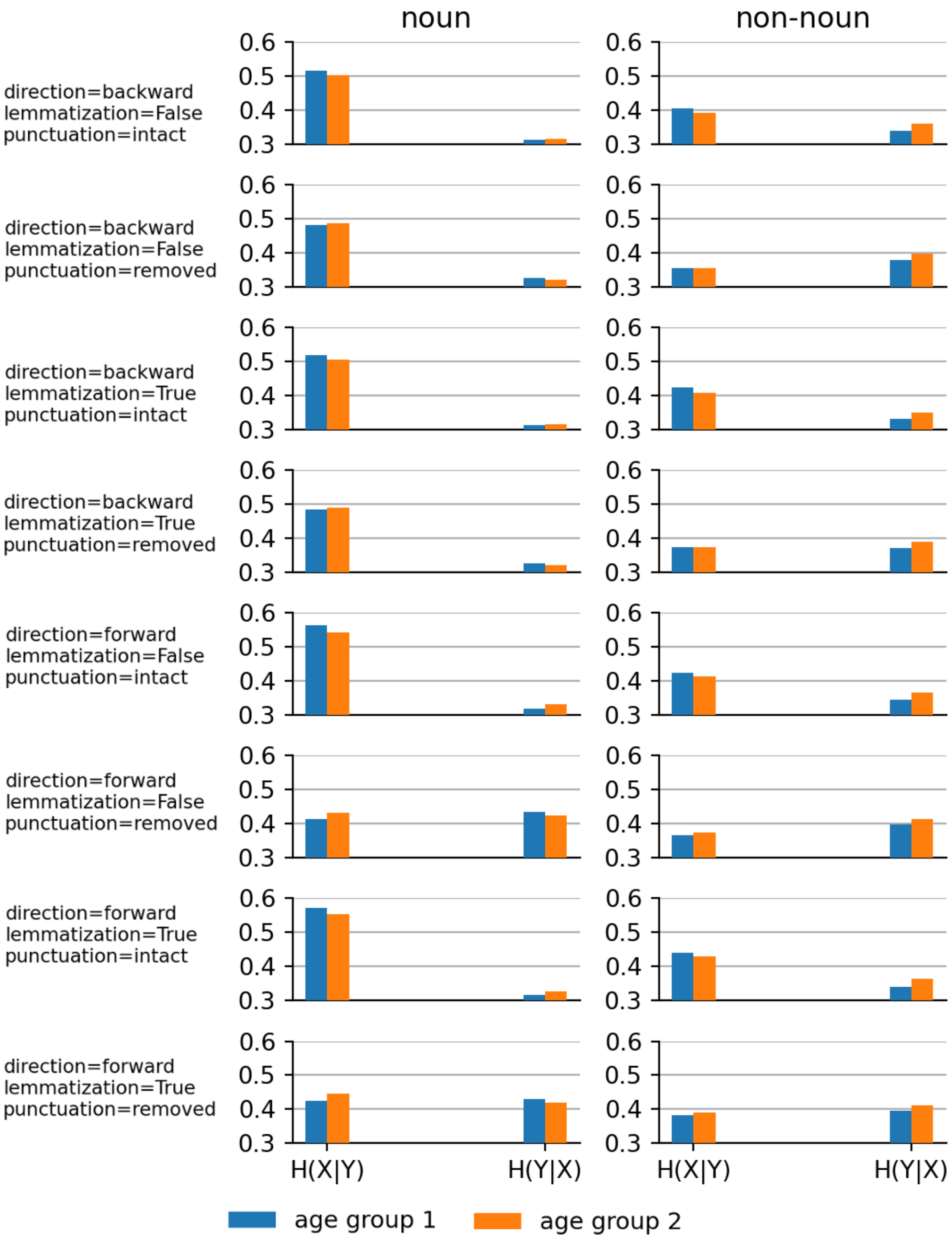


Figure 8.2: Comparison of conditional entropies between age group 1 and 2 for nouns (left panel), and frequency-matched non-nouns (right panel). The three factors, punctuation, lemmatization, and context direction vary top-to–bottom. The y-axis units are normalized bits (number bits in the conditional entropy divided by the number of bits in the joint entropy).

entropy is higher for age group 2. Furthermore, in accordance with the idea that punctuation symbols are strong entropy-maximizing contexts, I found a greater difference in $H(X|Y)$ in conditions in which forward co-occurrences (rows 4–8) as opposed to backward co-occurrences (rows 1–4) were collected.

Next, I compared $H(Y|X)$, shown in the second group of bars in the left panels. The theoretical framework that underlies entropy-maximizing contexts does not make specific predictions here, but this set of results are nonetheless valuable for better understanding how lexical distributional patterns change with age. I found that $H(Y|X)$ is overall lower and differences between age groups are in the opposite direction compared to $H(X|Y)$ across all eight conditions. In addition, in the forward co-occurrence conditions, $H(X|Y)$ is greater when punctuation was removed. This is in agreement with the idea that punctuation symbols are easily predicted given sentence-final nouns, and that it is much more difficult to predict the first word of the subsequent sentence compared to punctuation.

Interestingly, the patterns I have so far discussed also hold when tracking co-occurrence statistics of non-nouns (shown in the right panels of Figure 8.2). Despite that the entropy estimates are overall smaller, their qualitative pattern across conditions is very similar (except for a handful of exceptions) compared to those observed for nouns. This clearly demonstrates a corpus-wide shift in co-occurrence statistics that extends beyond the noun category, and could potentially influence learning of many other lexical classes.

8.2.3 Interim Discussion

Overall, the results of comparing $H(X|Y)$ and $H(Y|X)$ across age groups show that (when punctuation is left intact) over the course of developmental time, it becomes easier to predict nouns given their lexical contexts and increasingly more difficult to predict the lexical contexts that surround nouns. More importantly, I have shown that a theoretical framework based on the idea of entropy-maximizing contexts can account for the pattern of age-related differences in $H(X|Y)$ in input to children. As such, I think the idea that (English-learning) younger children hear nouns in entropy-maximizing contexts more frequently than older children, is a viable hypothesis for explaining the age-related increase in fragmentation of the noun category.

From a bird's eye view, the entropy-maximization hypothesis connects the concept of fragmentation (from linear algebra) to the concept of entropy (from information theory), and therefore allow us to make deeper connections between redundancy, next-word prediction, and, ultimately, SPIN theory and the conditions under which lexical atomicity emerges. In particular, the conditional entropy formalism allows us to succinctly state the connection between fragmentation and the statistical structure of lexical co-occurrence data: Fragmentation occurs when words that belong to the same lexical category (e.g. nouns) tend to occur in contexts that predict individual category members, as opposed to in entropic contexts that do not pick out individual category members.

The results shown above also align well with the findings presented in the previous chapter, in which no age-related difference in fragmentation was found when co-occurrence matrices were normalized. In order to have any effect, entropy-maximizing contexts must be relatively rare. If every context were, in theory, entropy-maximizing, then each column of the co-occurrence matrix would be as minimally populated as possible, and this would prevent contexts from actually maximizing their entropy in practice. Indeed, the simulations show that entropy-maximization in input to younger children only predicts the empirical results under a limited range of conditions — those in which the number of entropy maximizing context is kept relatively small relative to the total number of contexts (approximately less than 10 contexts, see Figure 8.1). This would explain why the age-related difference in fragmentation disappeared when normalizing the co-occurrence matrices prior to comparison (Chapter 7). Normalization removed differences in the amount of total variance that each column could account for, which likely eliminated the potentially disproportionate contribution of a potentially small set of contexts — potentially those that are entropy-maximizing. Consider

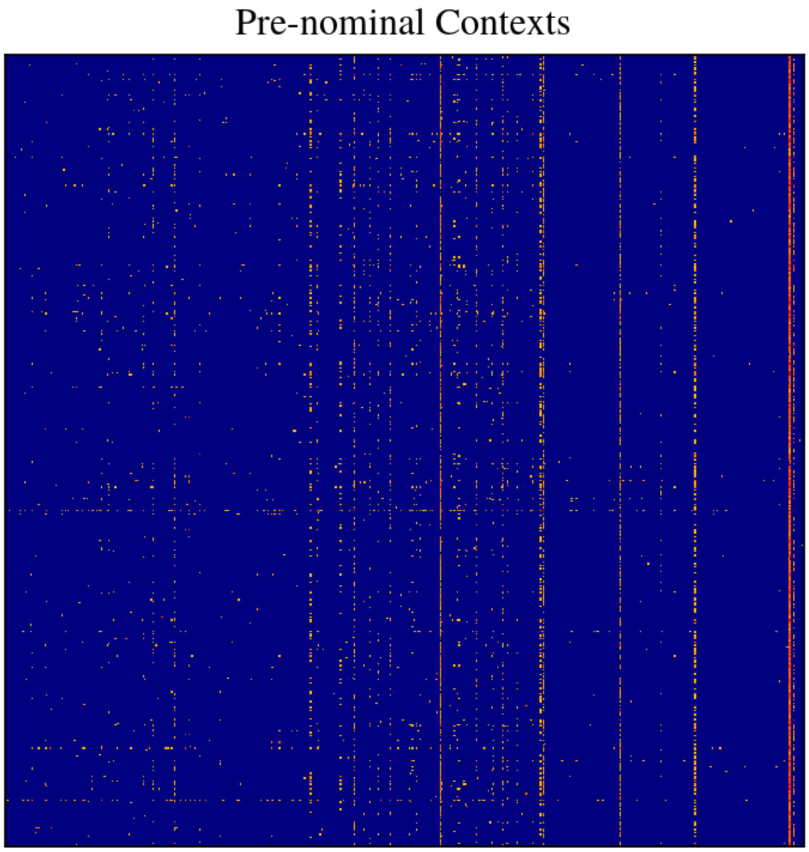


Figure 8.3: The noun co-occurrence matrix for sub corpus 1 of AO-CHILDES. Rows are labeled by nouns, and columns are labeled by adjacent left-context words. The vertical bands clearly demonstrate the presence of so-called ‘entropy-maximizing’ contexts (e.g. *the*, *a*, *your*) in input to English-learning children.

also Figure 8.3, a heatmap visualization of the co-occurrence matrix of sub-corpus 1; entropy-maximizing contexts are clearly visible in the brightly colored vertical bands, and confirms that they are rare relative to other contexts that are much more sparsely observed.[8]

An additional benefit of understanding fragmentation from the point of view of conditional entropy is that it provides insight into the *speed* with which a learning system might discover a lexical category. As the number of entropy-maximizing contexts that are diagnostic of membership in some category, A, increases, the likelihood that each converges on its maximum entropy is reduced, given some constant unit of time. It will take more co-occurrence observations to achieve the same level of entropy relative to another category that is diagnosed by fewer entropy-maximizing contexts. As mentioned previously, when the total number of co-occurrence observations is held constant, entropy is more likely to converge on its maximum value when the number of distributions that are sampled is kept to a minimum (or the number of outcomes per distribution is reduced). This insight is particularly important when developing and analyzing incremental learning systems like the RNN, which must accumulate evidence one step at a time before making inferences about the structure of the input. The rate of evidence accumulation is critically dependent on the number

[8] Of all target word left-contexts, 65% involve one of the top-10 most entropic contexts in sub-corpus 1; in contrast, only 60% involve one of the top-10 most entropic contexts in sub-corpus 2.

of entropy-maximizing contexts, the number of category members, and ρ, the proportion of co-occurrence observations that involve entropy-maximizing contexts. The slower the accumulation of evidence during early training, the more likely it is that the RNN considers inaccurate, and potentially long-lasting maladaptive, hypotheses about the structure of its input.

There are two more technicalities that need to be mentioned. First, I revisit the distinction between partial and perfect redundancy, first discussed in Chapter 5 and 6. Second, this chapter would be incomplete without mentioning synergy, an information-theoretic concept that is closely related to redundancy, and is especially relevant to the language domain where multiple words may interact in ways such that the whole is not the sum of the parts. I discuss each in turn below.

8.3 Perfect vs. Partial Redundancy

A potential limitation in the corpus analyses presented so far is that none of them address the difference between partial and perfect redundancy discussed in Chapter 6 when I first introduced SPIN theory. The aim of SPIN theory is to describe the conditions in which lexical atomicity is *guaranteed* to emerge in the RNN. As such, it does not distinguish between partial and perfect redundancy — no amount of redundancy in left-contexts about upcoming right-contexts is tolerated. As mentioned previously, this requirement is almost certainly too strict, given that useful (while not perfectly atomic) lexical semantic representations may nonetheless be learned when SPIN theory is violated (i.e. when redundancy is low, albeit not zero). On this view, modelers should not take SPIN theory too literal; it is meant as a target worth aiming at, not as a set of conditions that must be satisfied before anything useful can be learned (see discussion on lexical atomicity in Chapter 4).

Violating the counterbalancing requirement is almost certainly inevitable when working with natural language data. Thus, the question of primary concern for modeling the construction of form-based lexical semantic categories with the RNN is not how to *guarantee* the fulfillment of the counterbalancing requirement, but how much useful lexical semantic category knowledge can be learned *despite* violating the counterbalancing requirement. A strict counterbalancing of natural language data to eliminate any redundant information in left-contexts would, in most situations, be untenable. A perfect counterbalancing of left-contexts with respect to some target semantic category structure would require the exclusion of many words and phrases, or addition of countless more. Therefore, any heuristic interventions or inductive biases that might bring us closer to fulfilling SPIN theory are potentially useful, even if they do not get us all the way there. The staged training regime proposed in Chapter 6 and discussed at length in the next chapter, is one such approach.

Before discussing the staged training strategy, it is important to investigate whether the distinction between partial and perfect redundancy buys us anything for better understanding the age-related increase in the fragmentation of nouns in AO-CHILDES. Remember that this distinction was important for explaining the behavior of the RNN in the artificial language learning simulations reported in Chapter 5. The results clearly showed that full atomicity could be achieved in the partial-redundancy conditions, and failed only in the perfect-redundancy condition. However, it is not clear how generalizable these findings are to natural language corpora, where the co-occurrences matrices are much more sparsely populated.

8.3.1 FTP1 Combinations

I now turn to an idea developed in Chapter 6 regarding 'perfect-redundancy' pockets, which are hypothetical partitions of a corpus that happen to isolate occurrences of left-context + probe combinations where the left-context provides exactly the same information about a category-relevant right-context as the probe word. In such combinations, the left-context perfectly predicts the probe word, and thereby inherits all the distributional semantic properties of the probe word. I call such combinations FTP1 combinations, which stands for Forward Transition Probability (TP) = 1.0. This is true regardless of the backward transition probability, because the RNN examined in this work is uni-directional, processing its input left-to-right. For instance, if the word *big* always precedes *gorilla*, it inherits all of the predictive properties of *gorilla* regardless of the number of times that *gorilla* is preceded by *big*. For the interested reader, however, I also report the backward TP for each identified FTP1 combination.

Perfect-redundancy pockets may be created, for example, when an arbitrary corpus partition isolates one ore more target word that (i) occurs infrequently, (ii) is consistently used as part of a compound, (iii) is consistently modified by the same infrequent adjective, or (iv) simply due to chance given a sufficiently diverse corpus and a sufficiently small partition size. Whether a partitioning of AO-CHILDES creates perfect-redundancy pockets is the question I address in this analysis. To begin, I split the corpus into 8 partitions, because this is the number of partitions used to train the RNN on AO-CHILDES in Chapter 10, and then pre-processed each partition using the same tokenization strategy used to pre-process the input to the RNN. Details regarding tokenization are discussed in Chapter 10. For each partition, I counted the number of FTP1 combinations and then computed the backward TP (i.e. the proportion of times that the probe word is preceded by the left-context).

8.3.2 Results

The results are shown in Table 8.1. I reported only the results for partition 1 and 8 of AO-CHILDES as these are the first partitions that an RNN is exposed to when trained incrementally.[9] I found 36 FTP1 combinations in partition 1, and 33 FTP1 combinations in partition 8. Only the top-10 FTP1 combinations are shown, sorted by their backward TP. Not surprisingly, there are only a small handful of FTP1 combinations with a backward TP of 1.0 — there are 3 in partition 1, and 1 in partition 8. Given that the number of FTP1 combinations in partition 1 and 8 are comparable, it appears that perfect-redundancy pockets are unlikely to play an explanatory role in the event that atomicity differs between RNNs trained incrementally in age-order (partition 1 first) and RNNs trained in reverse age-order (partition 8 first).

However, the analysis of FTP1 combinations should be taken with a grain of salt. In order to actually impede atomicity of learned lexical semantic representations in the RNN, an FTP1 combination must be followed by a category-relevant distributional signal. Only then, can it be said that a left-context is perfectly redundant with a target semantic category signal. Because it is difficult to tell prior to running RNN simulations which right-contexts are diagnostic of semantic category membership, I did to conduct such an analysis. That said, there is no reason to think that the right-contexts of identified FTP1 combinations differ systematically in their amount of semantic content between the two partitions.

8.3.3 The Relation between Fragmentation and Redundancy

It should be noted that the relationship between fragmentation and information theoretic quantities such as redundancy and conditional entropy is not perfect — they are related, but not identical. For example,

[9] In Chapter 10, the RNN is trained either on age-ordered partitions or reverse age-ordered partitions of AO-CHILDES. In the former condition, the RNN is first trained on partition 1, and in the latter condition, it is first trained on partition 8.

LC	P	LC Fr.	P Fr.	Backw. TP
_gran	pa	1	15	0.066
_booster	seat	5	66	0.075
_training	baseball	1	13	0.076
_living	room	16	172	0.093
_wooden	bench	1	10	0.100
_tweedle	bug	8	79	0.101
pine	cone	7	22	0.318
_cement	mixer	1	1	1.000
roast	beef	11	11	1.000
_spelling	axe	1	1	1.000

CL	P	LC Fr.	P Fr.	Backw. TP
_sour	grape	1	18	0.055
sport	jacket	2	21	0.095
gn	ant	6	62	0.096
lda	zebra	1	5	0.200
_thorn	berry	1	4	0.250
_huckle	berry	1	4	0.250
vet	vest	1	4	0.250
lets	quilt	1	2	0.500
_wom	pa	18	27	0.666
_woody	woodpecker	1	1	1.000

Table 8.1: Top-10 forward transition probability = 1.0 (FTP1) combinations in partition 1 (left) and partition 8 (right) of AO-CHILDES, sorted by the proportion of times that a left-context occurs with one and only one probe in the same partition. LC stands for left-context, P stands for probe, Fr. stands for frequency, and TP stands for transition probability. An underscore in front of a left-context means that the word is treated as a whole word, rather than a sub-word by the BPE tokenizer.

fragmentation is a bivariate relationship (i.e a relationship between two variables), and, in this work, redundancy is a trivariate relationship (i.e a relationship between three variables). While fragmentation describes a relationship between between target words and either their left or right contexts, redundancy is concerned with the amount of information that two of these items provide about the semantic category of the target word (the third variable, category membership). It does not make sense to ask whether two variables provide redundant information — information about what? Because, in this work, the third variable is always semantic category membership, I often use the term redundancy without spelling out this fact explicitly.

In addition, while fragmentation introduces redundancy, this does not tell us whether the redundancy that was introduced is partial or perfect. To illustrate, consider that the adjective *small* always, and only, precedes the word *hamster*; in this case, *small* inherits all of the distributional semantic properties of *hamster* such as reliably predicting upcoming words that are also shared by other members of the ANIMAL category. In contrast, if *small* were to also precede other words, the relationship between *small* and *hamster* would be considered to be one of partial — but not full — redundancy. To tell the difference, additional information-theoretic analyses beyond fragmentation must be conducted. This additional step is important because only the introduction of perfect redundancy is a true violation of the counterbalancing requirement.

8.4 Interaction Information: Measuring Synergy and Redundancy

One last technicality needs to be mentioned. Strictly speaking, redundancy is not the only statistical phenomenon that can drive fragmentation. In information theory, there is a related concept, synergy, which, just like redundancy, describes a particular kind of interaction between more than two variables. In a three-variable system, redundancy occurs when a third variable provides information about the first variable that is already provided by the second variable. In contrast, synergy occurs when a third variable provides additional information about the first variable that is not provided by the second variable, and would not be predictive in the absence of the second variable. While both kinds of interactions can drive fragmentation, I have focused on redundancy in this thesis because it is easier to explain and understand, and because it highlights the negative side of next-word prediction in the RNN. Synergy, on the other hand, is precisely what next-word prediction excels at, and was developed for.

Let's explore the difference in greater detail. Sensitivity and redundancy can both result in the formation of chunk-level representations in the RNN, and each is therefore a potential impediment to lexical atomicity. However, the difference is that, while sensitivity to redundancy does not produce adaptive chunk-level representations, sensitivity to synergy tends to produce adaptive chunk-level representations, and is therefore desirable (a feature not a bug). To illustrate this, consider, again, the difference between adjective-noun phrases and noun compounds. In sentence (a) below, the phrase '*teddy bear*' is a coherent whole, despite being composed of two lexical items. In sentence (b), the noun phrase '*young bear*' can be decomposed without (much) loss of information pertaining to the meaning of the chunk.

(a) My teddy bear broke.

(b) The young bear fell down.

Here, *teddy* and *bear* can be said to interact synergistically with *broke*, because neither *teddy* nor *bear* alone is likely to predict the verb *broke* — only together does the association makes sense. In contrast, the information that *young* and *bear* provide about the upcoming word '*fell down*' is a better example of a redundant interaction, provided that either *young* or *bear* alone are reasonably predictive of *fell down*, and that adding one or the other provides only little additional predictive power. Put difficulty, the specific combination, *young* and *bear*, does not provide much additional information that might be useful for prediction. Keep in mind this example is contrived; a more complete demonstration would require showing additional sentences in the training data to establish precisely what each word predicts alone and in combination with others.

There is another way to view the difference between redundancy and synergy from the point of view of chunk-level formation in the RNN: Whereas redundancy *passively* contributes to chunk formation, synergy does so *actively*. For instance, a compound noun self-maintains its chunk-level representations in the RNN given that it is the compound as a whole that tends to be predictive of upcoming linguistic material. On the other hand, redundancy, while able to introduce chunk-level representations by accident, does not self-maintain those representations. They are eventually broken apart into their components, provided sufficient amount of training data.

8.4.1 Methods

In this analysis, I quantified the degree to which the relationship between probes and their left and right contexts in child-directed input is redundant, synergistic, or neither. To pull this off, I computed the interaction information. This quantity is a generalization of the mutual information to more than two random variables (Jakulin & Bratko, 2003). It enables the quantification of the amount of information that is shared between any two variables (e.g. probe word and right-context) that is shared by a third variable (e.g. left-context). The interaction information can be understood as quantifying the amount of information that is shared by each variable (McGill, 1954). It is close to zero if the variables do not interact with each other, positive if they interact synergistically, or negative if they contribute redundant information about each other.[10]

[10] The following example illustrates redundancy in a three-variable system: Consider that the presence and absence of rain and lightning are highly dependent on each other (each predicts the other). Further, the presence or absence of a storm (e.g. the third variable) interacts negatively with rain and lightning, because it reduces their dependence. Now, consider we ask whether it is raining, given that it is lightning. Knowing that there is a storm already provides us with information about whether it is raining, so knowing that it is lightning is redundant information. The interaction information would be negative.

To quantify how the interaction information changes with age in child-directed input, I first split AO-CHILDES into two equal sized sub-corpora, and collected 77k occurrences of probe words and their left and right contexts in each. The resulting 3-item windows were treated as observations of 3 discrete random variables (probe words, their left contexts, and their right-context). Next, for each sub-corpus, I computed the interaction information of these three variables. In addition to reporting the results after normalization (dividing the interaction information by the joint entropy; normalization = joint-entropy), I also reported the raw interaction information (normalization = None).

8.4.2 Results

The results, shown in Table 8.2 under the 'probe' word-list condition, are largely in agreement with the idea that there are fewer across-target interactions in input to younger compared to older children. The interaction information is lower for age group 1 regardless of the normalization used (joint entropy, None).

Normalization	word-list	Age Group 1	Age Group 2
None	probe	0.314	0.464
joint-entropy	probe	0.023	0.033
None	control	-0.156	0.029
joint-entropy	control	-0.011	0.002

Table 8.2: Interaction Information. When the input consists of three discrete random variables as it does here (e.g. left-context, probe word, right-context), the interaction information is negative when one variable provides information that is redundant with information provided by a second variable about a third variable, and zero if no variable provides information about any other variable.

To better understand whether this age-related trend is specific to the 700 probe words or holds globally in AO-CHILDES, the same analysis was conducted on a size- and frequency-matched set of non-nouns, sampled randomly from the corpus. Using the same sub-corpora, both their left and right contexts were collected, and used as input to the interaction information. The results, shown in Table 8.2 and labeled as 'control', suggest that there is a more global trend in how consecutive words — not just nouns and their contexts — in input to children interact with each other across developmental time. From this, I conclude that there is a broad shift from redundancy towards synergism in the sequential lexical statistics of input to children. The shift from low to high synergy in the co-occurrence structure of nouns can be considered, albeit with caution, as part of that larger trend.

8.5 Summary

In this chapter, I connected the concept of fragmentation with concepts in information-theory to establish a more robust theoretical link between fragmentation and SPIN theory. To begin, I discussed the relationship between combinatorial diversity and fragmentation. Combinatorial diversity undergoes a noticeable age-related increase across developmental time in AO-CHILDES, and makes it more likely that perfect or near-perfect redundancy is created in the input to older children. However, I argued that univariate measures like (lexical and combinatorial) diversity are insufficient for interpreting or explaining fragmentation. As a potential alternative, I introduced the entropy-maximizing hypothesis, which states that noun contexts in younger children are less fragmented because nouns occur more frequently in entropy-maximizing contexts compared to older children. Entropy-maximizing contexts signal the noun category while providing little to no semantic information about which noun might occur. A simulation based on this idea showed that the uncertainty in predicting a noun given its linguistic context should be larger in input to younger relative to older children provided that the entropy-maximizing hypothesis holds. The quantification of the conditional entropy of actual noun-context pairs in AO-CHILDES showed that this is indeed the case, and thus provided empirical support to the entropy-maximizing hypothesis. In the second half of the chapter, I teased apart technicalities related to the information-theoretic angle on understanding fragmentation, namely perfect vs. partial redundancy, and redundancy vs. synergy. First, I showed that, while perfect-redundancy does occur when partitioning AO-CHILDES, it affects input to the youngest and oldest children in AO-CHILDES equally. Second, I showed that synergy is another driver of chunk-level representation formation in the RNN, and that there is a global shift towards synergy in input to children across developmental time.

I am now in a position to discuss the staged training strategy, proposed in Chapter 6, which builds on the concepts of fragmentation and conditional entropy, and with the end goal of promoting lexical atomicity in the RNN in a developmentally plausible manner.

Chapter 9

Staged Learning

This chapter addresses a prediction derived from SPIN theory that should be of broad interest to scholars working at the intersection of machine learning and language acquisition. The hypothesis I propose in this chapter is that the order in which data is presented to the RNN influences the way in which lexical semantic distributional statistics are encoded in the network. In particular, this chapter provides theoretical arguments why training on age-ordered input to children should yield more atomic lexical semantic representations, and empirical findings that validate these ideas. Specifically, I leverage insights of SPIN theory to derive a cognitively and developmentally plausible intervention that seeks to induce a bias for atomicity as early during training as possible. An early bias for atomicity should have long-lasting positive consequences for how category-relevant semantic information is encoded. In particular, a bias for atomicity should help concentrate category-relevant information on the static input-to-hidden weights in the network, where knowledge can be readily accessed and re-used for downstream tasks, such as DECAF. All of the simulations presented herein have been previously made available in a peer-reviewed publication (P. A. Huebner & Willits, 2021a).

The ideas presented in this chapter are related to a classic proposal in the field of language acquisition research, namely that language input to children is staged in order to facilitate acquisition Levelt (1975). The claim is that children are not initially confronted with all the complexity and variety of their native language, and that caregivers introduce novel forms and meanings in a way that is helpful to young learners. Levelt (1975) argues that mothers are effectively narrowing the set of possible generalizations about the structure of language that a child is able to entertain. For example, Levelt (1975) writes that:

> ...the child is presented with grammatical strings from a miniature language, which is systematically expanded as the child's competence grows.

Much effort in language acquisition has been devoted to fleshing out this idea, and to obtain empirical support for this hypothesis (Freudenthal et al., 2006; Hoff-Ginsberg, 1986; Huttenlocher et al., 2002). In the following two chapters, I provide support from computational modeling that this proposal is still relevant today.

9.1 Titrating Separation and Integration across Training

Consider the two ingredients required for the construction of form-based lexical semantic category clusters in the RNN, separation and integration (Chapter 3). I have previously argued that separation is relatively cheap in the RNN, and that integration is relatively more difficult, due to the strong pressure by the next-word prediction objective to discover and exploit differences in word usage — with little regards for whether two words belong to the same semantic category. The counter-force, integration, is what pulls together representations of same-category members, and, ultimately, makes lexical atomicity possible. Without the ability to perform integration, the RNN would be little more than a pattern separator. If so, it would learn little abstract information such as the hidden similarity structure responsible for the patterns it has observed. Clearly, this is not the case. The input features in any connectionist systems with at least one hidden layer (such as the simple RNN) undergo a compression phase in which input features are re-represented in a latent (i.e. hidden) space that typically provides a more compact and/or abstract characterization of the raw input. This distinction is well-known among computational modelers and cognitive scientists. However, this point is easily missed when talking about category formation. Category formation is easily mistaken as a separation problem, rather than a problem that requires both separation and integration. It is not sufficient to be able to distinguish all words in a language (perfect separation); category formation requires integrating similar items so that representations of same-category members are more similar than representations of members that do not share the same category. I will show that the the RNN — at least during the earliest stages of training — tends to be more strongly driven by separation, than by integration, and that by carefully balancing these two opposing forces early during training, we can promote lexical atomicity.

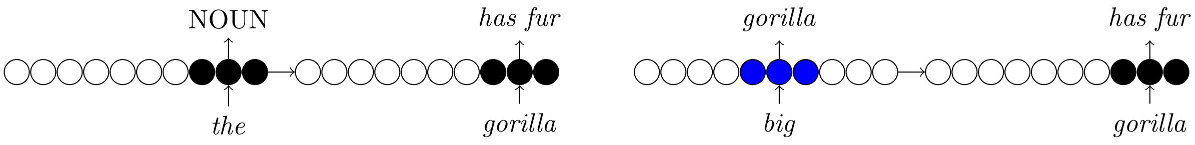
Before diving deep into the details of the proposed staged training strategy, a high-level overview is in order. Broadly, the idea is based on the aforementioned distinction between the two forces that shape the RNN's representational landscaped: First, separation is driven by context-sensitivity, and, second, integration is driven by architectural bottlenecks that promote abstraction. The major contribution of this work is the insight that these two forces may be more useful if deployed at different periods during training. When the RNN is trained as a language model, its embedding space is warped (stretched along one or many dimensions) in correspondence with the statistical properties of the input, to minimize next-word prediction error. Depending on the statistical properties of the items of interest (i.e. their within- and between-category similarity structure) the resulting word embeddings may form distinct clusters or remain spread out with few or no sharp boundaries between clusters. Recent work suggest that clusters approximating part-of-speech categories emerge first, before they are sub-divided into finer-grained topical and/or lexical semantic classes (Saphra & Lopez, 2019). This corresponds with my intuitions about the distributional properties of POS classes and finer-grained semantic categories: The distributional evidence for POS classes like the noun category is more readily available and should therefore be exploited first to minimize next-word prediction error in the RNN. In contrast, semantic distinctions within the noun category (e.g. differences in the distributional properties of *dog* and *airplane*), are less frequent and therefore require more learning trials before they can be readily detected and encoded. In other words, during early phases of draining, the RNN does not yet possess sufficient distributional information that can be integrated to aid formation of semantic category clusters. During this early phase, the RNN is particularly vulnerable to pattern separation because it drives apart items belonging to the same semantic category, before the network has had a chance to form semantic category clusters in the first place. The network's context-sensitivity is thus most problematic during early phases of training, when (idiosyncratic) category-irrelevant distributional signals have the potential to separate same-category members that should be represented close together in representational space.

I propose that in order to learn more atomic lexical representations (and therefore to support DECAF), the RNN must concentrate on structural (i.e. syntactic) regularities *before* learning finer-grained semantic regularities (e.g. selectional preferences of verbs). Put differently, I propose that in order learn useful lexical representations, the RNN must be prevented from over-fitting on idiosyncratic lexical relationships during early stages of training, and instead be preferentially tuned to abstract structural relations while ignoring correlated lexical relationships. This proposal is an attempt to move away from very general claims about the advantages and disadvantages of context-sensitivity and towards a more subtle understanding of the interplay of different constraints on learning at different times during training. This idea is not novel, but has been historically neglected. Emerging evidence supports the view that separation and integration are equally important, albeit at different stages of training. For example, clearly separating the stages during which knowledge about word-order and semantic content is learned, improves the ability of distributional models to capture the meaning of words in context (Erk & Padó, 2008). This idea is intimately related to work in children's language acquisition; specifically, scholars investigating the syntactic bootstrapping hypothesis, claim that children's abstract knowledge of argument structure can scaffold learning the meanings of words (e.g. distinguishing between broad semantic classes of verbs).

A guiding principle of this work is to move away from general claims about context-sensitivity being inherently problematic or necessarily preventing the formation of abstract category knowledge in connectionist systems. Instead, I will show that the lexical semantic representations learned by the RNN can be made more performant in a downstream categorization task (i.e. improved atomicity) if context-sensitivity is carefully titrated against integration rather than removed altogether. To do so, I propose a staged training strategy that leverages insights of SPIN theory (Chapter 6) and the longitudinal structure of age-ordered input to children (Chapter 2, 7, and 8).

9.2 Staged Learning

An RNN that is exposed to distributional signals that cue both a target semantic category and subordinate category will encode each signal in a super-imposed fashion. More specifically, exposure to a sequence such as *the big gorilla has fur*, where *big* cues a subordinate category within MAMMAL, and *has fur* cues the MAMMAL category, will result in a lexical representation of *gorilla* that will resemble that of other members of the MAMMAL category, but also be distinct from other members of MAMMAL that do not co-occur with *big*. To prevent this splintering of the MAMMAL category, these two cues must be encoded separately. But how? Based on SPIN theory, I propose a staged training regime where signals that cue superordinate categories are separated from subordinate category signals by separate training stages. In the first stage, the network is exposed to signals that cue the target category such as *has fur*, but in the absence of subordinate cues such as *big*. Doing so would preserve the coherence of the target category MAMMAL in the internal lexical organization of the RNN. Once an atomic encoding of the MAMMAL category is established, training (gradually) enters a second stage in which semantic cues that signal subordinate category distinctions are used to continue training the model. Importantly, this second stage of training should leave intact the atomic organization established previously. Fragmenting left-contexts, that previously would have resulted in non-atomic organization, should now have a very different effect on the RNN; instead of mixing super- with subordinate category information by using the same set of hidden units to encode both types of information, subordinate category regularities should be encoded by a separate bank of hidden layer units. Thus, the effect of staged learning is to learn two non-overlapping (i.e. orthogonal) sub-spaces in the high-dimensional hidden layer state space. In this way, information captured by the first sub-space does not co-vary with the information captured in the second, orthogonal, sub-space.



(a) Stage 1.

(b) Stage 2.

Figure 9.1: A schematic illustrating how an incremental training strategy and ordered presentation of data may promote more atomic lexical semantic representations of a set of target words such as common nouns. In stage 1 (a), the RNN is presented with data where target words (e.g. *gorilla*) are not accurately predicted by left-contexts, and are followed by semantically informative right-contexts. This results in one bank of hidden units (shown in black) that encodes the target semantic category. In stage 2, redundant information about the target category provided by left-contexts (e.g. *big*) are encoded in a non-overlapping bank of hidden units (shown in blue). The result of this staged presentation is that the RNN separates processing of words that provide redundant semantic information by using two orthogonal sub-spaces, one for processing the left-context, and another for processing the target word. Although both *big* and *gorilla* predict *has fur*, the result is that each word is handled separately by the RNN's hidden layer. Note: Each panel depicts one RNN with 10 hidden units unrolled across 2 time steps. Time steps are separated by the horizontal arrow between the two hidden layers.

9.2.1 Orthogonal Sub-Spaces

To illustrate this idea more clearly, consider Figure 9.1, which depicts a hypothetical RNN with 10 hidden units unrolled across two time steps (separated by the horizontal arrow in each panel). The goal is to learn atomic lexical semantic representations of nouns that belong to one of two semantic categories: MAMMAL and VEHICLE. In stage 1, right-contexts of nouns that cue a target semantic category are encoded by a small subset of all possible hidden units, shown in black. By training on data with minimally fragmenting left-contexts, this subset of units will encode variation between the two target semantic categories. Semantically uninformative left-contexts like *the* will activate this subspace in a way that does not favor one semantic category over another; an uninformative left-context should activate the sub-space such that the RNN predicts members of both MAMMAL and VEHICLE with equal probability. This is illustrated by the label 'NOUN' at the output of the RNN at the first time step in the left panel. In this first stage, the RNN's predicted probability distribution is not yet semantically differentiated, and therefore predicts next-words that are compatible with all possible nouns. Only when this dynamic is firmly established, does training transition to stage 2. In this stage, redundant semantic information may be introduced to the training data without sacrificing atomicity. Note that this staged progression mirrors the age-related increase in fragmentation of pre-nominal context in child-directed input. The result of staged learning is that semantically informative left-contexts such *big*, which predict variation *within* target semantic categories (e.g. large vs. small mammals), are encoded by a set of hidden units that does not overlap with those units that encode variation *between* semantic categories (MAMMAL vs. VEHICLE). To distinguish these two sets of units, the former is shown in blue.

The takeaway message is that in order to distinguish which of two competing cues signals membership in a target category or a subordinate category, the RNN must be presented with each cue separately, in a stage-like fashion. The reason is that separating these two types of information across training ensures that the language modeling objective cannot minimize prediction error by leveraging both types of cues simultaneously. Instead, in order to learn atomic lexical semantic representations, the RNN must first minimize the prediction error associated with predicting category-relevant next-words based only on a target word, rather than both the target word and its left-context. Once prediction error associated with category-relevant relationships is more or less completed, it should be possible to introduce left-contexts into the training data which provide redundant information about the target semantic category structure without sacrificing atomicity. At this point, left-contexts which provide (incidentally) redundant information are no longer useful for predicting semantically informative right-contexts, as the error associated with predicting right-contexts has already been minimized.

By following this recipe, it is in principle possible, that the RNN will learn two orthogonal sub-spaces, each useful at different time steps: The sub-space learned in stage 1 encodes information useful for predicting right contexts of target words, and the sub-space learned in stage 2 encodes information useful for predicting target words. The idea is that when these two sub-spaces are learned separately, they allocate non-overlapping hidden units, which can then be used at separate time steps during processing.

9.2.2 Caveats

I can think of at least three limitations regarding this proposal: First, the example above is an idealization of the actual learning dynamics of the RNN. In practice, it is highly unlikely that the RNN will use distinct subsets of hidden layer units to encode distinct types of information; rather, it is more likely that the RNN will approximate this behavior in a graded fashion. However, the closer the dynamics of the RNN is to this idealization, the more atomic learned lexical semantic representations will be. Second, it is important to note that staged learning does not disentangle regularities that correlate with the target category structure and those that correlate within subordinate category structure; instead, staged learning simply prevents entanglement in the first place. Obviously, a clear stage-like separation of information is often not tenable in the natural world, and this means that staged learning cannot be the only ingredient towards ensuring atomic internal organization. In my view, staged learning is but one of many ingredients that can be of use to statistical sequence learning systems, such as linguistically and/or cognitively informed self-attention Ke et al., 2018; Shen, Lin, et al., 2018. Third, the staged training regime requires that left-contexts of target words are essentially made semantically vacuous. This might have unforeseen negative consequences for learning about the semantic properties of pre-nominals. Given that, in stage 1, pre-nominal contexts would be sampled from the same distribution independent of the choice of noun or the right-context, a child simply would have no distributional evidence for distinguishing the meaning of, say, '*my dog*' and '*your dog*'. The benefit of counterbalancing left-context in stage 1 is to establish a single noun category, but comes at the cost of being able to statistically tease apart left-contexts. Only later, during stage 2, does statistical differentiation of pre-nominal contexts become possible.

9.3 Evidence for Early Semantic Property Inheritance

In the remainder of this chapter, I provide empirical evidence for crucial components of the theory discussed above. My aim is to establish the feasibility of the approach in more structured language learning simulations before applying it directly to RNNs trained on child-directed input. My first experiment was developed to probe for 'semantic property inheritance' by the left-contexts of target words, that is predicted by SPIN theory. In addition, I examine the assumption that semantic property inheritance is most problematic during early training.

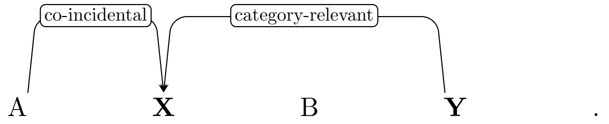


Figure 9.2: The sequential structure of sequences in the artificial corpus. The relationship between X and Y is the category-relevant relationship that the RNN should use for predicting the upcoming Y-word. The co-incidental relationship between A and X is governed by α. When α is large, A and X are more likely predict each other, and as a consequence, A will encode redundant information about the semantic category of the upcoming Y-word. Items in B are sampled randomly.

9.3.1 Artificial Language

First, I created an artificial corpus, similar to the ones in Chapter 5. I did so by sampling sequences from an artificial language which — like natural language but in a more controlled manner — has sequential dependencies that were correlated with a set of externally defined semantic categories. Each corpus consisted of 100,000 sequences, each of which consisted of exactly four words (see Fig. 9.2). The vocabulary was split into four equally sized and disjoint sets, which I refer to as syntactic categories. Each sequence consisted of one word from each syntactic category, occurring exactly in the order A X B Y. Each category, denoted by upper-case letters, was composed of 32 items. I refer to their members as A-words, X-words, B-words, and Y-words, respectively.

As before, I consider X-words to be the target words; it is their representations that are examined. Each X-word belongs to one of four semantic categories, which is defined by the category-relevant relationship between X and Y-words. This relationship is symmetric, which means that Y-word are also divided into four semantic categories. The corpus was designed such that each Y-words had an equal probability of occurring with each of the X-words from the matching semantic category. B-words were sampled randomly and thus did not serve as a reliable cue to semantic category. The critical manipulation concerned the distribution of A-words. A-words varied with regard to the amount of redundant information they provide about the semantic category membership of the X-words that occur in the same sequence. In the control condition ($\alpha = 0.0$), a corpus was created where A-words were not redundant with the semantic category, co-occurring with all words from all semantic categories equally. In the full-redundancy condition ($\alpha = 1.0$), the A-word was, like the X-word, a perfect cue to the semantic category of the upcoming Y-word. In the partial-redundancy condition ($\alpha = 0.5$), each A-word behaved identically to the control condition half the time, and identically to the full-redundancy condition the other half of the time.

The artificial language may at first appearance bear little relation to natural language. However, it has an intuitive relation to English sentences, and which is worth sharing to readers who prefer specific examples to make this relationship explicit. Recall that the language produces strings of the form A X B Y. We are now in a position to imagine pseudo-English examples for such a sequence. Given that each set strictly obeys the positional rules, we can think of them roughly as part-of-speech or phrasal categories (e.g. VP), and their items as representing members thereof. For example, I conceptualized X-words as English nouns, and Y-words as English verb phrases that express actions selectively associated with the entities referred to by the nouns. Items in A and B may be thought of as pre-nominal expressions, and relative clauses, respectively. Thus, we can think of a sequence in this corpus as declarative constructions like (b-d):

(a) A X B Y

(b) [The] doll [over there] [looks funny]

(c) [The red] doll [over there] [looks funny]

(d) [The red] doll [in your hand] [looks funny]

I use brackets to join multiple English words into a single item. Because the purpose of this thesis is learning lexical representations, X-words should be thought of as individual words. All other items in the vocabulary can be thought of as any linguistic expression (e.g. inflectional marker, morpheme, lexeme, phrase). For example, B-words could be an inflectional suffix, like the the English plural marker *-s*.

9.3.2 RNN Training

As before, the RNN was used in a standard language modeling procedure. At each time step, corresponding to one word in one of the artificial corpora, the RNN was trained to predict the next word in the corpus. The weights were updated after each fourth word (corresponding to the end of each sequence) using stochastic gradient descent with a constant, empirically determined, learning rate. By ensuring that there were always exactly four items in the RNN's memory before updating the weights, the training methodology is equivalent to backpropagation-through-time with gradient truncation applied every four time steps.

Instead of tuning the hyper-parameters on the language modeling objective, hyperparameters were chosen that resulted in best performance on the downstream categorization task directly. Because I am interested primarily in performance on this task, this tuning strategy reduced any bias that would have resulted from optimizing the language modeling objective. Hyper-parameters are shown in 9.1; I replicated all experiments using the LSTM to investigate whether the LSTM too is vulnerable to the maladaptive consequences of semantic property inheritance during early training.

hyperparameter	simple RNN	LSTM
window size	7	7
hidden layers	1	1
hidden units	64	64
learning rate	0.4	1.0
optimizer	SGD	SGD
batch size	64	64
steps	3K	3K
non-linearity	*tanh*	*tanh*
initialization	uniform	uniform

Table 9.1: Hyper-parameters used to train the simple RNN, and the LSTM.

9.3.3 Quantifying Atomicity

To quantify the atomicity of the RNN's learned lexical semantic representations, I evaluated how well the representations performed in a downstream semantic categorization task. In this task, performance is based on how well the RNN's learned clustering of X-word representations matched the externally-defined semantic category structure. Given that the task can be solved perfectly by simply tracking the semantic relationship between X and Y-words, performance should be highest if the RNN encoded all category-relevant statistics in the representations of X-words and not in the representations of any other words. That is, a network that learned to track only the relationship between X and Y-words will acquire representations of X-words that capture all the necessary information to achieve a perfect score. However, if any of the information necessary for categorization is encoded by the representations of A-words, then the performance of A-words in the semantic categorization should also yield non-zero performance — and potentially reduce that of X-words. The latter scenario is maladaptive, as information that could otherwise be represented atomically, is instead represented interactively at the processor. Thus, atomicity was operationalized as 1) perfect categorization of X-words, and 2) minimal categorization of A-words.

The measure of categorization performance is based on the highest possible accuracy of correctly deciding whether the representations of two words learned by the RNN belong to the same category. The algorithm for computing balanced accuracy is the same as described in Chapter 5. I have reproduced it below for reference.

The model's semantic category judgements are based on a similarity matrix S obtained by computing all pairwise similarities[1] between all A-word pairs, or all X-word pairs. To obtain the model's learned lexical representations for one set of words, I retrieved the set of weights that connects the localist encoding of a given word to the hidden layer (i.e. the embedding vector). Each similarity in matrix S was used to make a 'same vs. different' judgment within a signal detection framework, tested at multiple similarity thresholds (r = 0.0 to 1.0 with step size 0.001) to determine the threshold for maximum accuracy. If two words with indices i and j belong to the same category, and if $S_{ij} > r$, a hit is recorded, whereas if $S_{ij} < r$, a miss is recorded. On the other hand, if the two words do not belong to the same category, either a correct rejection or false alarm is recorded, depending on whether $S_{ij} < r$ or $S_{ij} > r$. At each threshold, I computed the balanced accuracy by taking the average of sensitivity and specificity. The measure of interest is the balanced accuracy at the similarity threshold which yielded the highest value. I used this process to compute a balanced accuracy score for each model, at consecutive intervals during training. The atomicity of the learned representations can then be determined by inspecting the extent to which the balanced accuracy is high for X-words but low for category-irrelevant A-words.[2] Chance-level performance on this task would produce a balanced accuracy of 0.5.

I should remind the reader again that I am not interested in semantic categorization *per se*. Instead, I use the performance in the semantic categorization task as a proxy for the performance of the distributionally-mediated extension of category-associated features (DECAF). If semantic categorization is poor, then the learned lexical representations will likely be of little use for performing DECAF. In sum, the categorization accuracy is an indirect measure of the success that children would have if using the lexical semantic representations learned by the RNN for extending learned meanings during word learning.

9.3.4 Results

Figure 9.3 shows the results of RNN language modeling simulations for corpora differing in the value of α. The right panel shows that perfect categorization of X-word representations is achieved when A-words are either not redundant or only partially redundant (red and blue lines), but not fully redundant with X-words (black line). Perfect categorisation means that X-word representations are organized into clusters that perfectly corresponds to their target categories. The left panel illustrates categorization performance of A-word representations that were learned alongside X-word representations. At the end of training, in

[1] As a measure of similarity between two vectors, I used the cosine of the angle between them.

[2] To be fair, the RNN has no notion of category-relevance, and cannot infer category-relevance from the training data alone. The point of this experiment is therefore not to demonstrate a failure of this sort, but to better understand the dynamics of temporal credit assignment.

both the no-redundancy and partial-redundancy conditions, the balanced accuracy for A-words is at chance. Thus, one can say that the networks trained in these two conditions have acquired fully atomic lexical representations for X-words. However, in the full-redundancy condition (black line), in which A-words are fully redundant with X-words, the target category structure is only partially captured in the lexical representations of X-words. This is shown by the imperfect categorization of X-words in the right panel. This failure is accompanied by imperfect categorization performance by A-word representations, shown in the left panel. This pattern of results exemplifies low atomicity: Neither the representations of A-words nor of X-words alone has captured the target category structure perfectly; instead, the target category structure must be encoded in the *interaction* between the two representations.[3]

The findings reported above replicate those presented in Chapter 5 using a slightly different artificial language corpus. However, here I am most interested in a temporary lack of atomicity during early training in the partial-redundancy condition ($\alpha = 0.5$). What I am looking for in this experiment is evidence for semantic property inheritance during early training in the partial-redundancy condition, which is most representative of the structure of natural language sequences. Indeed, the left panel shows that during early training, A-words temporarily inherit semantic properties of X-words. Peak semantic property inheritance is reached around training step 450 (blue line). Interestingly, semantic property inheritance was quickly dissipated, as the RNN began to focus exclusively on the predictive nature of X-words. The key insight here is that the RNN is shopping around, so to speak, for predictive cues, and initially considers both A and X-words simultaneously. It takes some time until the RNN learned that X-words are sufficient for predicting the upcoming Y-word, and that it need no longer rely on the redundant semantic cue provided by A-words. While the RNN in this experiment was able to overcome the temporary 'shopping around' phase, an RNN trained on a corpus natural language may be less successful, and may suffer long-lasting negative consequences.

9.4 Is Atomicity Long-Lasting?

The results above indicate that an early intervention (for inducing atomicity) might be most effective, as this is the phase in which the RNN is still determining which items are most predictive of other upcoming items. However, this raises the following question: Once atomicity has been induced early during training, how long would atomicity last? Would the effect be temporary, or endure even after exposure to natural language sequences rife with incidental redundancy? I have argued that, an early atomicity induction should be long-lasting due to principles underlying error-based associative learning. In particular, a phenomenon called 'blocking', which has been observed time and again in the associative learning literature, predicts that once an association has been established, it is difficult to replace. Generally, blocking refers to a phenomenon where a previously learned association between a stimulus and outcome reduces the likelihood that another, equally predictive, stimulus can elicit prediction of the outcome in the absence of the originally paired stimulus (Waldmann & Holyoak, 1992). Importantly, it has been observed that novel predictors do not replace an established association, and often, do not trigger a behavior associated with prediction of the outcome variable in the absence of the more established predictor. This framework predicts that a learned association between a target word and an upcoming semantic cue would endure throughout the learner's lifetime, despite subsequent exposure to novel predictors.

[3] I verified this claim by computing the balanced accuracy for contextualized representations generated by inputting A X sequences to the RNN. Categorization using these contextualized representations was at ceiling.

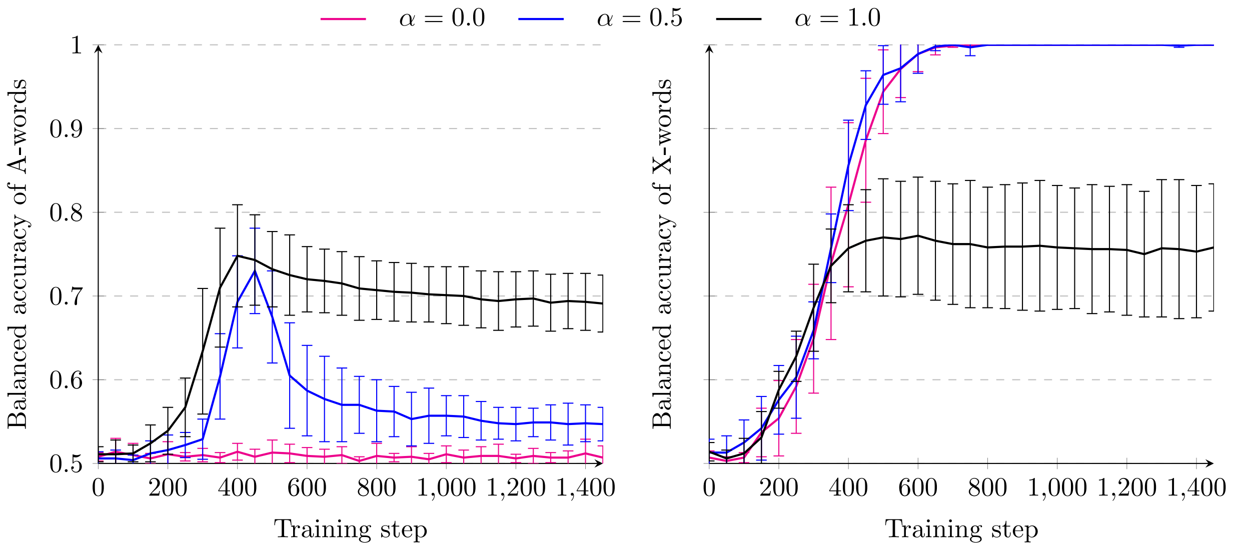


Figure 9.3: Semantic categorization accuracy (y-axis) of RNNs trained on corpora varying in the amount of redundant information about the target category structure provided by A-words. Each line represents the average performance across 10 RNN simulations, error bars indicate standard deviations, and the dotted vertical line marks the end of pre-training. The balanced accuracy measures the degree to which the lexical representations of A-words (left panel) or X-words (right panel) capture the externally-defined semantic category structure.

To test this prediction in a naturalistic setting, I trained the RNN on AO-CHILDES using one of two pre-training conditions: In the baseline pre-training condition, the network is simply trained on a random sample of 7-word sequences from AO-CHILDES for 37K training steps.[4] In the atomicity-bias pre-training condition, the RNNs were instead trained on an experimenter-crafted mini-corpus designed to induce perfect atomicity, also for 37K steps. In this condition, the networks were trained on two-item sequences of the form X Y, where X is the set of 700 probe words (Appendix A) and Y is a set of non-word items that perfectly cue the semantic category membership of the preceding probe words. Thus, the relationship between X and Y in this mini-corpus is conceptually identical to the relationship between X and Y in all artificial languages described previously — the distribution of Y-words is perfectly diagnostic of the semantic category membership of X-words (in this case, probe words). The reason I chose 37K steps is because this is the minimal number of steps needed to induce full atomicity (perfect balanced accuracy for lexical representations of probe words).

I conceptualized the pre-training phase in the atomicity-bias condition as an opportunity for the network to acquire a stable processing dynamic that minimizes semantic property inheritance by category-irrelevant left-contexts. Compared to a randomly initialized network that is heavily influenced by redundancy, the RNN pre-trained in the atomicity-bias condition comes equipped with an already ideally organized lexical semantic category knowledge at the input-to-hidden weights that it can immediately deploy. Because the relationship between probe words and upcoming semantic cues is already implicit in the organization of weights at the input-to-hidden weights, the network in the atomicity-bias condition need only adapt the parameters of the

[4] The number of steps is identical to the number of total mini-batches used during training, or equivalently, to the number of total weights updates computed.

processor (i.e. recurrent and output weights) to the structure already available at the input-to-hidden weights. Contrast this with the RNN in the baseline condition, which must simultaneously learn parameters associated with the processor and parameters associated with the lexical semantic representations at the input-to-hidden weights. As discussed previously, the more that these two learning problems can be temporally separated, the more atomic the resulting lexical representations should be.

Note, that after pre-training, all RNNs were trained on AO-CHILDES for the same number of steps, until completing approximately 3M total steps. No incremental training strategy was used.[5] The only incremental strategy is the separation of training into pre-training and standard-training. In the standard-training phase, networks in the baseline condition simply continued training on AO-CHILDES sequences. In the atomicity-bias condition, the learned lexical representations of probe words were transferred to otherwise randomly initialized RNNs, which were then trained on the same sequences of AO-CHILDES as models in the baseline condition. As before, 10 RNNs were trained in each condition.

The results are shown in the left panel of Figure 9.4. Note that the vertical dashed line indicates the boundary between pre-training and standard-training. Networks in the atomicity-bias condition (red line) were able to preserve their lexical semantic category knowledge over the course of millions of training steps, with only slow degradation that appears linear in the number of training steps. Furthermore, at the end of training, performance in the atomicity-bias condition is still well above that of networks trained on AO-CHILDES only. This finding demonstrates the long-lasting stability of atomic internal organisation in the face of complex lexical interactions present in naturalistic child-directed input. This finding lends credibility to the idea that an early intervention could have long-lasting beneficial effects for learning more atomic lexical semantic representations.

There is another reason why early atomicity induction is useful, form the point of view of lexical semantic development in children. A child who, for one reason or another, has acquired highly atomic lexical semantic representations very early during learning, would be able to reap the benefits for a long time after. For instance, such a child would be in a better position to perform the distributionally mediated extension of category-associated features (DECAF). This could potentially set up a virtuous cycle, where the child is able to learn novel word meanings more quickly using DECAF, and in turn, use his or her newly learned semantic features when extending learned meanings to more novel words.

9.5 Disentangling Information Trapped in the Processor

In the final experiment of this chapter, I have set out to test the validity of the assumption that temporally separating the learning of different kinds of statistics can actually improve the atomicity of learned lexical semantic representations. Specifically, the separation I have in mind is between word-level and chunk-level statistics, as discussed above and in Chapters 1 and 4. One way to test this assumption is to force the RNN to adapt one set of randomly initialized parameters to another previously learned set of parameters. To achieve this, it is possible to re-initialize only a portion of previously learned parameters at some point during training, while leaving intact others. In this way, the RNN is forced to re-learn the randomly re-initialized parameters from scratch — this time, constrained by knowledge that has been preserved elsewhere in the network.

[5] At each step, a 7-word sequence was randomly sampled from the corpus without replacement. Once sequences have been exhausted, a new epoch began, and the process repeated until 3M steps were completed.

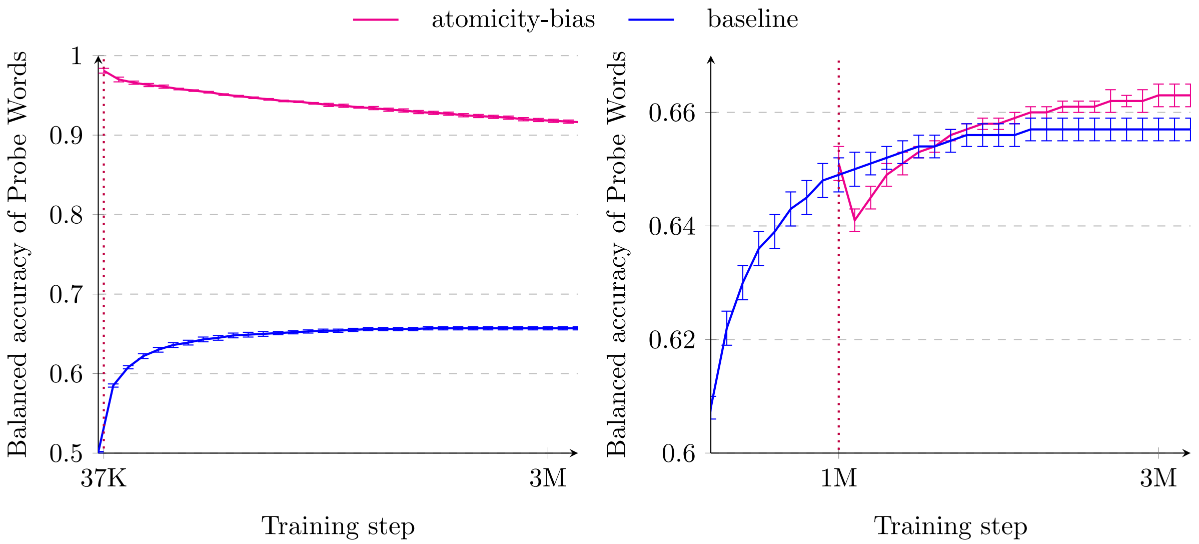


Figure 9.4: Semantic categorization accuracy, in units of balanced accuracy (y-axis), across training time (x-axis). Each line represents the average performance across 10 RNN simulations, and error bars indicate 95% confidence-intervals. **Left Panel:** Is Atomicity Long-Lasting? Two groups of 10 RNNs were trained on approximately 3M steps of child-directed language randomly sampled from AO-CHILDES, but with different initial knowledge and experience. In the atomicity-bias condition, the RNNs were first trained with 37K steps on simple artificial 2-item sequences (of the form X Y) to induce perfect knowledge of the target semantic category structure and with perfect atomicity. In the baseline condition, the RNNs were first trained with 37K steps on sequences of child-directed input - which neither provides perfect knowledge of the target task nor strong atomicity. **Right Panel:** Disentangling Information Trapped in the Processor. RNNs were first trained for 1M steps (13 epochs) on AO-CHILDES, and then separated into one of two groups (separation is indicated by ······). In the atomicity-bias condition (——), all the parameters of an RNN were randomly re-initialized except for lexical representations of probe words. In the baseline condition (——), training on AO-CHILDES continued uninterrupted for another 2M steps (27 epochs).

To pull this off, I trained 10 RNNs in each of two conditions. In each condition, the networks were trained non-incrementally on AO-CHILDES as before. While, in the baseline condition, each RNN simply keeps training in ordinary fashion past 1M steps, RNNs in the atomicity-bias condition receive an intervention at 1M steps. This intervention randomly re-initialized all learned parameters in the RNN except for the lexical semantic representations of probe words learned at the input-to-hidden weights. This strategy was designed to induce atomicity by removing any learned interactions between probe words and the contexts in which they occurred. That is, after re-initialization, a network cannot use its previous knowledge about the contexts in which probes occur to predict upcoming semantic cues — its only knowledge is that encoded in the learned lexical representations of probe words. Ultimately, the reason that re-initialization is supposed to promote atomicity is because it separates learning into two phases: In the first phase, lexical semantic representations are shaped; then, after 1M steps, any chunk-level statistics relevant to semantic categorization that might have become trapped at the processor (i.e. hidden layer) is removed at re-initialization. In the second stage, the RNN must re-learn weights for processing that work well with the semantic category knowledge already established at the input-to-hidden weights.

The results are shown in the right panel of Figure 9.4. The semantic categorisation accuracy of RNNs in the atomicity-bias condition (red line) initially drops below the baseline (blue line), then increases rapidly, and finally surpasses the baseline (average balanced accuracy is 65.71 ± 0.17 [6] in the atomicity-bias condition and 66.29 ± 0.17 in the baseline condition). The initial dip occurs because the randomly re-initialized networks must re-learn all parameters not responsible for representing probe words, including all other lexical representations. Counter to intuition, this removal of knowledge turns out to have long-term positive consequences for atomicity, as predicted by SPIN theory, and in particular, the staged learning framework described in this chapter. The success of this — rather dramatic — intervention is in accordance with machine learning research on techniques for combating maladaptive interactions such as sparsity-regularization (e.g. L_1-norm regularization), dropout, weight-decay, and other techniques.

9.6 Summary

In this chapter, I used SPIN theory to derive an intervention for promoting atomicity in the RNN language model. The proposal consists of a staged training strategy that combines incremental training with ordered presentation of data separated into two distinct stages, modeled after the longitudinal organization of surface-level lexical statistics of child-directed input across developmental time (Chapters 7 and 8). Each stage encourages the formation of distinct processing dynamics. In stage 1, the category-relevant association between target words and upcoming semantic cues is prioritized by training on counterbalanced left-contexts; in stage 2, left-contexts are emphasized, and no longer require counterbalancing. The prediction is that the atomicity induction that should occur during stage 1 helps the RNN to concentrate category-relevant lexical statistical in the lexical representations of target words at the input-to-hidden weights, and without being corrupted by category-irrelevant chunk-level statistics.

Next, I provided empirical support for three portions of this argument. First, I validated the existence of 'semantic property inheritance', the capturing of semantic category-relevant distributional information by the left-contexts of target words, and showed that it peaks during early training. Second, I showed that once atomic lexical semantic representations have been induced, they maintained their atomicity across millions of training steps despite being faced with incidental redundancy due to idiosyncratic natural language input. This is remarkable, given the vulnerability of incrementally trained neural networks to catastrophic forgetting (McCloskey & Cohen, 1989), the forgetting of knowledge learned long ago (e.g. during pre-training). This means that once an atomic internal organization is acquired, it is relatively stable. Third, I showed that the idea of temporally separating learning into two stages is viable, as doing so can improve lexical atomicity, as evidenced by improved performance in a downstream semantic categorization task. In particular, I showed that by disentangling lexical semantic information at the processor (i.e. re-initializing all but the weights learned for lexical semantic representations of probe words), semantic categorization accuracy is greater at the end of training compared to networks that did not undergo random re-initialization. Additional experiments and discussion can be found in the published version of this work (P. A. Huebner & Willits, 2021a).

So far, the picture of learning nominal semantic categories in the RNN is as follows: Knowledge of a noun's semantic category can be understood as an operation that transitions from a state of 'noun-ness' to a state that differentiates among sets of nouns. This operation is implicit in the noun's lexical semantic representation. SPIN theory states that such a principled transition operator is not learned unless the RNN encounters, during training, many nouns in an undifferentiated state of 'noun-ness' (in semantically uninformative left-context). This implies that learning semantic category membership of nouns is strictly

[6] mean ± margin of error, with $\alpha = 0.05$

dependent on the presence of a stable superordinate category — the noun category. In other words, the early discovery of part-of-speech (POS) classes might 'scaffold' learning of semantic distinctions that exist within a POS class. The implications of this idea are significant for language acquisition: If training on input to younger children first induces a more categorical encoding of nouns, the RNN should be more robust against correlated cues during later stages in training, and as a consequence acquire more atomic lexical semantic representations.

Chapter 10

The Age-Order Effect

In this chapter, I explicitly test a prediction derived from SPIN theory and developed in the last few chapters, namely that training an RNN on age-ordered child-directed input will yield more atomic lexical semantic representations than an identical model trained on the same data but in reverse order. Importantly, in this chapter, I examine the non-contextualized lexical semantic representations that are learned at the input-to-hidden weights in contrast to the contextualized representations that are dynamically generated at the hidden layer (Chapter 3).

10.1 Methods

10.1.1 Training Procedure

All simulations reported below use an incremental training regime. Briefly, AO-CHILDES was split into 8 equally sized partitions, 1-8. The RNN iterates over partitions rather than over the full corpus with each new epoch. Doing so preserves the age-ordering of transcripts during training, and allows the RNN to take advantage of reduced fragmentation of the noun category during early training. All RNNs were trained 12 times on each partition; once training on one partition completed, the RNN did not see that data again during training. This procedure results in 900,000 total training steps and is identical to iterating over the full corpus 12 times. To examine how the size of local iterations influences learning, I varied the number of AO-CHILDES partitions. While I report only the results obtained for RNNs trained on 8 partitions, the qualitative results are robust against the number of partitions evaluated (8, 128, 256).

A single training step consists of a single forward and backward pass on a batch of 64 sequences. Batches within each partition are shuffled once at the start of training, in order to randomize the order of batches across different RNN simulations while preserving the age-ordering of partitions. Hyper-parameters are identical to those reported in Chapter 4.

In order to test the generalizability of the results in this chapter to other RNN variants, I repeated each experiment using an LSTM in place of the simple RNN. In theory, the simple RNN is capable of maintaining information about items located an arbitrary number of steps into the past, but due to the instability of the gradient across time steps, this is extremely improbable in practice. To overcome the difficulty of learning longer-distance dependencies, numerous extensions of the RNN have been proposed. For example, the Long Short-Term Memory (LSTM; Hochreiter and Schmidhuber, 1997) uses gating units to control the flow of information into and out of the hidden state. Because the hidden state in the LSTM is never squashed by a non-linearity, this allows for more efficient gradient propagation across time steps.

10.1.2 Probe Words

To examine the atomicity of lexical semantic representations, I chose the same target words used in Chapter 3. As a reminder, the target semantic category structure consists of 700 common nouns frequent in AO-CHILDES, each of which is assigned exactly one of 27 semantic categories such as MAMMAL and VEHICLE. For a complete description of the semantic category structure, see Appendix A. To distinguish these from other target words, including less frequent nouns, I will refer to them as probe words. The total number of probe word occurrences in AO-CHILDES is approximately 200,000.

10.2 Results

10.2.1 The Right-to-Left Effect

Before examining the effect of age-ordered training, I first tested the assumption that left-contexts are semantically informative in AO-CHILDES. This is critical to the hypothesis I have been developing, in which left-contexts are thought to provide redundant information already provided by the probe word about upcoming semantic cues. If this turns out not to be true, there would be no grounds for claiming that left-contexts provide redundant information about semantic category membership, and an intervention based on counterbalancing left-contexts would be ineffective. While the corpus analyses in Chapter 7 and 8 already provide preliminary evidence for this, the assumption is straightforward to verify by training the RNN to predict consecutive tokens within AO-CHILDES partitions right-to-left. The left-to-right condition is the standard condition, in which tokens are presented in the canonical order in which English words are produced; however, in the right-to-left condition, the ordering of tokens in the input to the RNN is reversed. If semantic categorization performance is above chance when trained right-to-left, we can conclude that left-contexts provide semantic information about the target semantic category structure of probes. This follows from SPIN theory, which states that the RNN can only encode semantic category membership in a given target word if the cue to membership occurs in the right-context of the target word. Put differently, categorization performance is sensitive only to information that follows a target word in the order in which the input is processed. By training the RNN right-to-left, it is possible to exploit left-contexts for learning about semantic category membership; performance should be above chance only if they are useful for diagnosing semantic category membership. I trained 10 simple RNNs and 10 LSTMs in each condition (left-to-right vs. right-to-left).

The results are shown in Figure 10.1. In both conditions, the RNN achieves performance that is well above chance (chance-level is 0.5). Strikingly, learned lexical representations contain richer semantic information when trained right-to-left. For brevity, I will refer to this effect as the 'right-to-left' effect. Together, these results clearly demonstrate that both left and right contexts of probe words in AO-CHILDES are informative about the target semantic category structure; this means that each is likely to provide redundant information already provided by the other.

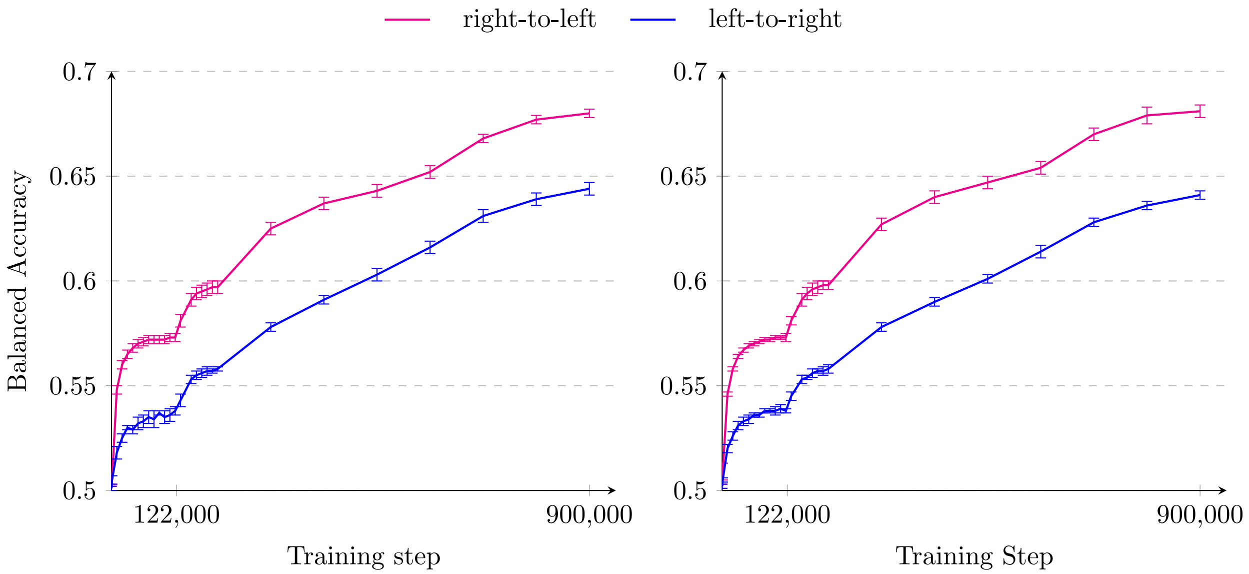


Figure 10.1: Semantic categorization accuracy (y-axis) at consecutive intervals during incremental training (x-axis) on AO-CHILDES, for the simple RNN (left panel) and LSTM (right panel). Categorization accuracy is operationalized as the balanced accuracy of correctly judging whether each possible pairing of probe word *lexical* representations are same-category members or not. Each line represents an average across 10 RNN simulations, error bars indicate 95% confidence-intervals. RNNs were either trained on age-ordered input by processing input in the left-to-right (——) or right-to-left (——) direction.

10.2.2 The Age-Order Effect

Finally, we are in a position to test the prediction that atomicity of lexical representations of probe words will be greater when the RNN is trained incrementally on age-ordered partitions of AO-CHILDES compared to in reverse order. Given (i) the results of my corpus analyses which demonstrate that input to younger children better approximates the counterbalancing requirement, (ii) the right-to-left effect which shows that left-contexts provide redundant information about the target semantic category structure of probes, and (iii) my theoretical argument that redundancy impedes the formation of atomic lexical semantic representations in the RNN, age-ordered training should yield more atomic lexical semantic representations, and this difference should manifests as an improvement in downstream semantic categorization using lexical representations obtained at the input-to-hidden weights.

The results are shown in Figure 10.2. Each curve plots the average categorization performance of 10 RNNs across training. The left panel contains results for the simple RNN, and the right panel contains results for the LSTM. In agreement with my predictions, training in age-order results in greater categorization performance at the end of training, despite having been exposed to exactly the same amount of data. This is true regardless of which architecture is used. For brevity, I will refer to this effect as the 'age-order' effect. Zooming in, I found that this improvement in performance is due to a spike in performance while training on the very first partition of AO-CHILDES, which contains input to the youngest children in the American-English section of the CHILDES database. The end of partition 1 is marked by a dotted vertical line at step 122,000 steps. This pattern of results is in agreement with the idea that once atomicity is established early during training, it may be preserved for a long time after. See the discussion of orthogonal sub-spaces in Chapter 9 for details.

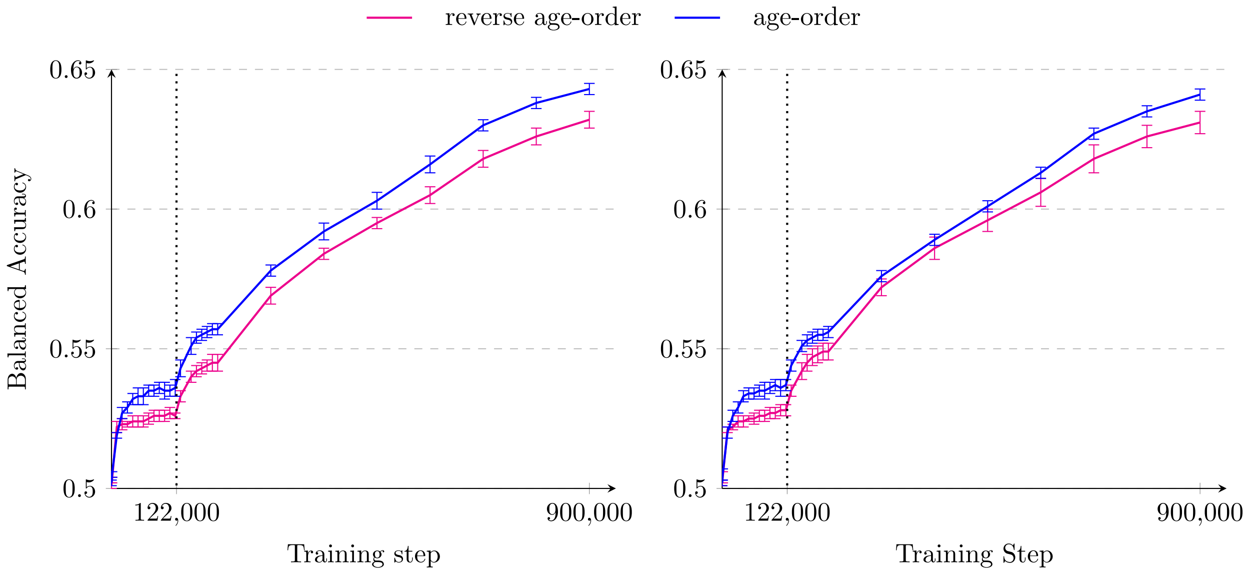


Figure 10.2: Semantic categorization accuracy (y-axis) of *lexical* representations at consecutive intervals during incremental training (x-axis) on AO-CHILDES, for the simple RNN (left panel) and LSTM (right panel). Categorization accuracy is operationalized as the balanced accuracy of correctly judging whether each possible pairing of probe word representations are same-category members or not. Each line represents an average across 10 RNN simulations, error bars indicate 95% confidence-intervals, and the dotted vertical line marks the end of the first AO-CHILDES partition. RNNs were either trained in age-order (—) or in reverse age-order (—).

10.2.3 Lexical vs. Contextualized Representations

It should be noted that, for incrementally trained RNNs, the total amount of semantic category knowledge that is encoded in the contextualized representations is greater than in the lexical representations, just like the results reported in Chapter 6 for non-incrementally trained networks. At the end of training, the difference in performance between contextualized and lexical representations for the simple RNN is slightly more than two accuracy points (average balanced accuracy = 0.664 vs. 0.643). This difference is even more pronounced in the LSTM (average balanced accuracy = 0.692 vs. 0.641). This follows from ideas developed in Chapters 5 and 6, namely that when a semantic cue is redundantly predicted by a left-context, the RNN tends to treat the left-context and the target word as a compound cue with the consequence that lexical semantic information related to the target word becomes ‘trapped’ at the hidden layer. Not only does this finding provide external validation for SPIN theory, but it shows just how much of a performance gap there is between contextualized knowledge that is dynamically generated by the processor and static knowledge readily available at the input-to-hidden layer. Surprisingly, only a small fraction of semantic category knowledge is trapped in the processor of the simple RNN; considerably more is trapped in the processor of the LSTM. This difference is likely related to the presence of gating units in the LSTM that control the flow of information within and between time steps.

10.2.4 The Benefit of Training in Age-order

In Chapter 6, I proposed that training in age-order might (partially) restore the performance gap in contextualized vs. non-contextualized representations due to the early promotion of lexical atomicity. We can now test this idea: Does training in age-order produce lexical semantic representations that perform equally well in semantic categorization as contextualized representations learned by an otherwise identical RNN trained on randomly ordered partitions? Not quite. The average gaps in balanced accuracy are shown in Table 10.1. When training on AO-CHILDES partitions in random order, the balanced accuracy for contextualized representation is still higher than the balanced accuracy of lexical representations when training in age-order (average difference = 0.018). Notably, however, the original gap (average difference = 0.029) has been cut in half by training in age-order. The same gap proved much larger to bridge in the LSTM, due to its ability to capture much more semantic category information in its contextualized representations than the simple RNN.

	Knowledge Gap	
	simple RNN	LSTM
(shuffled, contextualized) - (shuffled, lexical)	0.661 - 0.632 = 0.029	0.683 - 0.631 = 0.052
(shuffled, contextualized) - (age-ordered, lexical)	0.661 - 0.643 = 0.018	0.683 - 0.641 = 0.042

Table 10.1: Gap in balanced accuracy due to entanglement of semantic category knowledge at the hidden layer. Training on age-ordered partitions improves balanced accuracy relative to training on shuffled partitions, but does not fully restore performance of lexical representations to level previously observed for contextualized representations (trained on shuffled partitions). Values shown are averages across 10 simulations per condition.

10.2.5 Incremental vs. Standard Training

While the results clearly demonstrate that age-ordered training is superior — with respect to lexical atomicity — to training on randomly ordered partitions of AO-CHILDES, models trained non-incrementally nonetheless under-perform all incrementally trained models. Consider, for instance, the average categorization accuracy of non-incrementally trained RNNs discussed in Chapter 3. As mentioned there, the average balanced accuracy of contextualized representations was 0.679 ± 0.005 (mean ± margin of error), and the average balanced accuracy of lexical representations was 0.651 ± 0.006 (mean ± margin of error), with n=10 and α=0.05. Each is at least one point larger than the average performance of models trained incrementally on age-ordered partitions.

What is the reason for this performance gap? The answer is straightforward: At the start of each new epoch, a non-incrementally trained network iterates over the entire corpus, rather than one partition at a time. As a consequence, models trained in standard fashion (i.e. incrementally) are able to re-visit all of the data at the start of each new epoch. This means that such a model is continuously given the opportunity to integrate each new example with previously acquired knowledge — reducing catastrophic forgetting (McCloskey & Cohen, 1989). To illustrate this in more detail, consider, for example, the longest separation (in units of training steps) between two observation at the point at which an incrementally and non-incrementally trained model have completed training. Further, assume, that the incrementally trained model was trained on 8 partitions, and that each model iterated 12 times either over a partition (in the incremental condition) or the entire corpus (in the standard condition). In both cases, each model has been exposed to the same number of total training steps; however, it has been 825,000 steps since the former model has last observed examples from partition 1, and no more than 75,000 steps for the latter. The incrementally trained model, therefore, has gone longer without revisiting observations made at the start of training, and this increases the likelihood that this information is forgotten and/or over-written.

A similar kind of observation has been made by Fremann and Lapata (2016) who found that an incrementally trained Bayesian model of category acquisition slightly under-performs an otherwise identical model trained using Gibbs sampling, which enables full access to the data at any point in training. However, I agree with the authors that while the incremental model is slightly less accurate, it is more cognitively plausible.

10.3 Differences in the Availability of Information about Semantic Category Membership

Although the age-order effect falls out of SPIN theory, there is an alternative explanation that must be ruled out. It is possible that the amount of information about the target semantic category structure changes across developmental time, with more information available in input to younger compared to older children (partition 1 vs. 8, respectively). This simple account would explain the initial improvement in performance for models trained in age-order compared to models trained in reverse. In particular, it suggests that models trained in reverse, which are initially iterating over partition 8, are not able to accumulate as much information about the semantic category structure. Given that this alternative interpretation competes with the idea that counterbalanced left-contexts should promote lexical atomicity, it is important to rule it out. I conducted three analyses, discussed below.

10.3.1 Semantic Category Knowledge in the Processor

In the first follow-up analysis, I extracted contextualized representations from the same RNNs reported above, as detailed in Chapter 3, and evaluated them using the same semantic categorization task.

The results are shown in in Figure 10.3. First, note that categorization of contextualized representations is above-chance at the beginning of training. This is due to the inherent tendency of randomly initialized RNNs to cluster similar input sequences. This fact is not relevant to interpretation of the results. In agreement with the idea that the age-order effect is not caused by access to richer semantic information in input to youngest children, I found no initial performance advantage prior to step 122,000 for RNNs trained in age-order. As before, the qualitative pattern of results holds true for both RNN architectures (simple RNN shown in left panel, LSTM shown in right panel). This pattern of results indicates that the initial advantage in semantic categorization of non-contextualized lexical representations during age-ordered training can not be due to the presence of more category-relevant semantic information in partition 1 compared to partition 8. Models trained in either condition are initially equally good at categorizing contextualized representations, meaning there is an equal amount of semantic information in partition 1 and 8. This suggests that the age-order effect is not due to how much information is available to the model, but *how the model makes use of that information*. An atomic encoding strategy would promote the concentration of category-relevant information at the lexical representation of probe words, rather than in their left contexts. The age-order effect shows that that is precisely where information is encoded by models trained in age-order.

Figure 10.3 points to another surprising result, not related to my predictions. Even when contextualized representations are used — when the RNN is able to harness all the available information in left contexts of probe words — there is still an improvement at the end of training for models trained in age-order. Unlike the age-order effect, this improvement in performance emerges only gradually. It appears that the early induction of atomicity when training on input to younger children first is not only beneficial for the purpose of learning semantically richer lexical representations, but also for learning semantically richer contextualized representations. It is possible that the atomic organization established early during training scaffolds learning

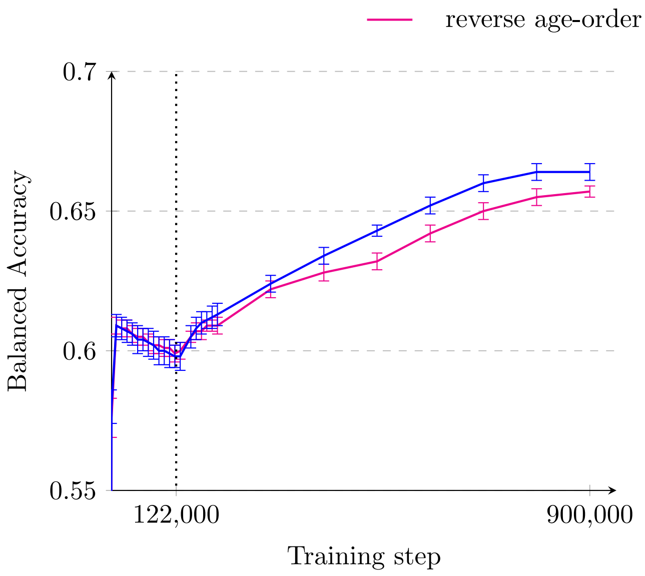


Figure 10.3: Semantic Categorization accuracy (y-axis) of *contextualized* representations at consecutive intervals during incremental training (x-axis) on AO-CHILDES, for the simple RNN (left panel) and LSTM (right panel). Categorization accuracy is operationalized as the balanced accuracy of correctly judging whether each possible pairing of probe word representations are same-category members or not. Each line represents an average across 10 RNN simulations, error bars indicate 95% confidence-intervals, and the dotted vertical line marks the end of the first AO-CHILDES partition. RNNs were either trained in age-order (——) or in reverse age-order (——).

to allow for more efficient capture of category-relevant semantic information. However, this is speculative, and requires additional research that is outside the scope of this work.

To further bolster the claim that the age-order effect is not due to age-related differences in the availability of information in the corpus, I examined whether there are any systematic differences in how well probe word contexts in input to younger vs. older children predict the target semantic category structure. In the two analyses below, AO-CHILDES was split into 2 partition, rather than 8 used to train the RNN.

10.3.2 Balanced Accuracy of Bag-of-Words Model

One way to compute how much distributional information about semantic category structure exists in a given corpus partition is to use the same measure used to compute semantic categorization performance in the RNN, the balanced accuracy. Instead of computing the balanced accuracy for probe word representations learned by the RNN, I computed the balanced accuracy for probe word representations learned by a bag-of-words (BOW) model that was fit on either partition 1 or partition 2 of AO-CHILDES. The BOW model represents a word as a vector where each element is the frequency with which a vocabulary word occurs with the word in question. The co-occurrence context is restricted to a sequence of words occurring left to a target word. The sequential information, however, is discarded by updating the frequency counter of a context word regardless of its position in the sequence.

The results are shown in the left panel of Figure 10.4. Semantic categorization accuracy for the BOW model fit on partition 2 reaches a higher peak, and drops off slower with context-size, compared to the model fit on partition 1. One way to interpret this finding is that the information about semantic category membership is preserved across greater distances in partition 2 compared to partition 1. This is likely related to the presence of longer utterances in partition 2, where semantic dependencies may span longer distances.

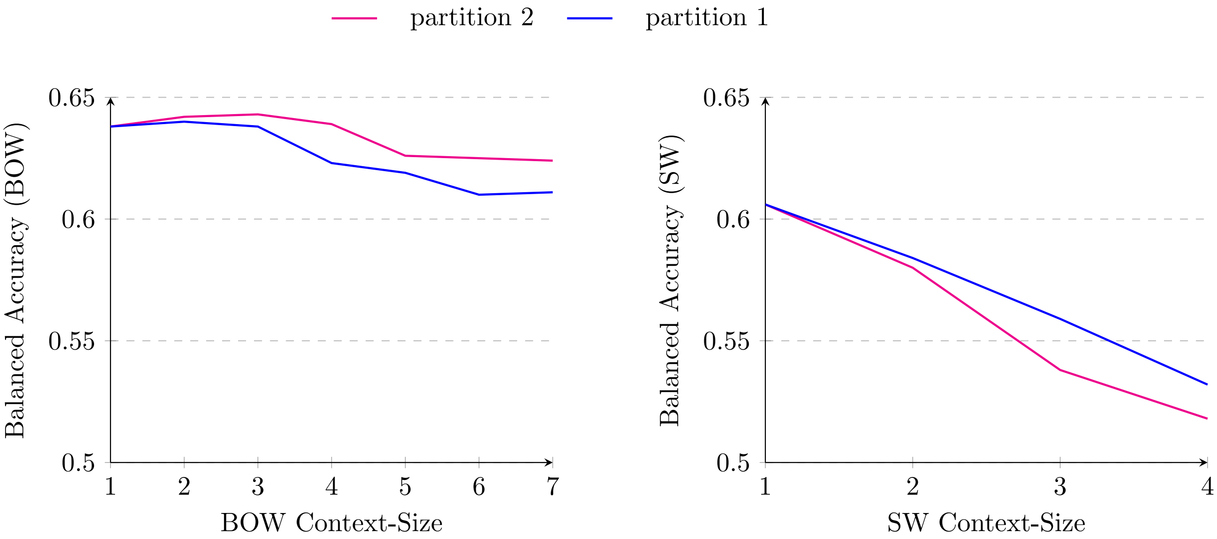


Figure 10.4: Semantic categorization accuracy (balanced accuracy) for BOW model (left panel) and SW model (right panel) with context-sizes varying from 1-7, and fit on partition 1 (——) and 2 (——) of AO-CHILDES. The higher accuracies for the BOW model fit on partition 2 indicates that category-relevant semantic information is preserved across longer distance in input to older children. BOW means bag-of-words, and SW means sliding-window.

10.3.3 Balanced Accuracy of a Sliding-Window Model

Representing probe words in a way that preserves word-order information can be achieved by using a sliding-window (SW) model. In contrast to the bag-of-words model, the sliding-window model slides a windows (1 through 4 used here) across the words in a corpus, and updates a vector of co-occurrence frequencies where each element represents both the identity of the word and it position in the sliding window. When the context size is 1, the bag-of-words model is a special case of the sliding-window model. When the context size is larger than 1, the size of each vector increases by a factor of the context size. Given a vocabulary of 8,000 words, a sliding-window model representation trained with a context size of 1 is 8,000 elements long; when the context size is 2, the representation is 16,000 elements long. As before, I trained one model on partition 1 and another on partition 2, and tracked the balanced accuracy as a function of the amount of input each has seen.

The results are shown in the right panel of Figure 10.4. The balanced accuracy tends to be larger for the sliding-window model trained on partition 1 when the context size is 3 and 4, but this trend does not hold for context sizes 1 and 2. This indicates that there is some, but not much, more information about the semantic category structure in partition 1 when word-order is preserved.

An interesting trend in this set of results is that the balanced accuracy drops considerably faster for the sliding-window models compared to the bag-of-words models. In other words, adding word-order information to the probe word representations weakens semantic categorization performance. This means that the RNN's sensitivity to the sequential structure of its input might negatively influence the construction of form-based lexical semantic representations. Surprisingly, these observations suggest that the best performing RNN should be the one that can learn to ignore the sequential structure, and focus on semantic dependencies somewhat independently of the distances they span. Clearly, this cannot happen in an RNN explicitly trained to predict sequences. This kind of observation suggests that the RNN may not be ideally suited to learning lexical semantic clusters. An alternative model that is less constrained by sequential structure is Word2Vec. Indeed, when trained on the same input as the RNN, Word2Vec does perform better on semantic categorization (P. A. Huebner & Willits, 2018), but not by much. That said, it is remarkable that the RNN can capture semantic category membership as well as it does despite being strongly influenced by word-order.

In sum, the follow-up studies revealed that the question about whether input to younger or older children contains more distributional information about lexical semantic category membership depends on whether or not word-order is considered in the analysis. How a system deals with sequential statistics, is therefore, a crucial determinant of its success in discovering distributional patterns indicative of lexical semantic category membership. On the one hand, it could be argued that, on the basis of the sliding-window experiment, there is more distributional information about the target semantic category structure in partition 1 compared to partition 2 of AO-CHILDES. On the other hand, when word-order information was removed (bag-of-words model), the special status of partition 1 disappeared. This is consistent with the idea that greater combinatorial diversity reduces the discoverability of distributional cues of semantic category membership. Syntactically more complex language may obscure distributional patterns that depend on consistent ordering of words. When context words are ordered in more variable patterns, it should become more difficult for statistical learning systems to detect those patterns and use them to cluster probe words by semantic category.

10.4 Learning Dynamics

In this last set of follow-up analyses, I consider potential differences in the learning dynamics of RNNs trained on different orderings of AO-CHILDES. The goal is to test whether the mechanistic account I have put forward — that training on data with less fragmentation of the noun category first promotes long-term lexical atomicity — is a viable explanation of the age-order effect. The key idea that underlies my account, and which I will leverage in subsequent analyses, is that pre-nominal contexts and probes 'stick together' to a greater degree in language input to older compared to younger children. This 'stickiness' (akin to fragmentation) in the data should also manifest in the dynamics learned by the RNN: A network trained in reverse order is first trained on partition 8, where pre-nominals and nouns are 'stickiest', according to my corpus analyses. It follows that any category-relevant semantic information about a probe word may also 'stick' to the representation of the left-context in which a probe occurs. This would explain why more information diagnostic of semantic category membership is encoded in the lexical representation of probe words when training in age-order: Semantic information is less likely to leak backwards across time steps where it can 'stick' to left-contexts. The challenge is to be able to identify this 'stickiness' and to devise a quantitatively measure to enable model comparisons.

10.4.1 Fragmentation and Effective Dimensionality

One method for quantifying the 'stickiness' between probes and left-contexts in the representational space of RNNs, is to adopt the measure of fragmentation defined in Chapter 7 and apply it to the predictions made by the RNN at the output layer. While I have talked about fragmentation as a property of co-occurrence matrices, it can be applied to any other kind of data that can be formatted as a matrix. The reason that fragmentation, as a quantitative evaluation, is useful for identifying 'stickiness' is because left-contexts that stick to particular probes are, by definition, only predictive of the probes that they precede, and are therefore not diagnostic of the noun category as a whole. In contrast, left-contexts that do not stick to probes are more likely to be shared by other nouns, and therefore promote an encoding of a single coherent noun category. It follows that the greater the number of sticky left-context + probe word combinations, the less coherent the encoding of the noun category will be, and the greater the number of sub-noun clusters in the representational space of the RNN. Fragmentation is ideally suited to measure this, as it is based on singular value decomposition (SVD) which decomposes a vector space into independent factors/dimensions in the order of the variance accounted for by each dimension. Given that the matrix representing a vector space has not been centered, the first singular dimension represents the factor that is most common to all items in the vector space. If, for instance, each vector represents representations of probes/nouns learned by an RNN, the first singular dimension will encode the variance due to each probe word belonging to the same category, the NOUN category. Put simply, in this example, the first singular dimension may be understood as a 'prototype noun representation". The variance accounted for by the prototype is the first singular value. All subsequent singular dimensions are associated with monotonically decreasing singular values, as the dimensions that each encodes becomes increasingly less important for describing the overall organization of the data, the vector space of all probe words. Subsequent dimensions may be understood as sub-categories of the prototype noun category, or sub-categories of sub-categories, and so on. Typically, the last few dimensions encode very little variation, and are often considered to reflect noise in the data. To quantify the extent to which the noun prototype predominates in organizing the representational space of probe words relative to smaller sub-noun clusters, it is possible to compute the ratio of the first singular value (representing the noun prototype) and all remaining singular values (representing properties not prototypical of the noun category). Fragmentation is simply the opposite concept: How much variation in the data is not accounted for by the first singular dimension? Borrowing from Chapter 7, the formula to compute fragmentation, given the vector of singular values, s, obtained via SVD, is

$$Frag = \frac{(\sum_i s_i) - s_1}{\sum_i s_i} = 1 - \frac{s_1}{\sum_i s_i} \tag{10.1}$$

As input to the computation of fragmentation, I collected contextualized representations of all 700 probe words at the output layer using the following procedure: For each probe word, I extracted all sequences of size 7 in AO-CHILDES that end with the probe word. Next, all sequences ending in a given probe were input to the RNN, and the resulting output layer states were averaged. I extracted representations at the output layer because the output layer most closely reflects the behavior of the network as a whole — it is the output layer that interfaces most closely with the error signal that guides learning. Contextualized representations not only include information about a probe word, but also all the information contributed by 7 items that precede each probe word in the corpus. I stacked each contextualized representation to form a 700 by 8,000 matrix where rows are labeled by probe words (700) and columns are labeled by all tokens in the vocabulary of the model (8,000). I did not normalize or center the resulting matrix before applying SVD. Next, fragmentation was computed. I repeated this procedure at consecutive intervals during training to understand how the network's learning dynamics are shaped across training.

What am I looking for in the results? First, I do not simply want to know whether fragmentation is overall lower or higher depending on the order in which a network is trained. The reason is that a large value does not tell us whether fragmentation is due to formation of target semantic categories or other groupings irrelevant to the target category structure. It is therefore not clear whether fragmentation is indicative of adaptive or maladaptive differentiation. However, given that I am most interested in the period of training that encompasses iterations over the first partition only (prior to step 122,000), when there is considerable opportunity to over-fit on idiosyncratic left-context + noun co-occurrences, I consider high fragmentation during this stage to reflect a pre-mature splintering of nouns into clusters that are largely irrelevant to the target semantic category structure. In other words, I consider high fragmentation of learned representations of the noun category *during early training* as evidence that a network has differentiated probes that may belong to the same or similar target semantic category.

An alternative way to understand what fragmentation of output layer representations means is to consider how error minimization works in RNN language models. A model that has become sensitive to subtle statistical relationships between nouns and their left-contexts is able to minimize next-word prediction error more quickly, due the ability of many left-contexts of probe words to make accurate across-target predictions (predictions about right-contexts based purely on information in left-contexts). Such a network, in turn, will produce more distinct next-word probability distributions, each catered to a specific situation. On the other hand, a network trained on input where left-contexts provide little across-target information, will produce very similar next-word probability distributions, reflecting the fact that it does not yet know how to differentiate different occurrences of nouns given their left-contexts. Such a network will produce predictions that respect that all probe words are nouns, meaning its predictions are relatively insensitive, at first, to distinctions *within* the noun category. Fragmentation picks up on the difference between these two networks, as it is, roughly-speaking, a measure of the average multi-variate dissimilarity between row-vectors (next-word probability distributions) and the 'prototype' vector that best summarizes all rows (prototype next-word prediction). The former network that produces more dissimilar/differentiated next-word probability distributions at its output layer is assigned higher fragmentation compared to the latter network which tends to produce very similar next-word probability distributions. The theory I have proposed in Chapter 6 and extended in Chapter 9 predicts higher fragmentation for networks trained in reverse age-order on AO-CHILDES while training on partition 8 (prior to step 122k). The results are shown in the top panels of Figure 10.5. The top left and top right panels track the fragmentation of contextualized representations of probe words at the output layer of the simple RNN, and LSTM, respectively. In agreement with theory, fragmentation is initially much higher for models trained in reverse age-order (red), and this is true regardless of the architecture. Interestingly, fragmentation spikes during training on the first partition (partition 8 for models trained in reverse age-order), and then returns to the same level exhibited by models trained in age-order after exiting the first partition. To illustrate this more clearly, the grey doted line marks the point at which training transitions from the first to the next partition of AO-CHILDES.

One downside of using the formula for fragmentation defined above is that it only considers the relative size of the first singular value, while disregarding the relative differences in the sizes of subsequent singular values. Larger singular values are associated with singular dimensions that account for larger amount of variance in the data, and are therefore more important than smaller singular values which often account for incidental variance that is not relevant for describing the underlying (factor) structure of a given dataset. In other words, fragmentation does not take into consideration how steeply singular values associated with progressively less important singular dimensions decrease. Data with a small number of dimensions each accounting for a large amount of variance would provide a similar numeric value for fragmentation compared to data with a large number of dimensions each accounting for a small amount of variance. Fortunately, there exists a formula for computing a similar quantity that is sensitive to this difference. The quantity is called 'effective dimensionality' (ED), and has been previously used to analyze the dynamics of RNNs by Farrell

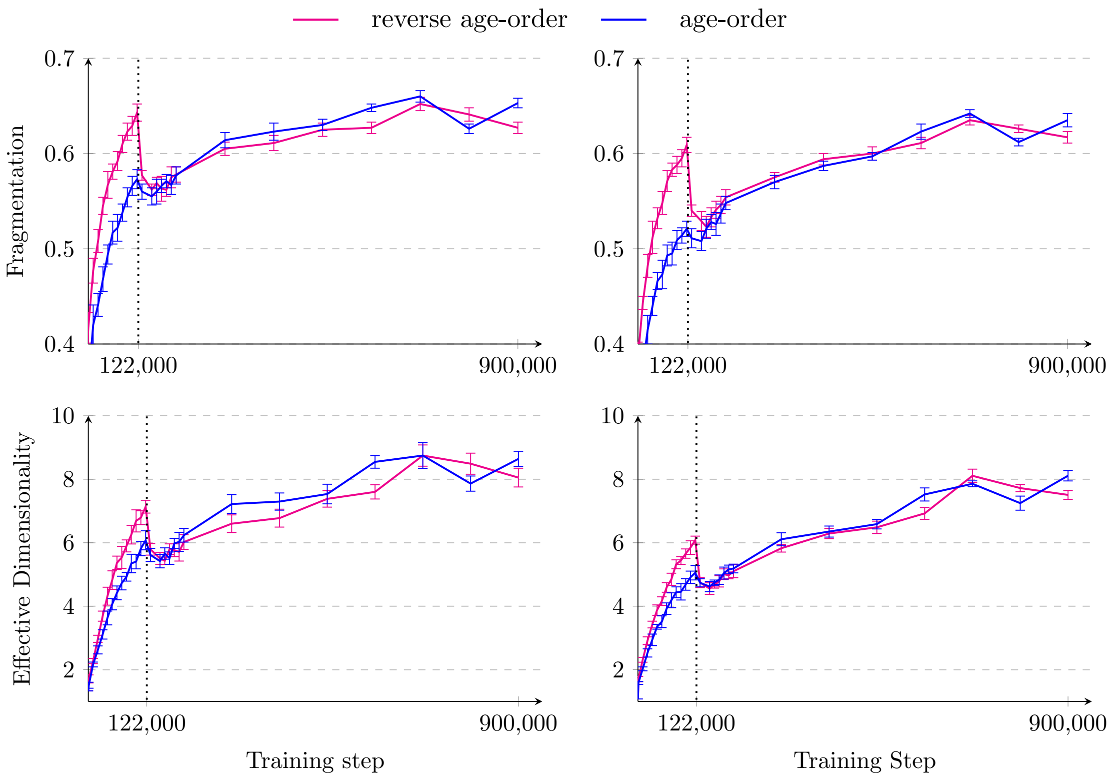


Figure 10.5: Fragmentation (top panels) and effective dimensionality (bottom panels) of RNN next-word predictions (y-axis) at consecutive intervals during incremental training (x-axis) on AO-CHILDES, for the simple RNN (left panel) and LSTM (right panel). Next-word predictions are computed for each 7-word sequence in the corpus that ends with a probe word. Each line represents an average across 10 RNN simulations, error bars indicate 95% confidence-intervals, and the dotted vertical line marks the end of the first AO-CHILDES partition. RNNs were either trained in age-order (—) or in reverse age-order (—).

et al. (2019). The formula to compute ED is given by

$$ED = \frac{(\sum_i s_i)^2}{\sum_i {s_i}^2} \tag{10.2}$$

Another desirable property of ED is that it can be used to estimate the number of functional (i.e. effective) dimensions used by a high-dimensional system to solve a particular task. While the total number of dimensions available to the RNN at the output layer is the size of the vocabulary (8,000), it is more likely that the RNN uses far fewer dimensions to perform well on the next-word prediction task. Larger ED, like fragmentation, is associated with a more complicated model, which use more dimensions to perform well on a given task. Higher ED is also associated with the greater potential of a system to over-fit on its training data, to be susceptible to chaotic dynamics, and to discover higher-dimensional solutions to complex problems (Farrell et al., 2019). Here, I use ED to measure the intuition that a network with a splintered encoding of the noun category uses more dimensions to be able to capture — both category-relevant and category-irrelevant — variation within the noun category. Similar to my predictions for fragmentation, I predicted that a model first trained on input to older children (partition 8) uses a more complicated, higher dimensional encoding of the noun category, and therefore should exhibit higher ED than a model first trained on input to younger children. To compute ED, I proceeded as before; I collected the same contextualized representations of all 700 probe words at the output layer of each RNN, stacked each vector vertically to construct a matrix, decomposed it with SVD, and then computed ED.

The results shown in the bottom panels of Figure 10.5 confirm my prediction. Overall, the results closely resemble the pattern of fragmentation (top panels). The number of effective dimensions steadily increases for both groups of models, but there is a dramatic spike in ED and then a return to comparable levels at the transition between the first and second partition (partition 8 and 7) only for models trained in reverse age-order.

The results of the fragmentation and ED analyses also point to a potential mechanistic account of why RNNs trained in age-order (blue) compared to RNNs trained in reverse order (red) on AO-CHILDES produce contextualized — as opposed to lexical — representations that also perform better in the semantic categorization task. While more speculative, the idea is as follows: Given the analyses above, it is clear that the noun category of RNNs trained on partition 8 compared to partition 1 first is splintered into smaller noun clusters, obscuring the fact that nouns are a coherent grammatical/distributional category, and resulting in chunk-level knowledge that is less likely to generalize beyond partition 8. Compared to a model trained on partition 1 first where nouns are less 'sticky' and should therefore generalize better to subsequent partitions, a model trained on partition 8 first will likely need to re-organize its initial clustering to a greater extent in order to be more compatible with data in other partitions of AO-CHILDES. Such re-organization is likely to come at a cost, such as forgetting of previously acquired semantic information. An RNN that exits training on partition 8 with a splintered encoding of the noun category, where each cluster is tied to its own distinct set of 'sticky' pre-nominals, is forced to learn semantic category divisions relative to the clusters it has already acquired. This means that each distinct noun cluster is semantically differentiated in parallel, with little regard to how differentiation in one cluster connects to another. In other words, semantic differentiation in a network trained in reverse age-order may need to distinguish between VEHICLE and MAMMAL in multiple locations in representational space, given starting clusters that mix and match members from each category. I think that a better strategy is to start with a single central noun cluster and then divide the central cluster exactly once for each category distinction, or in a hierarchical fashion whereby sub-clusters are divided in a principled manner.

10.4.2 Divergence from the Superordinate

An additional analysis was conducted to further lend empirical support to the idea that training in reverse age-order results in a splintered encoding of the noun category in the representational space of the RNN during early training. The following analysis is conceptually similar to fragmentation and ED, but is computed very differently, using tools from probability theory rather than linear algebra. If this analysis, which is based on a very different mathematical framework, provides similar qualitative results, this would produce convergent evidence, and make less plausible that the results obtained above are due to an artifact in my analysis. An analysis of this sort has not been previously been done; I will refer to it as 'divergence from the superordinate', and DS for short. It involves the the Kullback-Leibler (KL) divergence, a standard tool for computing the dissimilarity between tho probability distributions in probability theory. Similar to fragmentation and ED, which implicitly provide a measure of how similar individual output layer representations of probe words are to an underlying prototype representations, DS explicitly compares, using the KL divergence, each output representation, q_i with an explicitly computed prototype vector p_i. I proceeded as follows: At consecutive intervals during training, I extracted the contextualized representations of each probe word at the output

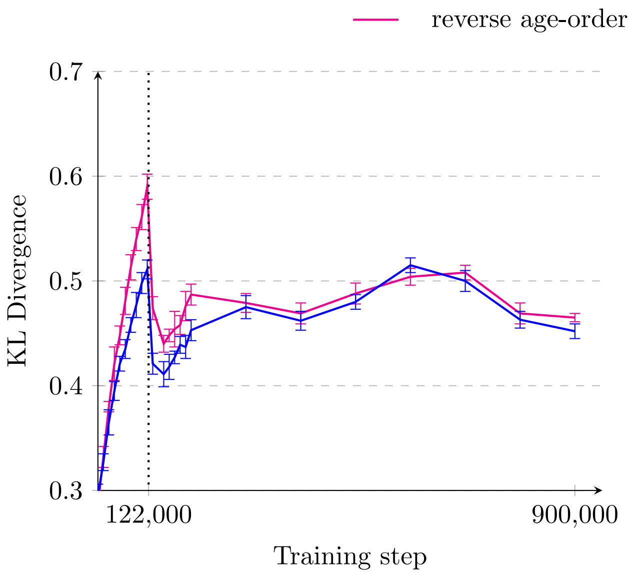


Figure 10.6: Divergence from Superordinate (y-axis) at consecutive intervals during incremental training (x-axis) on AO-CHILDES, for the simple RNN (left panel) and LSTM (right panel). Each line represents an average across 10 RNN simulations, error bars indicate 95% confidence-intervals, and the dotted vertical line marks the end of the first AO-CHILDES partition. RNNs were either trained in age-order (——) or in reverse age-order (——).

layer as described above. Next, each vector q_i is compared to a prototype vector q_i, which, similar to q_i, is a vector obtained by inputting all sequences in which the probe i occurs in AO-CHILDES, except that the probe itself is excluded from the sequence, so that only the left-contexts are input to the RNN. The result is a hidden state vector that represents all the information stored in the contexts of probe i. Next, I obtained all 700 lexical representations and average them to produce a representation of the 'average probe'. I then input this average lexical representation to the RNN which already contains the representation of all contexts in which probe i occurs to contextualize the 'average probe'. Finally, the hidden state is fed-forward to the output layer where I obtained the final prototype vector, p_i. The KL divergence between q_i and p_i was computed using the formula

$$D_{KL}(p_i||q_i) = \sum_{x \in X} p_i(x) \log(\frac{p_i(x)}{q_i(x)}) \tag{10.3}$$

where i indexes probe words, and x indexes output units in the next-word probability distribution X produced at the output layer of the RNN. I call this analysis 'divergence from the superordinate' because p_i can be considered a contextualized representation of the superordinate category of all probe words, which is just the noun category. This procedure was repeated for all probes. Finally, all 700 KL divergences were averaged, as defined in

$$DS = \frac{\sum_i D_{KL}(p_i||q_i)}{n} \tag{10.4}$$

where n is the number of probe words.

The results are shown in Figure 10.6. The qualitative pattern of the results is similar to those observed for fragmentation and ED, and is again independent of the architecture (simple RNN in left panel, LSTM in right panel). The early spike in DS for models trained in reverse age-order (red) indicates that their next-word predictions, at that point in training, diverged more strongly from next-word predictions that would be expected if each probe's lexical representation more closely resembled the representation of the 'average probe'. As evidenced by the sharp rise in DS in both panels in Figure 10.6, there is a huge amount of clustering of probe words during training on the first partition. After exiting the first partition, the RNN undoes much of its initial clustering, presumably because much of it was premature and based on idiosyncratic lexical relationships that do not generalize beyond the first partition.

Altogether, the follow-up analyses of RNN learning dynamics provide convergent evidence for the idea that RNNs trained in reverse age-order on AO-CHILDES produce next-word predictions in line with a premature splintering of probe words in representational space, and that this splintered organization is driven by the stickiness of left-contexts and nouns in language to older children.

10.5 Summary

Previous distributional models of lexical category learning collapsed language input to children of different age groups, making such models difficult to study from a developmental perspective. It is well known that children's language input changes across development, starting much more limited in complexity, and gradually diversifying into its adult form. These changes may have consequences for language learnability (J. L. Elman, 1993; Lany & Saffran, 2010; Neuman et al., 2011). As shown in Chapter 7 and 8, the distributional statistics in the input to younger children better signal the presence of the noun category compared to the statistics of speech to older children (1-3 vs. 3-6 years of age). Findings such as these suggest that current distributional models of lexical category acquisition are missing an important dimension: developmental time. In order for insights from modeling studies to be relevant to child language acquisition, models must learn under more realistic, developmentally plausible conditions. This chapter clearly demonstrates that developmental time and order of data presentation are important factors that guide and constrain learning. In the RNN, training on child-directed input yielded improved downstream performance in a lexical semantic categorization task if data is presented in the order in which children experience it across developmental time. Follow-up experiments (i) ruled out alternative explanations of the age-order effect based on age-related differences in noun density, and (ii) provided supporting evidence that the learning dynamics of early training are consistent with SPIN theory.

Chapter 11

Related Work

The remainder of this thesis does not present novel empirical or theoretical findings, but is an attempt to connect my findings to a broad range of disciplines such as psycholinguistics, language acquisition, and machine learning. In this chapter, I briefly review existing research that in one way or another (whether related to learning in humans or machines) raises similar points and/or concerns presented in the previous chapters. I begin by discussing work in machine learning, and end with behavioral studies conducted with children and adults. Regarding the latter, my work relates most closely to studies of word learning in children, where the goal is to discover what information and strategies children use to acquire the meanings of the words they hear.

11.1 Neural Network Learning Dynamics

In this section, I explore connections of my work to existing research on neural network dynamics. In particular, my focus is on incremental learning on ordered data, and the notion that learning outcomes and learning trajectories differ in important ways depending on the order in which data is presented to a neural network.

11.1.1 Starting Small

In addition to investigating the potential of the simple RNN as a viable model of the supplier of form-based lexical semantic representations to perform DECAF, a secondary aim of this thesis is to make sense of an unresolved debate concerning wok by J. L. Elman (1993) on 'starting small'. This debate was sparked by an observation that the order in which data is presented to the RNN during training constrains what the network can learn. More precisely, J. L. Elman (1993) observed successful learning outcomes only when the model was first presented with simplified sentences, containing few embedded clauses, before being exposed to more complex (i.e. longer) sentences. He coined the expression 'starting small' to refer to this phenomenon, and suggested that a similar phenomenon may govern children's language development. For instance, being exposed to grammatically simpler utterances may enable children to acquire complex grammatical structures that would be difficult to learn in the absence of simplified input. This idea can be understood in terms of limited processing abilities: Infants and children might be predisposed to learn what is just within reach of their developing processing capabilities; to achieve the best learning outcome, novel structures are presented contingent on the learner's skill-level.

In a follow-up examination using a larger range of artificial training data, D. L. Rohde and Plaut (1999) showed that 'starting small' does not always facilitate learning. In fact, the majority of their simulations showed that initially restricting the grammatical complexity of the input actually impeded learning in the RNN. The learning impairment was most severe when the input was made more naturalistic via addition of cross-clausal semantic dependencies. Only in extreme cases in which the input was devoid of such dependencies, did 'starting small' provide an advantage. D. L. Rohde and Plaut (1999) concluded that 'starting small' may be of little help for learning the grammatical structure of a natural language, due to the large number of semantic dependencies present in those languages. Furthermore, they speculated that the data does not need to 'start small', because the RNN itself 'starts small' — the RNN first learns adjacent before learning longer distance dependencies. At best, the authors found, the initial presentation of simplified training data is not necessary, and, at worst, can impede learning.

Importantly, however, D. L. Rohde and Plaut (1999) suggested that initial exposure to simplified input, while not beneficial for learning the grammatical structure of utterances, is likely beneficial for learning their meaning. This raises the possibility that 'starting small' can facilitate semantic development, and by extension, knowledge of lexical semantic category membership. In this thesis, I explicitly tested this possibility. This is a departure from the original work by J. L. Elman (1993) who was primarily interested in learning long-distance dependencies related to morpho-syntax.

Since the publication of D. L. Rohde and Plaut (1999), there has been virtually no follow-up work on 'starting small', nor detailed investigations of the role that order of data presentation plays in facilitating learning in artificial systems. It is surprising how few researchers have revisited the issue. This thesis is an attempt to revive the ideas proposed by J. L. Elman (1993), and to provide contemporary empirical support and theoretical arguments for why the study of the order in which learners are exposed to data is still relevant. This area of inquiry has the potential to bridge basic research in cognitive science with engineering-oriented disciplines (artificial neural networks), and is a starting point for comparing (language) learning in children and in artificial systems.

11.1.2 Curriculum Learning

Currently, many neural network researchers disregard altogether the potential longitudinal organization of their data by shuffling the order of the input with every training epoch. The rationale is to reduce the risk of overfitting. While overfitting is an important concern for deep learning practitioners — shuffling the input often eliminates unwanted statistical irregularities — alternative approaches are rarely considered (but see Bengio et al., 2009; A. Graves et al., 2017). I identified several notable exceptions, discussed below.

In his work on 'starting small', J. L. Elman (1993)showed that the simple RNN better predicts next-words in sequences consisting of complex embedded clauses when it is (i) first trained on simpler sentences with no embedded clauses, and only then (ii) trained on the full range of sentences. According to J. L. Elman (1993), natural language sequences can be broken down into smaller components, such as main and embedded clauses. It is often claimed that Learning these building blocks first might enable learning systems like the RNN to tackle more complex sentences in a step-wise fashion.

While attempts by D. L. Rohde and Plaut (1999) to replicate these findings have largely failed, one of the authors, D. L. T. Rohde (2002) noted in his doctoral thesis that staged training of his Connectionist Sentence Comprehension and Production (CSCP) model yielded improved learning outcomes compared to a shuffled training condition. The CSCP model is similar to the RNN, but a different artificial language was used to train it, and instead of next-word prediction, the model was evaluated in terms of acquisition of semantic knowledge. D. L. T. Rohde (2002) did not further comment on this finding; one reason is that training of the CSCP required several months to complete and retraining using different input conditions would have been prohibitively time-consuming. He noted however that such a result was in accord with D. L. Rohde and Plaut (1999) who claimed that 'starting small' should facilitate learning of the meanings of words, phrases and longer utterances.

Subsequent work concerning curriculum learning was conducted primarily in the area of machine learning and computational linguistics. For instance, Bengio et al. (2009) compared training on the full input with no apparent order to a curriculum learning strategy whereby input is separated into stages according to some measure of task difficulty. For example, a neural network trained to categorize 2D shapes benefited from a curriculum strategy whereby the first half of training examples were sampled from a training distribution with less variability in shape than the full training distribution. Similarly, a neural language model trained to categorize word sequences as grammatical, given samples of a target language, achieved a lower test error when trained incrementally on sections of the input ordered by the size of the vocabulary. The authors suggested that a curriculum learning strategy acts both as a way to find better local minima and as regularization (performance improvement was evident primarily on the test data with little improvement on the training data). Moreover, Bengio et al. (2009) argued that curriculum learning may speed training because a network spends less time predicting examples that are too difficult.

In his doctoral thesis, Mikolov (2012) discovered that ordering chunks of several standard written text corpora based on the perplexity obtained by a 2-gram model, trained on the same data, and excluding any chunks with very large perplexity, improved sequence-learning considerably. This kind of manipulation of training data is roughly equivalent to sorting the data by the amount of uncertainty about an upcoming word. Explaining the motivation for his experiment, Mikolov (2012) noted that complex patterns in the data are based on simpler patterns and that "these simple patterns need to be learned before complex patterns can be learned". This is in line with the 'representational trajectory hypothesis' proposed by A. Clark (1994). Reviewing Elman's work on 'starting small', A. Clark (1994) suggested that restricting the early input to simple examples might direct a learner's representational device into a direction more suitable for learning more complex linguistic abstractions. Specifically, A. Clark, 1994 asserted that failure to learn lower order regularities in the training data should reduce the likelihood that higher order features that depend on those lower order features are learned.

More recent work has focused on more sophisticated methods for creating curricula for neural networks. For instance, A. Graves et al. (2017) studied the usefulness of various learning progress signals for selecting the next sample to train on, and found that some signals can lead to significant gains in curriculum learning efficiency compared to a uniform sampling approach. Prediction gain, which selects the next sample based on the change in loss due to training on that sample, was found to perform best for maximum likelihood training. This measure is used as an indicator of learning progress and is therefore calculated during training. In this way, the curriculum is updated in an online fashion, based on some measure of learning progress, rather than pre-computed. The researchers also noted that the uniform sampling approach presents a strong baseline, and that this may be due to an implicit curriculum inherent in gradient descent training. Because learning is dominated by gradients from tasks which are learned fastest, the direction of gradient descent tends to be in the direction of where the most progress can be made. As such, automating a curriculum may be best viewed as utilizing a learner's natural tendency to learn those tasks which at any given time during training result in the largest training progress.

Curriculum learning has also been explored in the area of natural language translation. As an example, Zhang et al. (2018) compared a diverse range of curricula for learning to translate German to English in a sequence-to-sequence neural network. A number of observations were made: First, Zhang et al. (2018) found that a simple word-frequency based criterion for ordering training examples was most advantageous — not only compared to a shuffled baseline, but also to a much more computationally expensive method which assigns difficulty ratings based on the error of another machine translation model. Further, using word frequency to assign difficulty ratings also outperformed other criteria, such as sentence-length (shortening training time by up to 30%). In contrast, ordering training examples by increasing sentence length only minimally improved convergence time, and in only one of ten different simulations. The success of other difficulty criteria was largely dependent on the initial learning rate. Zhang et al. (2018) concluded that:

> ...curriculum learning can improve convergence speed, but the choice of difficulty criteria is key... No single curriculum schedule consistently outperforms the others, and results are sensitive to other hyperparameters such as initial learning rate and curriculum update frequency.

Finally, P. A. Huebner et al. (2021) found that when a Transformer-based neural language model[1] was trained in age-order on a concatenated corpus designed to mimic the unfolding language environment of children and young adults, the network achieved significantly higher scores on a benchmark of grammatical knowledge than the same network trained in reverse order. The authors speculated that the age-ordered curriculum supported an early discovery of useful grammatical categories — much like the scaffolding effect of the discovery of the noun category on lexical representation learning in the RNN, reported in Chapter 10. Importantly, the model in the age-ordered curriculum condition, was first trained on AO-CHILDES, which would have exposed the model to a large number of high-entropy slots in high-frequency frames at the earliest stage of training (Bannard & Matthews, 2008; Cameron-Faulkner et al., 2003). This work shows that age-ordered training on child-directed input is not only useful for scaffolding the performance on semantic tasks (e.g. semantic categorization), but also tasks requiring judgment of grammaticality.

11.1.3 Staged Learning

There is now ample evidence that demonstrates that learning in neural networks is not homogeneous across training steps; rather, there are distinct periods during training marked by qualitative differences in what is learned, and how. One of the most striking demonstrations of this comes form work by Achille et al. (2018) who studied how degradation of the input to a convolutional neural network (CNN) trained to classify images, at different stages of training, influenced the network's classification performance at the end of training. While the CNN was able to recover from high-level degradations of the input during the earliest stages of training, the network did not recover from lower-level manipulations such as blurring of input images. In contrast, identical manipulations during later stages of training did not influence classification accuracy of the final model. The authors compared their findings to 'critical periods' previously observed in animal models of visual perception, and argued that the observance of similar critical periods in CNNs points to computational similarities between biological and artificial neural networks. To better understand their results, the authors analyzed the weights of various CNNs during training, and found that information rises quickly in the earliest stages of training, and then decreases. Achille et al. (2018) argued that the decrease in information captured

[1] P. A. Huebner et al. (2021) developed a new model they called BabyBERTa, which is a smaller version of a much larger Transformer called RoBERTa. BabyBERTa is a masked language model, meaning that it was trained to predict randomly masked words given both left and right sentential context as input, as opposed to next-words given only left context as input

in the weights of the CNN as training progresses prevents redistribution of information resources, and termed this phenomenon 'information plasticity'. More specifically, the authors proposed that strong connections are learned during the first few epochs of training which are optimal given the input data distribution, and that once learned, these strong weights do not change with additional training. If not established during this critical phase, such strong connections are difficult to learn with additional training, resulting in poor asymptotic behavior (i.e. performance at convergence).

More evidence for staged learning comes from T. A. Chang and Bergen (2022) studied how recurrent and other neural language models acquire individual words during training. The authors extracted learning curves (based on surprisal) for over 600 words from the MacArthur-Bates Communicative Development Inventory, and correlated them with empirical age-of-acquisition data. The results are evidence of distinct stages during training, during which language models adjust next-word predictions in a step-wise fashion. For instance, early in training, language model surprisals tended to reflect surprisals based on uni-gram frequencies, indicating that models had not yet learned to take into consideration contextual information when predicting upcoming words. Eventually, the models began to diverge from the uni-gram distribution, and continued to approach the bi-gram distributions. Interestingly, models only diverged from the bi-gram distribution when their surprisals matched surprisals based on bi-gram frequency. This incremental convergence on and divergence from higher-order n-gram distributions across training suggests that language models over-fit on lower-order n-gram probabilities before learning higher-order n-gram probabilities. T. A. Chang and Bergen (2022) concluded that language models expand their use of contextual information when making next-word predictions in step-wise fashion — learning to make more nuanced predictions with each step.

In a related line of work, researchers have examined which linguistic features an RNN language model learns at different times during training relative to other features. For example, Saphra and Lopez (2019), used an analysis method called Singular Vector Canonical Correlation Analysis (SVCCA) to compare the learned representations of an LSTM across training steps, and found that part-of-speech is learned earlier than semantic features, and topic. This is in line with ideas and simulations presented in this thesis which has emphasized the need to learn features relevant to grammatical class prior to learning features diagnostic of finer-grained lexical semantic distinctions (Chapter 6 and 9).

One of the most convincing demonstrations for the idea that neural networks acquire knowledge in stage-like fashion comes from formal analyses of deep linear networks by Saxe et al. (2019). Using singular value decomposition (SVD) of a dataset with hierarchical category structure, the authors obtained a complete analytic derivation of the development of internal representations in the deep network. Not only did their derivation accurately predict empirical learning rates, but it provides a strong mathematical argument why the network learned superordinate category distinctions before basic category distinctions. This pattern of learning progressively finer-grained distinctions is termed 'progressive differentiation', and the analytic derivation by Saxe et al. (2019) demonstrates that this phenomenon is an inevitable consequence of deep-learning dynamics exposed to hierarchical structure. Superordinate categories are learned faster than subordinate categories because

> ...items corresponding to broader hierarchical distinctions have stronger statistical structure, as quantified by the singular values of the training data.

This leads to waves of differentiation in deep — but not shallow — networks, and is compared to how children gradually acquire conceptual structure (Inhelder & Piaget, 1958; F. Keil & Sessar, 1979; Mandler & McDonough, 1993).

Similar evidence for stage-like differentiation during training was observed in a continuous-time RNN used to study the emergence of grid-cells, which are responsible for navigation in people and many other animals. Cueva and Wei (2018) showed that units in the RNN that correspond to grid-cells differentiate out of border-cells rather than emerging independently out of non-specialized units. Unlike grid-cells which are tuned to particular locations in a topological mental map in a grid-like fashion, border-cells are tuned to the perimeters of locations. More specifically, the authors showed that units corresponding to border-cells emerge first during training on a navigation task, and that units which eventually become grid-cells with additional training, initially have tuning properties similar to border-cells. This functional specialization is not unlike the progressive differentiation observed in the RNN by Saphra and Lopez (2019) and the deep network studied by Saxe et al. (2019). The findings of Cueva and Wei (2018) provide further evidence that specialized units do not emerge in a vacuum in artificial neural networks; rather, specialization appears to be a response to limitations on existing units to minimize error at a specific time during training. In sum, the evidence presented in this section points to the idea that there a sub-tasks that an untrained network may need to solve first before learning additional strategies to compensate for any remaining error.

11.1.4 Simplicity Bias

Emerging evidence from analyses of learning dynamics of neural networks has shed light on the biases imposed by stochastic gradient descent (SGD), the update rule used to train almost all modern neural networks, including the RNN. This work is based on the premise that SGD based learning systems favor acquisition of some aspects of the data over others, and that the choice of features a network pays attention to — especially when there are multiple diagnostic features — has important consequences for how it generalizes. For example, Shah et al. (2020) trained and tested CNNs on artificial image-based datasets with features that varied in complexity. The authors observed that many neural networks exhibit an extreme 'simplicity bias' where networks rely exclusively on the simplest feature while remaining invariant to all other — more complex — predictive features. Further, the authors showed that this bias could explain why small shifts in the distribution of the data across training time can significantly degrade performance. If a model has learned only to pay attention to one predictive feature to solve a particular classification task, the model will fail to generalize to data where this feature is no longer predictive — despite the availability of other features previously available to the network. Lastly, Shah et al. (2020) showed that the simplicity bias can also hurt generalization in the absence of a distribution shift: Many models preferred to encode simpler features at the expense of more complex features even when these simpler features have less predictive power than the more complex features. In related work by des Combes et al. (2018), this maladaptive tendency of SGD based training of neural networks was termed 'gradient starvation'. The authors explained the failure to utilize more predictive, but also more complex, features in terms of the learning trajectory through weight space: Learning simpler rules first may inhibit the acquisition of more complex rules, because simpler features tend to be more frequent and therefore dominate (i.e. 'starve') the gradient directed at less frequent, more complex features.

11.1.5 Prior Knowledge, Facilitation, and Interference

Many psychologists agree that depending on the nature of the representations that are shaped by training, some learning problems are facilitated, and others are more difficult. The work herein supports this view from the view of connectionist modeling. These result fit with a small, but long-standing literature showing effects of prior knowledge on learning in neural networks.

For instance, in an effort to study children's age of acquisition of words using orthographic and phonological representations as input to a neural network model, Zevin and Seidenberg (2002) found that later learning is facilitated by prior knowledge of the input dimensions relevant to the word-learning task. Similarly, exposing a network to lots of examples from a training distribution can promote the discovery of abstract structures useful during structured generalization tasks. For instance, Calvo and Colunga (2003) modeled children's rule-based generalization in the RNN, and found that pre-training the network on pseudo-random sequences is crucial to promote the discovery of an abstract rule — 'repeat what came before', e.g. AB → B, BA →. Without such pre-training, the RNN did not generalize the abstract rule to novel items. Contrary to claims made by Marcus et al. (1999) that RNNs cannot generalize abstract rules defined to novel items, this work shows that with appropriate pre-training exposure, the RNN does learn abstract rules, including duplication (i.e. 'repeat the item that occurred in the previous slot').

Findings such as these are not specific to recurrent networks; training an auto-encoder to classify artificial 2D stimuli, Roark et al. (2022) observed that the category that was learned the best was the category that was most closely aligned with the the long-term regularity experienced during training. Classification performance was worst for categories that were misaligned with the long-term regularity, and intermediate for other categories. In other words, having extensive experience with regularities along a particular dimension supports the discovery of novel categories that vary along this dimension. Roark et al. (2022) called such dimensions 'axes of high variance'. Collectively, demonstrations like these clearly demonstrate a bias to carve novel categories and concepts from existing categories. If a neural network is tasked to classify input using a novel label which requires distinguishing items along axes that are orthogonal to axes of high variance, we can be fairly confident that the network will not be able to do so quickly, given that it must reconfigure itself to detect differences along these dimensions — a process that is much slower than using existing units sensitive differences along axes of high variance. This is closely associated with 'catastrophic interference', a phenomenon that occurs when a neural network is trained on novel examples that do not conform to the task or distribution previously learned. Contrary to facilitation due to pre-training or exposure to input with variation along dimensions useful for classification, catastrophic interference has exactly the opposite effect: If novel examples are in conflict with existing knowledge, neural networks tend replace previously learned knowledge, rather than integrate it with new information (McCloskey & Cohen, 1989).

Recently, Mannering and Jones (2021) found that catastrophic interference is relevant in the domain of distributional modeling: The authors found that Word2Vec, a widely used 1-layer network that represents co-occurrence associations of a target word with other words in a high-dimensional vector, almost completely forgets a word's dominant sense (*bank* → *money*) if subsequently presented with examples of its subordinate sense (*bank* → *river*). Incremental training on non-stationary input, therefore, is a two-edged sword: While some data may scaffolds learning, in other instances, it has the potential to erase existing knowledge — a phenomenon that is extremely unlike biological systems.

New evidence in the field of computational linguistics has revealed that the data that an RNN is exposed to during early training has important consequences on its ability to generalize long-distance dependencies. Inspired by the finding of J. L. Elman (1993) that 'starting small' helps learning dependencies across embedded clauses, Saphra and Lopez (2020) studied how LSTMs learn the long-distance rule between the two English function words *either* and *or*. These two connectives are often separated by long clauses, as in '*Either the gorilla will attack, or she will remain calm*'. In line with ideas presented in this thesis, the authors found that initial exposure to sentence fragments that are later used as intervening spans, speeds learning of the long-distance rule between *either* and *or*, but are worse at generalizing to novel intervening spans. In other words, familiarity with intervening spans is detrimental to generalizing the long-distance rule to unfamiliar intervening spans. Follow-up analyses of their networks revealed that LSTMs exposed to familiar intervening spans learned representations for the word *either* that were highly inter-dependent with the familiar spans — the networks had scaffolded their learning of the long-distance rule using familiar spans. As a consequence, these networks relied on familiar intervening spans to reliably predict the long-distance target *or*. In contrast, exposing LSTMs to a diverse range of long, and low-frequency intervening spans improved generalization of the long-distance rule to novel intervening spans. These findings are somewhat counter-intuitive given the unspoken assumption in computational linguistics that learning complex rules might be facilitated when representations of simpler patterns are already available. Saphra and Lopez (2020) explain their findings in terms of a 'familiar conduit' (i.e. a familiar intervening span):

> ...the idea is that it is easier for the RNN to predict, given [the word *either*], the first word in a familiar conduit ... and let the words in the familiar conduit predict each other, and predict the [the word *or*], as opposed to predict, given [the word *either*], [the word *or*] (in a delayed manner), which requires memorizing the prediction across the entire conduit (which would require adapting already learned representations for items in the familiar conduit).

The authors concluded their work by noting their findings are relevant to the development of training curricula for RNNs. Specifically, they suggested that measures of length, and syntactic depth are likely inappropriate measures for curriculum learning. Instead, high slot diversity (number of intervening spans) appears to be more promising for scaffolding the learning of long-distance rules. This conclusion aligns with work presented herein, which demonstrate that counterbalancing left-contexts of target words during early training has long-lasting benefits for lexical semantic categorization.

11.1.6 Promoting Structured Decomposition of Input

Another line of related work concerns the development of training data that explicitly promotes structured decomposition of the input into independent components. An example of this is the training strategy developed by Hill et al. (2019) that endows a simple connectionist model the ability to perform well on analogical reasoning tasks by forcing the model to encode abstract relational structures. This is achieved by by training the model on carefully controlled pairs of stimuli that contrast relational structures that cannot be solved by taking non-relational shortcuts. Similarly, work by Vankov and Bowers (2020) showed that combinatorial generalization can be induced in a simple feed-forward connectionist architecture by forcing the model to predict so-called VARS vectors that highlight the combinatorial structure in the data, not unlike how input to younger children emphasises regularities at superordinate levels in the lexical category structure of natural language data. Together, these findings demonstrate that with the right pressure to represent knowledge in a more componential fashion (e.g. atomically), networks are able to perform well on difficult tasks that were previously solvable only by models with built-in relational structure and/or the ability for rule-based manipulation of symbols. Both methods use only the input data to achieve these goals, as opposed to model-based interventions.

11.1.7 Isotropy in Semantic Vector Space

The analyses of RNN learning dynamics presented in Chapter 10 are closely related to the idea of 'isotropy' in semantic space. Broadly, isotropy simply means uniformity in all orientations, and in the context of semantic modeling, isotropy refers to uniform usage of all the dimensions of a vector space (Cai et al., 2020; Mu & Viswanath, 2018). When features relevant to grammatical class are considered as part of the learning process, the fact that, say, nouns, occur in similar contexts, results in the formation of a highly anisotropic semantic space, such that representations of nouns are spread only along a select few dimensions. That is, distributional information relevant to noun category membership constrains the otherwise uniform spread of representations along multiple dimensions. Whether this is advantageous or not depends on the task a model is used for: If the model is used to discover grammatical classes, anisotropy is potentially desirable; however, recent work suggests that for modeling lexical semantic tasks, isotropy — greater utilization of existing dimensions — is preferred (Bullinaria & Levy, 2012; Cai et al., 2020; Mu & Viswanath, 2018). For example, Bullinaria and Levy (2012) observed that removal of the first 100 singular dimensions of a co-occurrence matrix improved performance in four lexical semantic tasks. The authors concluded that singular dimensions (aka. principal components) that capture large variances in corpus data are often "contaminated with aspects other than lexical semantics". Similarly, Mu and Viswanath (2018) found that the removal of all but the top singular dimensions from representations learned by Word2Vec improved performance on lexical semantic similarity, categorization, and analogy tasks. This leaves unanswered an important question: What exactly do the top singular dimensions encode? The work in this thesis proposes a potential answer, namely that learning in distributional systems is not exclusively guided by information about lexical semantics, but also by information relevant to grammatical categorization, and word-order. While it is true that such factors can be considered nuisance variances to be removed at the end of training for the purpose of improving performance on lexical semantic tasks, I argue that constraints imposed by these factors during the earliest stages of training are helpful for organizing the representational landscape in which lexical semantic representations emerge during later stages of training. The findings in Chapter 10 are particularly revealing in this respect. The analyses of fragmentation and effective dimensionality in the learned representations show that anisotropy is in fact desirable during early training on child-directed input. Strong anisotropy during early training reduces the likelihood that the network encodes category-irrelevant associations that can have long-term negative consequences on subsequent learning.

11.2 Human Learning Dynamics

The theory developed in this work was designed to explain the behavior of the simple RNN, but the ultimate goal is to understand the statistical learning system that children might use to acquire form-based lexical semantic representations to support their word learning (e.g. DECAF). The purpose of the following sections is to relate the theory presented in this work to learning in people, and especially children. In particular, I focus on whether and what kind of scaffolding effects (such as the one observed in Chapter 10) have been documented in studies of human learning. My examination of these questions will focus primarily on learning in children, and in the domain of language, but learning in adults and in non-linguistic domains are also considered.

11.2.1 Scaffolding Effects

SPIN theory predicts that in order to learn more atomic — and therefore more generalizable — lexical semantic representations, children benefit by acquiring coherent part-of-speech (POS) classes prior to learning about lexical semantic categories. For instance, to learn semantic categories of nouns, children must first isolate and disentangle purely grammatical information such as word-order from the input they hear, in order to establish the noun category. After doing so, children should be able to make more rapid progress learning semantic variation within the noun category. While this prediction was borne out in simulations, is there any behavioral evidence that such scaffolding takes place in children?

Inspired by the work of J. L. Elman (1993) on 'starting small', Poletiek et al. (2018) tested the hypothesis that an incrementally ordered presentation of samples from an artificial language can help people learn recursion faster. In particular, the order of presentation that was examined was growth according to structural complexity and growth according to string length. The results showed that learning of center-embedded structures was facilitated when participants learned from stimuli that were presented in the order of growing structural complexity. Participants in this condition were better able to generalize what they have learned about simple units in the artificial grammar to more complex units. This facilitation is likely due to scaffolding provided by initial exposure to structurally simple units.

Another line of evidence for scaffolding effects comes from statistical learning studies. For instance, Lew-Williams et al. (2011) evaluated the ability of English-learning infants to segment an Italian speech stream. It was found that familiarization with a sample speech stream alone was not sufficient for allowing infants to detect word boundaries. Instead, successful detection of word boundaries was detected only when infants were familiarized with the same speech in combination with words heard in isolation. Because isolated words are a frequent occurrence in child-directed speech (Brent & Siskind, 2001; Fernald & Morikawa, 1993), the authors suggested that one word utterances may play a role in preparing infants for future language tasks. Indeed, prior work showed that there is a benefit of exposure to one-word utterances on word recognition in sentences (Gout et al., 2004; Houston & Jusczyk, 2000) and on later vocabulary development (Brent & Siskind, 2001). The authors explain that words spoken in isolation "pop out" and therefore provide salient markers in fluent speech for word segmentation. This finding connects with the results of my corpus analyses which show that nouns in input to younger children tend to pop out better (Chapter 7), and that hearing nouns in more consistent and shared contexts first has long-term advantages for learning about semantic category membership.

While the work of Lew-Williams et al. (2011) showed that knowledge about single words can influence subsequent knowledge, what about co-occurrence relations between words? After all, the theory developed in this work is not about properties that are tied to individual words (e.g. their sound), but about how words are defined in relation to each other. Is there evidence in the behavioral literature that infants benefit from learning simple co-occurrence statistics first? Work by Lany and Gómez (2008) showed that this is the case. The authors asked whether exposure to adjacent dependencies would facilitate learning of related nonadjacent dependencies. Briefly, the experiment was conducted as follows: Infants 12 month of age were familiarized to an artificial grammar consisting of two-word sequences in which words from either category A or B occur in sequence-initial position, and words from either category X or Y occur in sequence-final position. There were two familiarization conditions: In the control group, words in category A and B did not predict the category of the next word, whereas in the experimental condition, words in category A and B did predict the category of the next word (A was consistently paired with X and B with Y). Following an 8 minute familiarization, infants were habituated to 3-word sequences which followed the same structure as those heard in the experimental condition, except that AX and BY sequences were separated by words in a novel category C (e.g. ACB and BCY). During testing, infants were exposed to sequences which violated the nonadjacent dependency seen during habituation. Successful learning of the nonadjacent dependency was quantified as a significant increase in mean listening time for the test trials compared to the last two habituation trials. Because 10 month old infants typically fail at learning nonadjacent dependencies, it was not surprising that infants in the control familiarization condition did not show learning. However, infants in the experimental condition who were exposed to adjacent dependencies between A and X and B and Y, did learn the nonadjacent dependency. The authors concluded that learning nonadjacent dependencies is facilitated by exposure to simpler instances of such structure. This means that infants can generalize from their knowledge of simple structural relationships (e.g. A predicts X) to more complex relationships (e.g. A predicts X after some intervening C).

The importance of the starting conditions has also been recognized in infants' acquisition of adjectives. For example, T. H. Mintz and Gleitman (2002) found that novel adjectives were not readily acquired by 36- and 24-month olds without first being provided rich referential and syntactic information about the meanings of the novel adjectives. Specifically, infants only acquired novel adjectives if they were used to modify nouns referring to specific and familiar objects rather than vague objects (e.g. labeled by *thing* or *one*). T. H. Mintz and Gleitman (2002) concluded that their findings "favor an account of lexical acquisition in which layers of information become available incrementally, as a consequence of solving prior parts of the learning problem."

More support for the notion that prior experience can influence learning comes from an artificial grammar learning study conducted by Marcus et al. (2007). In this study, infants heard sequences of artificial sounds (e.g. musical tones, animal sounds) in which items in a particular position are repeated in a rule-like pattern (e.g. ABB or ABA). Next, children were tested on their ability to discriminate pattern-consistent sequences from pattern-inconsistent sequences. The authors found that successful discrimination was contingent on whether they had first heard the same sequences instantiated with speech sounds. Put differently, infants were better at discriminating novel pattern-consistent and pattern-inconsistent sequences when they had previously heard the same patterns in sequences composed of speech sounds. The authors concluded that speech is special to infants, in that it can "catalyze" learning. An alternative interpretation, also provided by the authors, is that speech sounds are already highly familiar to infants, and that familiarity with elements in the input may give learners access to less salient dependencies between those items. On this view, prior exposure to a particular set of items would allow the learner to uncover the most salient dependencies between those items, and thus free cognitive resources for the detection of less salient dependencies. Scaffolding effects have also been observed outside of statistical learning studies, such as in concept acquisition. On such demonstration comes from Kotovsky and Gentner (1996) who showed that prior exposure to lower-order relations supports acquisition of higher-order relational concepts. Whereas 6- and 8-year-olds were successful at recognizing higher-order relational similarity across different dimensions such as size and saturation, 4-year-olds only succeeded if the stimuli to be compared shared lower-order features, such as size.

More behavioral evidence for scaffolding effects on learning comes from a visual category learning study conducted by Horst et al. (2005). The authors were specifically interested in how categorization unfolds over time. In a visual familiarization task, 10-month-olds were exposed either to exemplars characterized by a common function or appearance. When learning exemplars characterized by a common function, infants were initially most sensitive to the common feature, and acquired individual features of exemplars later. On the other hand, when learning exemplars characterized by a common appearance, infants were initially most sensitive to the features that were unique to each exemplar, and only learned the common feature later. In both cases, subsequent learning was scaffolded by prior experience.

11.2.2 Category Refinement

There is reason to believe that the course of category learning during early childhood shares deep similarities with the emergence of lexical category knowledge in the RNN. For instance, category learning studies have revealed that children tend to (i) carve novel categories from previously acquired superordinate categories, and that (ii) young infants initially prioritize learning compact prototype representations of categories. I discuss each in turn. First, Taylor and Gelman (1989) studied how young children incorporate new word meanings into their developing lexicon. Two-year-olds were first taught new labels for toys that already had known labels, and then asked to select the object with the novel label from an array of 4 toys. Based on their selections, the author noted that 2-year-olds tended to interpret the novel label as referring to a subordinate of the known category label. Based on these observations, Taylor and Gelman (1989) concluded:

> Although not all alternative explanations have been ruled out, these results suggest that, from a very young age, children may spontaneously form language hierarchies when they hear a novel word for an object that already has a familiar name.

Second, a study conducted by Younger (1990) found that 10-month olds are more likely to judge a novel prototype as more familiar than exemplars previously seen during a familiarization phase. Judgements made by 13-month-olds exposed to the same exemplars showed the opposite pattern: Previously seen exemplars were judged as more familiar than an unseen prototype. The author concluded that 13-month-olds are more likely to form an abstract prototypical representation of previously seen exemplars and discard idiosyncratic features of exemplars. The authors also suggested that as more memory resources come online over developmental time, learners gain greater capacity for remembering item-specific features tied to specific exemplars. This is consistent with the learning dynamics of the RNN which initially encodes dimensions shared by all members of a category, and only then begins to remember additional dimensions that distinguish members of the same category.

11.2.3 Biases in word Learning

The findings in this thesis suggest that in order to learn useful lexical semantic representations in the RNN, the network would be helped by constraints that operate at the lexical level. To accomplish this, we may turn to cognitive psychology for inspiration concerning how prior knowledge guides and constrains lexical semantic development in children. For instance, Waxman and Markow (1995) conducted a series of novelty-preference experiments with 12-month-olds and 4-year-olds, and observed that both nouns and adjectives focused infants' attention on object categories. Waxman and Markow (1995) explained their finding by arguing that children begin the process of acquiring word meanings by assuming that all novel words initially refer to object categories, as opposed to object properties. The authors observed that only later during development do children begin to differentiate novel labels as object category labels or object property labels. This starting assumption preferentially guides children's learning towards the discovery of nouns, much like the statistical structure of language to English-learning children initially emphasizes the discovery of the noun category, as shown in Chapters 7 and 8.

Many other word learning biases have been proposed: Syntactic cues (R. W. Brown, 1957; Gelman & Taylor, 1984; Katz et al., 1974), lexical contrast (Au & Markman, 1987; E. V. Clark & MacWhinney, 1987), and selective attention to certain perceptual cues (S. S. Jones et al., 1991a). While many of these biases are meant to constrain the set of candidate word-referent pairings considered by infants, similar and/or additional biases could be straightforwardly extended to the language-internal distributional domain, where the candidate pairings are among words. In fact, there is much overlap in the ideas that have been proposed in the psychological literature concerning biases for word-referent association learning and the ideas proposed in this thesis concerning the need to distinguish category-relevant from category-irrelevant lexical associations. For example, the idea that previously acquired knowledge about individual examples can organize learned knowledge in ways that may promote encoding of specific aspects of future examples (possibly at the expense of others) has a long history in the word learning literature. A classic demonstration is found in the work by S. S. Jones et al. (1991a). In a series of word learning experiments, the authors observed that children selectively attend to some stimulus dimensions (e.g. shape and texture) over other dimensions when generalizing learned object labels to novel objects. Importantly, the authors argued that the learning bias that children use to guide their generalizations emerges from experience with previously learned word-referent associations that have been abstracted to highlight statistically predictive perceptual dimensions. S. S. Jones et al. (1991a) proposed that this abstract knowledge, in turn, guides lexical semantic development, by highlighting what kinds of semantic properties are relevant to what kinds of categories.

More recently, behavioral work with English-speaking college-aged participants showed that language-internal cues can facilitate word-referent mapping. In particular, Monaghan and Mattock (2012) showed that the presence of highly frequent function words (e.g. determiners and prepositions) reliably pick out referring expressions (e.g. nouns) in child-directed multi-words utterances, and suggested that this may aid word-referent mapping in a cross-situational word learning task. Crucially, in this task, participants heard multi-word utterances with potentially multiple referring expressions; whereas in one condition, referring expressions were reliably marked by function words, in another condition they were not. The authors showed that the presence of function words resulted in greater word-referent mapping performance at the end of training, despite greater linguistic complexity. Based on their findings and previous work on the distributional learning abilities of children, the authors concluded that that children's "nascent distributional knowledge has an impact on learning word meanings" (Monaghan & Mattock, 2012). Interestingly, their experimental study also showed that early during training, the grammatical structure of the language began to positively influence learning before many of the word–referent mappings had been learned — consistent with the motivation for the staged training regime proposed in Chapter 9.

11.2.4 Resistance to Knowledge Restructuring

The age-order effect is a testament to the lasting effect of knowledge acquired in the past on performance evaluated long after initial exposure. This observation is consistent with an emerging literature on the persistence of implicit statistical knowledge on task performance even when previously acquired implicit knowledge is no longer predictive of current task structure. As an example, Kóbor et al. (2020) showed that, under implicit learning conditions, processing continues to be shaped by previously acquired transitional probability structure even after that structure has been removed from participants' learning environment. In a four-choice reaction time task that required prediction of upcoming stimuli in a visual stream, unpredictable transition probabilities were processed according to prior expectations, despite evidence that such expectations were no longer relevant. Instead, participants required significant amounts of additional exposure to the unpredictable structure in order to update their prior expectations — exposure to more stimuli than required to acquire prior expectations in the first place (Kóbor et al., 2020).

An understanding of the difficulty of knowledge restructuring, especially in the domain of category learning, has existed at least since Lewandowsky et al. (2000). In their study, participants were trained in a categorization task that involved judging whether a fire was wind-driven or slop-driven based on textual descriptions of various fire situations. Importantly, the task could be solved in one of two ways, (i) using an expedient — but less precise — strategy, based on a single feature, and (ii) a more complex — but perfect — strategy, based on the conjunction of two features. By default, participants adopted the expedient strategy, unless instructed to use the more complex strategy with the help of a diagram explaining how to do so. Surprisingly, Lewandowsky et al. (2000) found that participants who had been shown the diagram halfway through training, were unable to re-structure their knowledge by adopting the more complex strategy. Rather, participants tended to continue using the more expedient strategy based on a single predictor. Even when participants were encouraged to use the more complex strategy by showing them an adaptive display, participants still failed to do so. Only when the diagram was presented prior to start of categorization training, did participants adopt the more complex strategy. The authors explained their findings by appealing to simple associative learning principles. These, and other studies demonstrating the persistence of previously learned simple predictors (Edgell & Morrissey, 1987) closely correspond to the staged learning dynamics of the RNN proposed in Chapter 9.

The idea that experience with specific samples of language may result in the formation of adaptive or potentially maladaptive expectations has also been discussed in the production literature. For instance, language learners rapidly learn the phonotactic regularities of their native language — statistical constraints on the set of syllable-to-syllable transitions that are possible in a language. These phonotactic constraints, once encoded, may be more or less compatible with the statistics of a novel language: Having learned, for example that /s/ frequently precedes /i/ in English, people have little difficulty producing the word *sing*, which is compatible with previously acquired phonotactic constraints. However, when people learn a new language that violates the phonotactic constraints of their native language, people tend to make more production errors compared to others with native exposure to the target language (Flege & Davidian, 1984; Yavaş & Someillan, 2005). Production of novel words that violate previously learned phonotactic constraints is difficult and has been shown to require sleep consolidation, during which, it is thought, knowledge re-organization takes place (Anderson & Dell, 2018; Dell et al., 2021). Moreover, Anderson et al. (2019a) showed that learning to reverse a previously learned novel phonotactic regularity is more difficult than learning the same regularity in the absence of exposure to the reversed rule. The authors accounted for their findings using simple associative learning principles, and concluded that phonotactic learning is incremental. The latter point is important because it means that previously acquired knowledge that is relevant to the same task must first be 'unlearned' in order to make room for new — conflicting — knowledge. This theory of learning contrasts sharply with approaches based on algebraic rules, where moments of declarative insight are explicitly encoded without the need to 'make room' or update previously learned connections.

Similarly, numerous studies have shown that learning second-language speech sounds is influenced by previous language exposure. Specifically, the amount of conflict between one's native language sounds and novel sounds determines the difficulty involved in learning the novel sounds (Best et al., 1988; Lotto et al., 2004). As an example, the distinction between the English /r/ and /l/ speech sounds is difficult for native Japanese listeners, whose language does not distinguish sounds along this dimension. Neural network models are often invoked to provide insight why this difference in learning difficulty exists (Roark et al., 2022). Because novel information is represented in the same substrate used to store knowledge of previous experiences, the effectiveness at encoding depends on the degree to which dimensions relevant to the novel information are already represented. The age-order effect (Chapter 10) is a clear example of this — without clearly representing dimensions along which nouns co-vary, the semantic category relations between nouns (i.e. do *gorilla* and *boat* belong to the same category?) are entangled with non-noun related dimensions that make it difficult to recover semantic distinctions in a principled manner. Roark et al. (2022) explain the influence of prior experience as such:

> ...long-term experience with a native language enhances representation of dimensions that are

> relevant to that experience and diminishes representation of dimensions that are irrelevant. The resulting effect is that input categories that align with learners' existing representations, maximizing distinctions that need to be made, are readily learned and input categories that are orthogonal to those representations are difficult to learn.

Similar findings have been documented in the speech segmentation literature (Finn & Kam, 2008).

11.2.5 Blocking

Another psychological phenomenon that the work in this thesis is relevant to is 'blocking'. Broadly speaking, blocking refers to a situation where a single learned predictor of some target (e.g. event) remains the dominant predictor even after additional predictors have been associated with the same target (Kamin, 1967). It has been considered a principle of associative learning, which readily accounts for blocking in terms of prediction error minimization. If some context reliably predicts a target with little error, it is difficult to associate additional contexts with the same target given that there is little error left to warrant additional learning (Lewandowsky et al., 2000). In other words, subsequent predictors — even though they may reliably co-occur with a target — will not acquire much predictive power for that target on its own; instead, in order to activate the target, the original predictor must also be active. For example, consider that A signals the target event T during initial training, and subsequently an additional context B is introduced that signals T in combination with A. At test, B will not reliably signal T on its own (e.g. no response is elicited), despite being equally predictive of T as A. While blocking plays a powerful role in human contingency judgments (Williams et al., 1994), it has also been observed in studies of category learning (Gluck & Bower, 1988; Shanks, 1991). Blocking is conceptually identical to the explanation proposed for why staged learning in the RNN can immunize the network against encoding irrelevant, spurious lexical associations (Chapter 9). By training the RNN to minimize prediction error as much as possible during early training (when first-order relations between nouns and category-relevant upcoming words are emphasized), there is little error left during later training that the network could use to overwrite atomic associations with (potentially maladaptive) chunk-level statistics.

11.2.6 Transfer Asymmetry

As the results presented in Chapter 10 show, the order in which stimuli are learned has long-lasting consequences for how the RNN generalizes learned knowledge to a novel task. Similar observations have been made in studies of visual categorization. For instance, Schyns and Rodet (1997) studied how people generalize knowledge gained during a series of visual categorization trials as a function of the order in which trials were presented. Participants were trained to categorize visual stimuli consisting of one of 3 types of features, X, Y, and XY, where X and Y are shapes that are diagnostic of membership in different categories, and XY is the concatenation of the two shapes, and is diagnostic of membership in a third category. There were two conditions: In the first, participants were trained to categorize stimuli in the order X → Y → XY. In the second condition, participants were exposed to learning trials in reversed order; they were trained to categorize stimuli in the order XY → X → Y. At the end of training, both groups of participants had seen exactly the same stimuli. At test, participants were shown a novel visual stimulus where the shapes X and Y were both presented, but without being concatenated; call this shape X-Y. Interestingly, participants in the first condition categorized the stimulus X-Y as an instance of XY; in contrast, participants in the second condition categorized X-Y as either an instance of X or Y, but not X-Y. This pattern of result indicated to Schyns and Rodet (1997) that participants who had been trained in the order X → Y → XY learned to rely exclusively on the separate features X and Y, whereas participants who had been trained in the order XY → X → Y had additionally learned the compound feature XY. Importantly, this revealed an asymmetry in how people transfer learned knowledge: Participants readily learned the individual features X and Y when the compound feature XY had been presented first, but participants did not acquire the compound feature when they had already learned each feature, X and Y, separately. Further, the authors were able to account for this transfer asymmetry using a 2-layer network that learns via error minimization (Widrow & Hoff, 1960). The results demonstrated the need to separate learning into separate stages where first-order relations (e.g. X ↔ membership in category 1 , or Y ↔ membership in category 2) are learned separately in order to prevent learning of potentially maladaptive higher-order dependencies (e.g. XY), just like in the RNN.

11.2.7 Concept Combination in Adjective-Noun Phrases

The staged learning strategy proposed in Chapter 9, needed to satisfy the requirements of SPIN theory for learning atomic lexical semantic representations, essentially separates learning of co-occurrence relations among nouns and upcoming words, and pre-nominals and upcoming words. This is needed to prevent entanglement of pre-nominal with nominal distributional statistics. Interestingly, the language acquisition literature suggests these two learning stages are naturally separated in children: It appears that children do not readily integrate meanings of adjectives and nouns in adjective-noun phrases until they have gained basic competency with each separately, and especially nouns. For instance, comprehension studies show that preschoolers are not yet capable of reliably integrating adjective-noun combinations during online processing, and instead over-rely on information from just one of these constituents (Ninio, 2004; Thorpe et al., 2006). For instance, preschoolers, rely on just the noun when both adjective and noun information is required for disambiguation, or just the adjective when noun information is not required. Moreover, preschoolers are especially drawn towards an exclusively noun-based interpretation of adjective noun phrases. Russian speaking 6-year-olds, for instance, did not use information from the pre-nominal adjective to narrow down their fixation on a set of nine referents differing in color after hearing the adjective-noun phrase '*red butterfly*' (Sekerina & Trueswell, 2012). These studies suggest, albeit with some speculation, that preschoolers do not process adjective-noun combinations in such a way that would result in entanglement of their semantic properties. This would distinguish preschooler learning dynamics from those observed in the RNN, where frequent adjective-noun combinations are encoded as a unit, thereby mixing semantic properties across representations of the adjective and noun. Interestingly, according to SPIN theory, such a processing strategy would be beneficial if children are processing distributional statistics incrementally as does the RNN; selectively ignoring the adjective when processing an adjective-noun combination would ensure that (distributional) semantic properties are concentrated at the representation of the noun, and not inherited by the adjective. Such a filter is precisely what is required to meet the counterbalancing constraint (Chapter 6), whereby predictive dependencies between adjectives and nouns are avoided — either by not being available in the input, or by ignoring them during processing.

A more recent study has shown that children's tendency to over-rely on the noun when processing adjective-noun phrases is due to their limited, still-developing, online processing abilities Davies et al., 2021: When presentation of stimuli was slowed, 3-year olds succeeded at integrating adjective and noun meanings in an online reference resolution task, whereas faster presentation prevented successful integration. This link between processing speed and lexical semantic development is further supported by a longitudinal study conducted by Fernald et al. (2006), in which faster processing speed at 25 months was associated with more rapid vocabulary learning across the second year. The authors suggested that infants who are faster at encoding speech can "attend more efficiently to subsequent information in the speech stream and thus develop a more robust network of lexical-semantic representations that can be accessed more reliably during comprehension" (Fernald et al., 2010). These findings suggest that it would be useful to purposefully limit processing in the RNN to reduce entanglement of semantic properties of adjacent words during early training. This direction would fit well with recent efforts incorporating inductive biases into neural network models that guide learning towards human-like solutions (D. Liu et al., 2021; Noelle & Zimdars, 1999; Shen, Lin, et al., 2018; Shen, Tan, et al., 2018).

One of the most striking contrasts between learning in the RNN language models and children is the strong reliance on multi-word integration by the former, and an emphasis on learning individual word meanings prior to integration of multiple word meanings by the latter. It is well known, for instance, that children's knowledge of noun meanings develops between 6 and 24 months (Bergelson & Swingley, 2012; Fernald et al., 1998), well before children start to integrate information from adjectives. It is not until their fourth year that children are able to efficiently and flexibly combine adjective and noun meanings in naturalistic contexts (Waxman & Klibanoff, 2000). Consistent with the proposal in Chapter 7, which suggests that pre-nominals (including adjectives) should be learned after the network has had a chance to learn co-occurrence relations between nouns and upcoming items, children 's learning of adjective meanings is protracted compared to nouns (Berman, 1988; Booth & Waxman, 2009; Gentner et al., 2001).

Many proposals have explained the longer path of acquisition: For instance, adjectives violate a potential word-learning bias, namely the whole-object assumption which states that a new label refers to a complete object (Markman, 1990). Additionally, adjectives occur with lower proportion in child-directed input (Sandhofer et al., 2000), and have greater semantic, syntactic and pragmatic variability. Another potential explanation, based on work herein and which partially overlaps with previous proposals, is that form-based lexical representations of adjectives may inherit the distributional semantic properties of nouns to which they are often bound in language. When an adjective modifies a noun, the set of plausible upcoming words is not only constrained by the semantic properties of the nouns, but also by the semantic properties of the adjective, and their (potential) interaction. When a distributional learning system like the RNN is used to learn lexical semantic representations, there is uncertainty about which dependencies are causally related to the noun, the adjective, or both. Because adjectives often restrict the meaning of nouns they modify, when the RNN encounters adjective-noun phrases, it may not learn much about the individual semantic properties of each individually. The meaning of *little*, for instance, is relative to the noun it modifies; *little* therefore does not have semantic properties useful for predicting upcoming words when there is no noun to provide such a yardstick. Similarly, the presence of a modifying adjective like *little* may constrain the set of words that a noun may co-occur with, and this masks a subset of the distributional semantic properties needed to arrive at a complete semantic representation of the noun in the RNN. This is another example where the distributional perspective can lend additional insight into developmental phenomena in language acquisition.

11.2.8 Serial Dependence in Human Visual Perception

One of the primary insights in this thesis is the strong inclination by the RNN to rely on — potentially category-irrelevant or causally unrelated — contextual information to minimize next-word prediction error. By so doing, credit assignment is spread across time steps, rather than concentrated on specific time steps corresponding to specific words. Whether or not this leakage of information across time steps is a useful property of the language processor, and whether this actually happens when people process language, is still debated. However, evidence for the presence of a similar kind of 'temporal smearing' of information in people comes from studies of human visual perception. Specifically, researchers in that field have identified a phenomenon known as ‘attractive serial dependence’, a bias whereby a current visually presented stimulus appears more similar to the previous one (Bliss et al., 2017; Fischer & Whitney, 2014; Liberman et al., 2014). What this bias demonstrates is that visual perception is not constructed by stitching together a series of static snapshots of the external world; rather, information at one time step influences how the next is processed, and perceived. Such contextual effects on visual perception would be straightforwardly accounted for by an RNN-like mechanism trained to classify time series of visual data. Because the output of the RNN at each time step is a function of accumulated information, it is likely that the RNN would judge the current input as more similar to previous inputs, like people. Recent work by Fornaciai and Park (2020) showed that attractive serial dependence can occur solely from memory interference, meaning that attractive biases are generated during memory-related processes, rather than perceptual processes. This finding is suggestive of an RNN-like ‘perceptual working memory’ responsible for integrating visual stimuli across time steps.

Chapter 12

Predictions and Implications for Language Acquisition Research

This chapter is concerned with implications of the findings presented in previous chapters for theories of language acquisition, and language teaching. Further, I discuss behavioral predictions that should hold true in children to the extent that an RNN-like statistical learning mechanism accurately describes aspects of the acquisition of form-based lexical semantic representations in children.

12.1 Implications

In the sections below, I discuss implications of my work for theories of language acquisition, the organization of the language system, and the roles that the input, caregiver, and limitations on processing have on how semantic knowledge is represented.

12.1.1 The Lexicon

In the language sciences, it is often claimed that people have access to a mental storehouse that contains their knowledge of all the words they are familiar with (E. V. Clark, 1995; R. Jackendoff & Jackendoff, 2002; Pustejovsky, 1998). This so-called lexicon is responsible for long-term memory of semantic and syntactic properties of lexical items, and the provider of content to the sequencing/processing system whose job is to combine lexical items into well-formed sentences. While this theoretical framework has many advantages — formal constraints and clear separation of functions among them — many questions concerning the acquisition of lexical knowledge remain. For instance, the recent successes of language models trained on naturalistic language corpora has challenged the necessity of a principled division of labor among systems responsible for storage of lexical knowledge and systems that use this knowledge during production or comprehension (T. Brown et al., 2020; Gulordava et al., 2018; Kuncoro et al., 2017; Sutskever et al., 2014). In these networks, the boundary between knowledge of individual words and knowledge of how to process word sequences is blurred due to the joint optimization of network parameters that store and process lexical items.

A primary goal of this thesis is to stimulate debate concerning whether a data-structure like the mental lexicon can emerge in a system that learns via next-word prediction. Scholars that defend the role of an RNN-like mechanism in acquisition have tended to stress the primacy of sequences over isolated occurrences of words in the construction of meaning (J. L. Elman, 2009, 2011). While such scholars do not deny that lexical semantic representations that capture isolated and full-featured word meanings may not readily emerge in the RNN, they argue that this is *not necessary* to account for human language abilities. The argument goes something like this: The disambiguation of word meanings, and especially the argument structure of verbs often depends on lexical context, such as the semantic properties of verb arguments. Context effects, at first blush, suggest the need for more sophisticated non-static mechanisms to lookup word meanings than the lexicon allows. That said, defenders of the lexicon have a simple way out of this dilemma: The conditions under which disambiguation occur are simply added to the lexical entry of the polysemous word (R. Jackendoff & Jackendoff, 2002; Pustejovsky & Boguraev, 1993). While this is a sensible approach, continuing to add contextual information to the lexicon would push the the lexicon beyond its original purpose as a storage of static, isolated word meanings. J. L. Elman (2009) argues that this blurring of lexical and processing knowledge needed to account for contextual effects on word meaning disambiguation dissolves the principled division of labor between lexical and processing knowledge for which the lexicon was originally envisioned. By continuing to add such conditions to the lexicon, it would eventually converge onto the responsibilities traditionally assigned to the processor. What J. L. Elman (2009) suggests is that we are better off not chasing a target whose premise is undermined by attempts to rescue it from alternative approaches.

While I agree with J. L. Elman (2009) on many points, the premise of this thesis is that there are some tasks that *do* seem to require knowledge of isolated word meanings, and that in order to succeed on these tasks, people must store word meanings in a format that make them readily accessible — akin to static entries in a lexicon. One of these tasks, as discussed in Chapter 1, is the distributionally-mediated extension of semantic category-associated features (DECAF). This task is essential for inducing approximate meanings of novel words in the absence of extra-linguistic perceptual information about word meaning (e.g. the referent is absent or does not straightforwardly manifest in perception). Because word learning tends to proceed one word at a time, this task appears to require accumulation of statistical knowledge about individual words rather than the chunks in which they often appear. To be clear, I am not suggesting that all words require static lexical entries nor that every language task draws on context-independent knowledge of word meanings. What I am suggesting is that distributionally constructed lexical semantic representations would benefit by being atomic to support children's inferences about the meaning of novel words. The weakest version of this argument is that (i) lexical semantic representations vary in their degree of atomicity, (ii) that the degree of atomicity is influenced by the statistics of the language that children hear, and that (iii) children are better off emphasizing those representations and those statistics that yield the most atomic lexical semantic representations, so that they can (iv) make the the most out of situations in which a novel word is presented in the absence of perceptual information about the word's meaning.

What then, does the work herein, contribute to the debate surrounding the presence or absence of a dictionary-like lexicon? This question is difficult to answer, because there is much disagreement about whether children learn isolated word meanings, or prioritize meaning at the level of larger chunks such as phrases and sentences, or both (Goldwater et al., 2009; McCauley & Christiansen, 2019; Ramscar et al., 2013b). An unsatisfying answer is that the RNN language model gives rise to a complex mixture of knowledge about the distributional semantic properties of isolated *and* contextualized words. In the RNN, the degree of atomicity depends almost entirely on the ways in which words are used in the language it has been exposed to. Some words are used with highly predictive contexts, while others occur in highly variable contexts, and the balance between the two contributes to the degree to which a word is encoded atomically. Interestingly, the findings herein show that the RNN has no detectable bias for atomicity — if anything, it is biased against atomicity and towards capturing chunk-level statistics in the form of compound cues. This raises interesting questions about whether a bias for atomicity can emerge at all given the highly context-sensitive nature of natural language sentences, or whether such a bias should be baked into the architecture before the system has had a chance to be molded by language experience?

The above question rests on a deeper question, namely when is atomicity most useful? It is possible that lexical atomicity is a temporary heuristic that is most useful during early language learning and is discarded as soon as children can rely on more sophisticated experiential knowledge to infer novel word meanings. Alternatively, atomicity may continue to be important for performing various language tasks beyond early acquisition, such as compositional semantic inference. An altogether different view would be that lexical atomicity does not start to become useful until much later. On this view, atomicity would not be something that is built into a system, but something that is nurtured — possibly with structured language instruction from caregivers and/or teachers. It is not out of the question that children do not begin the process of language acquisition expecting to extract isolated word meanings, but instead, extract whatever meanings they can, using chunks of language of whatever size they are able to identify. Some scholars have even suggested that children segment the speech stream into progressively smaller chunks, all the while storing larger chunks in their developing lexicon (i.e. 'chunkatory') (McCauley & Christiansen, 2019). This would fit well with the idea that atomicity is not the primary goal of lexical semantic development, and that on the path to atomicity, children first learn to represent the meanings of larger chunks of language, and only later learn to decompose chunks into smaller atoms of meaning.[1]

12.1.2 Ambi-categorical Words

To illustrate the dilemma that simultaneously learning lexical and sequential structure poses for theories of acquisition, consider the word *back*. This word is interesting because it is ambi-categorical, meaning that it can be used in more than one part-of-speech — *'Turn so I can see your back'* (noun) vs. *'Give that back'* (adverb). Given my conceptualization of lexical semantic representations in the RNN (i.e. the input weights aka. embedding layer), each word in the vocabulary is associated with only a single lexical representation, regardless of whether the word can be used cross-categorically. This does not necessarily impede the ability of the model to distinguish part-of-speech (e.g. is *back* used as a noun or adverb?), because the sentential context typically provides disambiguating information. In fact, the RNN trivially handles ambi-categoricality by learning lexical representations that can push the hidden state in one or another direction, depending on the state of the hidden layer one time step prior. This is one of the major advantages of contextualized processing, and why the RNN is considered by many scholars important to theorizing about language processing (J. L. Elman, 2011; J. L. Elman & McRae, 2019). Given a sentential context that favors a noun-reading, the lexical representation may push the RNN into a state indicating that this is the case. Similarly, the same vector may push the state to a region indicating an adverb-reading if the sentential context favors such a reading.

The potential for multi-purpose application of the same word vector essentially solves the ambi-categoricality problem in the RNN, but has potentially undesirable consequences on the organization of the lexicon — the semantic vector space at the input layer. Here, vectors serving multiple purposes, depending on the context in which they are used, do not conform to the same constraints as other mono-categorical words, and this

[1] Note, that this idea is not compatible with the hypothesis presented in Chapter 9, that atomicity is best induced early during training.

negatively impacts atomicity. If learned categories of words are based on their distribution relative to other words, then ambi-categorical words could cause a distributional learner such as the RNN to conflate category distributions and create semantic categories that contain mixtures of part-speech classes. This is potentially problematic. For example, if trained on a corpus where the word *back* is used as both a noun and an adverb, the vector for *back* will not neatly cluster with semantically similar words in the noun category, such as as *body* or *belly*, but instead be skewed in a different direction so as to fulfill its dual function. What we are left with is a system that can either be used to handle ambi-categorically and other contextual language phenomena, or be constrained to acquire atomic lexical semantic representations, but not both. I think this trade-off relationship between sensitivity and insensitivity to context is not a limitation of the RNN, but a fundamental limit on what is possible in any system (including children) that attempts to extract isolated lexical statistics embedded in the complex sequential structure of language.

Emerging evidence suggests that children do not encode ambi-categorical words in a single representation, but rather learn separate representations, one for each category (Conwell & Morgan, 2012), and that vowel quality and prosody are important cues for distinguishing their usage. This strategy preserves atomicity, but is not in line with the modeling work in this thesis where each orthographically identical token is treated as an instance of the same type, with exactly one lexical representation. To account for the behavioral data, the simple RNN would need to be extended with access to the raw speech signal or acoustic features for differentiating orthographically identical tokens. However, in this scenario, many more questions arise. For instance, what mechanism, if any, would be responsible for segmenting the input into discrete vocabulary items? If the units over which the RNN operates are not at the lexical level, training may not give rise to clearly distinguishable lexical representations, making the goal of lexical atomicity even more difficult to attain. If, on the other hand, the vocabulary were specified by the researcher ahead of time, such that different uses of ambi-categorical words were assigned separate vocabulary entries, this begs the question how such an assignment is learned in the first place.

It should be noted that cross-category word use is only one instance of a more general phenomenon, namely polysemy. When a word is polysemous, it has many possible meanings, which are contextually determined. Polysemy is far from the exception to the rule: It has been estimated that nearly half of all words in the English and Chinese language have at least dual meaning (Lin & Ahrens, 2005). Polysemy is therefore a consideration that must be taken seriously by models of lexical semantic development.

12.1.3 Functional Specialization

The demonstrations presented herein reveal a fundamental trade-off between sequence-level and word-level learning dynamics; while greater utilisation of contextual information is useful for modeling the probability distribution over possible sequences, individual lexical items become increasingly reliant on information provided elsewhere in the system and, in turn, are rendered less useful as input to downstream lexical tasks such as DECAF. Does this mean we are better off using different models that specialize in tasks that require either sequence-level meaning or lexical meanings but not both? The findings and idea presented herein would certainly support the idea of such functional specialization. Moreover, this idea would be in agreement with mounting psycholinguistic evidence that (i) children and adults store both individual words *and* multi-word sequences (Abbot-Smith & Tomasello, 2006; Ambridge et al., 2015; Arnon & Clark, 2011; Bannard & Matthews, 2008; C. L. Jacobs et al., 2016; Rubino & Pine, 1998), and the presence of functional differences related to language processing across the two brain hemispheres (Federmeier, 2007; W. W. Graves et al., 2010). While I remain uncommitted to a conclusive answer, the evidence presented herein does suggest that additional, and potentially more specialized machinery than the RNN may be needed to get the most out of DECAF.

To fully answer this question, we must first unravel a deeper question, namely what component of the language system is the RNN a model of? In other words, where does the RNN fit within the broader picture of the human language system? In the psycholinguistics literature, the RNN has been used to model an incredibly diverse range of phenomena (Brouwer et al., 2017; F. Chang et al., 2006; Christiansen et al., 1998; J. L. Elman, 1990; J. L. Elman & McRae, 2019; Linzen & Baroni, 2021; Magnuson et al., 2020; Misyak et al., 2010; D. L. T. Rohde, 2002). But how are theorists supposed to integrate this individual demonstrations? Is the RNN best understood as a monolith, that is simultaneously responsible for multiple tasks, or is the human language system better characterized by a collection of RNN-like mechanisms dedicated to individual tasks? One of the reasons I have chosen the RNN to model the supply of lexical representations to a broader system for performing DECAF is in part the model's ability to account for many language abilities. This raises many questions: Is the RNN merely a supplementary system that supports word learning and processing indirectly, or does it play a more direct role in online language processing, lexical semantic development, grammatical development, or all of the above? If the RNN were to occupy the role of an online language processor, would it also be responsible for learning? It seems the two go hand in hand — without learning, the RNN may be of little use as a processor. Another important question is how much an RNN-like mechanism would interface with the extra-linguistic conceptual and experiential domain? For example, would the RNN be supported by a separate system for acquiring and representing semantic categories and relations or would this knowledge be kept separate?

An in-depth discussion of some of these questions in the context of hemispherical specialization can be found in Federmeier (2007) who proposed that the left hemisphere specializes in top-down predictive processing, whereas the right hemisphere specializes in bottom-up integration of lexical knowledge. On this view, it is plausible that the RNN language model is best viewed as a specialized component of the human language system — akin to language systems in the left hemisphere — and that additional components, facilitating integration of lexical semantic knowledge, are needed to account for the full range and complexity of language. Further evidence of functional specialization in the human language system comes from neuroimaging work in which systems responsible for lexical versus combinatorial processing were dissociated. For instance, W. W. Graves et al. (2010) found evidence for hemisphere-level differences in activation in response to lexical vs. processing knowledge. More precisely, the authors found that right hemisphere tempo-parietal structures were recruited for combining individual noun concepts into a whole. In addition, prior to meaning combination, the authors found evidence for left hemisphere temporal activation involved in analysis of lexical meanings. In sum, there is considerable evidence that lexical semantic representations are stored separately from the language processor in humans. The RNN may be better suited as a model of the latter, with a separate system responsible for the former.

The question about functional specialization is intimately related to the debate regarding the role of word-order in the acquisition of form-based lexical semantic knowledge (M. Jones & Recchia, 2010; M. N. Jones & Mewhort, 2007). On the one hand, it is possible that children entirely discard information about word-order and instead track relations between words independently of their syntactic or linear distance in a sentence. On this view, semantic categories may be represented as distributions over bag-of-words, after having discarded information about word-order word-distance. Alternatively, word-order and word-distance might play a critical role, as such cues may reliably correlate with meanings in the natural world. A review of the literature suggests that word-order is helpful for some tasks, but not for others. For instance, P. A.

Huebner and Willits (2018) showed that Word2Vec, which does not explicitly use word-order information to learn lexical semantic representations, performs slightly better in a semantic categorization task compared to the RNN after both were trained on AO-CHILDES. On the flipside, the word pairs that Word2Vec considers similar are typically a mixture of syntagmatically and paradigmatically related word pairs, whereas the RNN is heavily skewed towards learning the latter. This result illustrates a fundamental trade-off: By removing explicit information about word-order, lexical semantic information is potentially enriched, but at the cost of blending paradigmatic and syntagmatic similarity, a distinction which is crucial to enable semantic feature extension, as discussed in Chapter 1 and previously proposed by M. Jones and Recchia (2010).

12.1.4 Lexical Access

Another implication of the current work is in research on lexical access. It has long been argued that the strength with which a word is encoded in semantic memory is a function of how often it was encountered. There is lots of evidence to back this up: High frequency words show an advantage in perceptual and production tasks, such as shorter response latencies and higher accuracy in tests of perceptual identification (Morton, 1969), word naming (Forster & Chambers, 1973), and lexical decision (Scarborough et al., 1977). The usual explanation is that with each new occurrence of a word, its lexical representation is strengthened, yielding faster lexical access time. However, it is possible that differences at retrieval-time do not reflect differences in the *strength* with which a word was encoded, but a difference in *how* a word was encoded (e.g. atomic vs. non-atomic). The RNN simulations presented in this thesis show that the semantic properties of a word may not be concentrated at a single location in semantic memory, but spread more diffusely across other parts of the network. As a consequence, it is likely that it would take longer to activate the semantic properties of a word that was encoded more atomically as opposed to a word that was encoded less atomically — without resorting to differences in the strength of encoding.

This explanation is in line with previous behavioral findings and modeling work. For instance, M. N. Jones et al. (2012) have proposed that variability in lexical access (as measured by reaction time in visual or spoken word recognition tasks) can be accounted for by a measure of a word's contextual diversity that takes into account both the number of contexts and their semantic overlap. Their measure, termed 'semantic distinctiveness count' predicts that lexical access is faster when the word has occurred in multiple semantically distinct contexts more highly than in more redundant contexts. In a series of experiments, the authors showed that it consistently provided a better fit to word recognition reaction times, and is a better predictor than pure word frequency. The idea behind the claim of M. N. Jones et al. (2012) is that words are less strongly encoded when encountered in redundant contexts, and these encoding-time differences manifest at retrieval-time.

12.1.5 Moving Beyond Context Diversity

A strong undercurrent of this thesis is the need for statistical analyses of language that consider the multivariate structure underlying co-occurrence patterns. Because the organization of learned knowledge in connectionist systems like the RNN is strongly influenced by the overall structure of the input (Saxe et al., 2019), measures related to, say, the average diversity of individual context words, are not sufficient to predict what a model will learn. The average context diversity of nouns in a corpus says nothing, for instance, about how nouns systematically relate to each other, nor how coherent the noun category is as a whole. Specifically, context diversity itself does not shed light on whether nouns tend to occur in more similar contexts. Both higher and lower context diversity are equally compatible with a more coherent — or less coherent —- noun category. What context diversity tells us is instead, is how long it might take a distributional learner to discover the category — the more diverse the contexts in which nouns occur, the longer it might take to integrate contexts to construct a single cluster. Instead, if the goal is to characterize the coherence of a category, a multivariate analysis is needed, such as the procedure used to calculate 'fragmentation' proposed in Chapter 7.

The need for more sophisticated statistical analyses is related to the importance of analyzing statistical associations from an information-theoretic perspective. For instance, context diversity alone does not address the question of how useful a context is for predicting a target word. Context diversity tells us little about predictive strength, because it doesn't take into account how often a context occurs with other target words, or how contexts tend to pattern with each other across multiple targets. Reliance on simple lexical statistics that either do not consider multivariate patterns, or informativity, can result in misleading conclusions. For instance, M. N. Jones et al. (2012) say the following about learning lexical representations in the RNN:

> Words that are low in contextual variability will be better supported by consistent contextual cues, and thus should be weighted less strongly in memory, since they will be more predictable in context. Conversely, words that are high in contextual variability should be represented more strongly in the lexicon, since they are less associated with any given context, and thus lack contextual scaffolding.

While this is usually a reasonable starting assumption, the authors do not talk about how differences in the informativity of context words might turn this statement on its head: High context variability *does* reduce reliance on contextual scaffolding when encoding a target word, *but only* when its contexts are also shared by other members of the grammatical category the target word belongs to (i.e. high atomicity). If the contexts in which the target word occurs are unique to the target word, than those contexts become redundant with the target word, and potentially suffer from 'semantic property inheritance' (i.e. low atomicity) discussed in Chapter 6. Thus, context diversity alone cannot be used to fully characterize network behavior (and certainly not the degree of lexical atomicity).

It should be noted that similar concerns have been raised by McClelland and Rogers (2003) who suggested that tracking only pairwise correlations between features and objects is not sufficient to account for children's early knowledge of the similarity between objects of the same kind. Instead, they propose that, in order to account for children's initial grouping of objects into broad conceptual distinction, statistical learning systems would benefit by tracking the coherent co-variation between features across large collections of observed objects. Their argument is backed up by extensive simulation work which shows that broad conceptual distinctions akin to those of pre-verbal infants emerge automatically in a connectionist architecture that is sensitive to coherent co-variation between features and objects — and without prior knowledge about which features are more relevant or perceptually salient for object categorization.

12.1.6 Two-Process Language Learning

Much of this work herein assumes that the semantic categorization step needed to perform DECAF is a task that is performed by directly accessing static lexical semantic representations. However, not all scholars agree that lexical semantic categorization — or any other semantic task for that matter — draws exclusively from knowledge stored in a static lexicon. Instead, categorization can be understood as a task that people learn to perform and which goes beyond static word knowledge, rather than a task that can be performed as soon as lexical semantic representations have become available in the lexicon. In other words, semantic categorization is not necessarily the exclusive burden of the lexicon, or, by extension the language model which may be partly responsible for producing those representations.

An alternative account, provided by P. Huebner and Willits (2019), is known as ‘two-stage’ or ‘two-process’ lexical semantic development. On this view, lexical semantic development is a hybrid of two distinct processes. The first process involves unsupervised associative learning, with the goal of acquiring representations of words that are useful for a wide range of goals. The second process involves explicitly learning to classify word-pairs as instances of one of several types of linguistic relations. Critically, this second process uses the representations of the first process as its starting point. The authors demonstrated this using a two-stage model, where the first stage is a distributional semantic model (e.g. HAL, Word2Vec, RNN), and the second stage is a transformation of previously learned representations into task-specific target spaces. In other words, the passive statistical knowledge acquired via self-supervised distributional learning is not directly available without the support of a supervised stage that potentially modifies how learned knowledge is accessed and/or transformed during a specific task. P. Huebner and Willits (2019) showed that a computational instantiation of this approach improves performance on multiple semantic tasks compared to using the representations learned by stage 1 directly. From this, the authors concluded that a task- or goal-oriented perspective holds promise for furthering our understanding of lexical semantic development.

What relates the two-process account of P. Huebner and Willits (2019) to work in this thesis is that implicit in their approach is the need to overcome a specific theoretical limitation of existing models of semantic memory, namely the assumption that a concept is represented as a single general-purpose vector. The authors share the intuition that capturing lexical statistics in language models such as the RNN is not the end-goal of acquisition, but merely a starting point for seeding representations by more task-specific specialized systems. This idea is also more broadly aligned with recent trends in computational linguistics and machine learning where pre-trained lexical representations are transferred from a general purpose distributional learning system to a more task-specific system such as a classifier. Rather than viewing this method simply as a research tool, P. Huebner and Willits (2019) suggested this approach holds promise for extending models of semantic representation to account for effects of task-specific influences on human semantic behavior.

At the extreme end of this continuum, J. L. Elman, 2009 goes so far as to argue that the entire notion of a static lexicon is misguided. With regards to semantic categorization, J. L. Elman, 2009 would likely claim that the information needed to perform this task is not necessarily encoded in the structure of the representational system where isolated words live, but in the connection weights of the system that has been trained to perform the task. Importantly, J. L. Elman, 2009 does not argue that people don’t know things about words; just that the notion of context-insensitive representations of words are neither necessary nor desired. Under this framing, semantic categorization is something that people learn how to do, not a fundamental property of the representations themselves.

12.1.7 Progressive Differentiation

One as of yet unanswered question in language acquisition research is whether learners start with ‘coarse grain’ representations of lexical categories and refine them as needed, or whether they start with narrow category representations and eventually collapse words into broader categories once sufficient evidence in favor has accumulated. An unspoken assumption at the foundation of this thesis is that category learning in the RNN is proceeds from coarse-grain → fine-grain, starting, for instance at the superordinate level (all nouns), and ending with fine-grained semantic distinctions between nouns. Whether the same learning trajectory applies to children is still widely debated among acquisition researchers. In this section, I discuss some work on this front. In particular, some of this work argues that aspects of children's semantic development can in part be explained by neural network learning dynamics trained incrementally on staged language input.

A number of factors have led researchers to propose that semantic memory is organized taxonomically, such as semantic organization in recall from memory (Bower, 1970), category-aligned systematic decline in memory performance in patients with semantic dementia (Warrington, 1975), and patterns of inductive inference about semantic features and category membership, both in development (E. V. Clark et al., 1985; F. C. Keil, 1981) and adults (Collins & Quillian, 1969; Rips, 1975). These and other findings have led cognitive scientists to propose a theory of concept acquisition known as progressive differentiation (F. C. Keil, 1981). According to progressive differentiation, superordinate concept categories (e.g. FISH vs MAMMAL) are acquired before finer-grained conceptual distinctions (e.g. trout vs salmon). This idea received much attention in the work by Rogers and McClelland (2004), who demonstrated that a feed-forward neural network first learns to differentiate between concept categories with the least feature overlap, before distinguishing concepts with greater feature overlap. This notion was formalized by Saxe et al. (2019) who plotted the learning trajectories of linear feed-forward networks obtained from closed-form solutions of their learning dynamics. Briefly, Saxe et al. (2019) showed that the learning dynamics of the network used by Rogers and McClelland (2004) is equivalent to progressively encoding singular dimensions that account for increasingly less variance in the mapping from input to output. Importantly, those dimensions that account for the most variance are learned faster than dimensions accounting for less variance. The authors also showed that progressive differentiation can account for many phenomena from the child language acquisition literature, such as U-shaped learning and periods of over-regularization.

The current thesis relates to this line of inquiry because it shows that language data itself - and not just the learning strategy employed by children or neural networks - may be organized to facilitate progressive differentiation. It is possible that as of yet unknown but beneficial consequences would result from combining learners that operate via progressive differentiation with input that is itself scaffolded in a way that supports progressive differentiation. Because the distributional patterns of nouns provide conflicting information about their membership in the noun category vs. membership in potentially numerous smaller sub-categories, input that progressively reveals the existence of finer-grained categories can help a learner discover the complex subcategory structure in a step-by-step fashion. An incremental learner exposed to this kind of input could partially avoid the interference between super- and subordinate category signals by discovering the statistics underlying each category at a time. On this view, SPIN theory predicts that in order to avoid such conflict, infants should acquire broader semantic distinctions, such as animate vs. inanimate nouns before acquiring finer-grained distinctions, such as nouns that refer to birds vs. nouns that refer to insects.

12.1.8 Optimizing Language Teaching

The question of what role the input plays in language acquisition is hotly debated. Clearly some input is needed, but how much and how do qualitative differences like type/token ratio, or more subtle statistical properties (e.g. compounding, noun modification, etc.) affect learning? There are numerous studies that highlight the role that individual variation in language exposure has on vocabulary growth (Hart & Risley, 1995; Huttenlocher et al., 1991; Rowe, 2012), which indicates that the quantity and quality of the language input can have a strong influence on learning outcomes. This kind of research has long been plagued by failures of replication, as well as difficult confounds like genetic similarity between caregivers producing the language input and children at the receiving end. However, recent findings show that input effects on language outcomes persist even in the absence of genetic confounds (Coffey et al., 2022). The presence of the age-order effect in the RNN further lends credibility to this line of work from a computational perspective. More importantly, the age-order effect suggest that it is not just the quantity of input that can influence learning, but also the statistical associations between words in the surface structure of language (e.g. redundancy, fragmentation).

One of the earliest work on the possible relationship between children's input and language learning outcomes was conducted by R. Brown (2013). His recommendation to concerned caregivers was as follows (p. 26):

> Believe that your child can understand more than he or she can say, and seek, above all, to communicate. To understand and be understood. To keep your minds fixed on the same target ... There is no set of rules of how to talk to a child that can approach what you unconsciously know. If you concentrate on communicating, everything else will follow.

The contemporary literature on this issue has changed little (Gallaway & Richards, 1994). There are no 'quick and dirty' methods to speed children's acquisition, beyond increasing and diversifying a child's language input. This is not controversial; what is hotly debated is whether there are qualitative aspects of language that are more favorable for children's language growth, and whether such qualitative aspects are specific to developmental periods. One proposal, the 'less is more' hypothesis, suggests that presentation of language material in simplified contexts (e.g. shorter sentences) may speed aspects of language learning such as syntactic and morphological development (J. L. Elman, 1993; Kersten & Earles, 2001; Newport, 1990). However, not all researchers agree that forcibly simplified input is a sensible approach to language teaching (Rafferty & Griffiths, 2010; D. L. Rohde & Plaut, 1999). More precisely, it has been suggested that over-extending the time in which simplified language input is provided may actually delay acquisition. For example, learners provided exclusively with simplified input may struggle to learn more complicated constructions in their language (Rafferty & Griffiths, 2010). On the other hand, providing more advanced examples of language from the beginning may delay learning due to the large number of hypotheses that would have to be considered (Cameron-Faulkner et al., 2003).

This presents a conundrum to the language teacher: How should samples from a language be chosen to maximize learning outcomes? I argue that in order to answer this question, a better understanding of how input to children changes across developmental time is needed. The findings reported in this thesis demonstrate that language to children is developmentally organized in a way that facilitates acquisition of atomic lexical semantic representations of nouns; this is likely not a co-incidence, but, instead, may indicate a deep relationship between children's input and the machinery by which the input is encoded. It is plausible that children's distributional learning system is initially tuned to particular aspects of the input, right from the start, in order to constrain their initial space of hypotheses. More speculatively it is possible that such a system might also adjust its tuning synchronously with changes taking place in the learning environment across developmental time. While historically, research has focused on how input quantity and quality at one point in development influence learning outcomes at a future point, a potential new direction is to examine how the learner continuously adjusts to the input that is provided, and potential interactions between stimulus history, learner, and how novel input is analyzed.

It is important to distinguish the age-order effect — an effect of stimulus history on lexical semantic development — from previous longitudinal investigations of children's language learning outcomes. Historically, the majority of such studies have investigated potential links between input and language abilities (e.g. word recognition, vocabulary size, syntactic knowledge), but not how input might guide or constrain children's lexical semantic knowledge. This is not surprising given that, historically, effects of language input on semantic development have been largely (and, perhaps, purposefully) ignored. This attitude towards language effects on semantic development is well captured by Newport et al. (1977) who wrote:

> The conditions for the acquisition of cognitive-semantic relations and those for the acquisition of grammatical functions which elaborate those relations seem different. We have nothing to say about how the child learns or develops with regard to the semantics of natural language. But we can say that, however this is done, it is accomplished with less close reliance on environmental linguistic support than is the acquisition of some properties of surface syntax. It is in this latter dimension that we find dramatic effects of an environmentally-dependent acquisition process.

In more contemporary research, the study of input effects on semantic development has been more fully embraced. For instance, Gelderloos et al. (2020) examined the extraction of semantic information directly from child-directed vs. adult-directed language, and how linguistic and acoustic factors might make this easier. To do so, the authors compared computational models of semantic development trained either on adult-directed speech or child-directed speech. Gelderloos et al. (2020) found that, while child-directed speech helps in the initial stages of learning, models trained on adult-directed speech eventually reach comparable task performance, and generalize better. Interestingly, improved generalization appeared to be due to differences in linguistic rather than acoustic properties between child-directed and adult-directed input, given that the same results were replicated when models were trained on acoustically comparable synthetic speech. What this observation has in common with ideas presented in Chapter 9 and findings presented in Chapter 10 is that presentation of child-directed input by itself does not aid semantic development in the general case; rather, the benefit of child-directed input appears to be the scaffolding provided at the earliest stage of learning. Once, such scaffolding is in place, the findings by (Rafferty & Griffiths, 2010) and Gelderloos et al. (2020) indicate that introduction of more representative (i.e. adult-directed) input better supports semantic development than continued exposure to less representative (i.e. child-directed) input. This is in accord with the 'Zone of Proximal Development' and the related 'Goldilocks principle': Learning is optimal when the task that is presented at each stage in development is continuously updated so that it is not too challenging nor too easy for the learner (Bruner, 1984; Hill et al., 2015; Wilson et al., 2019).

Another line of work has examined how word-order influences word learning. These studies make close contact with SPIN theory, and the finding that the representations learned by the RNN are heavily influenced by whether an item in the input predicts another item, or is predicted by it. For instance, Dye et al. (2017) examined whether word-order asymmetries in lexical distributional statistics might influence how people interpret the meanings of familiar or novel target words during reading. The critical manipulation is whether a word that is semantically similar to a target word is presented in the left or right context in text read by participants. Interestingly, participants rated the similarity between target words and words occurring in the left context higher compared to words occurring in the right context. This result confirms that the distributional semantic representations learned by people are influenced by the order in which word pairs were presented in sentences, and are therefore in accord with the idea that lexical semantic development and sequential language processing are linked — potentially via a common prediction-based mechanism. Together, these results suggest that people are sensitive to word-order when learning distributional statistics of language, and that word-order differences influence how distributional information is represented, and how quickly it is learned. Therefore, word-order should be more closely examined by future research that investigates language teaching.

12.1.9 Input vs. Intake

A good model of how people acquire semantic category knowledge must be able to distinguish different kinds of contextual signals available in the learning environment. While some signals provide information about the features that are diagnostic of a target category, other information is purely incidental, and may not be causally related to the target category. Over the course of development, people become tuned to these distinctions; stimulus dimensions that are relevant to a target category are selectively attended, while irrelevant dimensions are ignored or down-weighted (Gentner, 1988; Goldstone, 1994; Nosofsky, 1986). Some scholars have argued that in the absence of a theory of how natural stimuli are causally related, the task of learning natural categories in a bottom-up fashion from observation of perceptual similarity and/or co-occurrence alone is implausible, or at best heavily under-constrained (Goodman, 1972). Language learners are faced with an identical problem: The acquisition of knowledge of semantic category membership from co-occurrence data alone is difficult, in part due to the large amount of co-occurrence relations in natural language data, and the immense number of interactions between lexical items.

What all this suggests is that children potentially filter linguistic input they hear. In many circles this is known as the distinction between 'input' and 'intake': While a child may hear X and Y, the child may listen preferentially to X, if, for instance, he or she cannot process Y, or if Y does not capture their attention. A classic example is the finding that children primarily listen to high-pitched speech, speech which begins with the child's name, or is accompanied by eye contact or other gestures. These findings suggest not to draw strong conclusions about the role that caregiver speech plays in teaching language; it is equally (or potentially more) likely that it is the child that brings certain predispositions to bear on acquisition, such as by filtering the input prior to intake. On this view, caregiver speech is facilitative only insofar as it is able to make it past such filters. For instance, some have suggested that presentation of shorter utterances during infancy might scaffold aspects of acquisition; however, the input-intake dichotomy would suggest that it is the filter on infants' intake of longer sentences that might actually be doing the work. Considerable evidence has mounted in favor of maturational constraints on what input infants and children actually take in. Much of this evidence has focused on how children's processing of language differs from adults (Griffin et al., 1994; Hudson Kam & Newport, 2005; Ramscar & Gitcho, 2007). It is likely that the simple RNN would also benefit from similar filters on its input, which might help to reduce the entanglement of category-relevant and category-irrelevant statistical dependencies.

12.1.10 Caregiver Input and Semantic Development

Classic studies of the effects of variation in child-directed input produced by mothers ('motherese') on language growth have focused on acquisition from the point of view of grammatical development, whilst not paying attention to semantic development. Such studies showed that many factors related to syntactic complexity of children's language input do not predict subsequent syntactic competence. For instance, whether mothers use short or long sentences, wide-ranging or restricted sentence types, complex or simple sentences, do not have a noticeable effect on language growth on children. On the basis of these findings, many scholars have concluded that such linguistic adaptions may be the result of social and pragmatic pressures on caregiver-child interactions, rather than serving pedagogic purposes. Interestingly, only few have examined the potential impact of syntactic complexity on semantic variables, like semantic category knowledge. Many scholars did not do so because they did not think this possible, or important. However, the findings in this thesis suggest it is worth re-examining such null-effects in the context of lexical semantic development; it is quite possible that age-related changes in language-internal factors like MLU and lexical diversity support the acquisition of lexical semantic knowledge, as is the case in the RNN simulations presented in chapter 10. If so, this would help explain why some aspects of caregiver input that significantly depart from adult-directed language do not predict learning outcomes — previous scholar may not have been looking in the right place.

12.1.11 The Semantics-Syntax Interface

The possibility of syntactic factors scaffolding semantic development is tantalizing, but not widely adopted. Many scholars, especially those with a formal linguistic bent, often consider semantic and syntactic processing due to separate systems, with a clear, principled demarcation (R. Jackendoff & Jackendoff, 2002). However, if we are to take seriously the possibility that children acquire lexical semantic representations, in part, by tracking distributional information available in their input, such a principled distinction may not be upheld in all parts of the language system. If an RNN-like mechanism were to support word learning in children via the distributionally-mediated extension of category-associated features (DECAF), such a process would allow for arbitrary interactions between what is traditionally considered the domain of semantics and syntax. In fact, the age-order effect (Chapter 10) cannot be explained by a system that performs purely syntactic operations, nor one that performs purely semantic operations: The greater discoverability of nouns in speech to younger children is an observation of syntactic structure, and, yet, this *syntactic* phenomenon scaffolds — and therefore interacts with — the learning of *semantic* structure. The possibility for such an interaction has historically been neglected, and the work herein suggests that paying attention to incremental progress made during development has the potential to shed light on the interface between syntactic and semantic systems in acquisition.

12.2 Predictions

One of the most fascinating aspects of working with computational models is that they potentially offer novel predictions that can be tested by researchers. Based on the findings in this work, and in particular, those that have culminated in SPIN theory, and the age-order effect, I have derived several testable predictions about the role that input-related factors play in scaffolding children's lexical semantic development. If an RNN-like mechanism is indeed involved in children's acquisition of word meanings (DECAF), the predictions listed below provide a starting point for researchers interested in empirical validation.

Many previous studies on lexical semantic development have found evidence for the supporting role of conversational factors such as deixis, joint attention and shared gaze (Yu & Smith, 2012), on lexical semantic development (E. V. Clark & Garnica, 1974). However, comparatively little is known about the potential effect of language-internal factors, such as the lexical contexts in which nouns are heard. In particular, the findings in this thesis suggests it is worth examining whether language-internal factors such as combinatorial diversity (e.g. nominal modification, noun phrase complexity) and fragmentation can account for additional variance in lexical semantic development beyond conversational factors already known to researchers. If SPIN theory applies to distributional learning in children as much as it applies to learning in the RNN, it predicts that such correlations should be found. To the extent that these predictions turn out to be accurate, SPIN theory has the potential to generate novel recommendations for how to talk to children during the earliest stage of lexical semantic development.

12.2.1 Pre-Nominal Semantics

There is an irony at the heart of SPIN theory: It suggests that in order to learn more atomic lexical semantic representations, target words must not be preceded by semantically informative left contexts. While avoidance of redundant semantic information during early training (via age-ordered training) is in fact beneficial for learning more atomic lexical representations for nouns in the RNN, it remains to be seen whether a similar intervention also benefits noun learning in children. All else being, equal, I predict that this would be the case. Such an intervention would seem to be most effective during the first two years of life, but no later, because the strategic withholding of pre-nominal adjectives could have negative consequences on learning adjectives and how to integrate adjectives and nouns for interpreting complex noun phrases. It is likely that an RNN-like distributional learning mechanism is most active during the first two years of life, when children do not yet possess rich knowledge about the world that would otherwise support semantic category-based induction of novel noun meanings. At this early stage, distributional evidence may be the most readily available source of information, so that interventions designed to boost atomicity of form-based lexical representations can be prioritized with the least negative impact on children's ability to integrate semantic information across adjective-noun combinations.

Such an intervention would require careful consideration of multiple competing interests. On the one hand, it is better to avoid semantically informative pre-nominal adjectives for the purpose of constructing atomic lexical representations for nouns, but on the other hand, pre-nominal information should not be ignored altogether. Clearly, more work is needed to understand the precise trade-off that such an intervention would have for the atomicity of distributionally constructed lexical representations and experiential knowledge about the concepts denoted by adjectives and nouns. Based on the time course of the age-order effect, where the counterbalancing pre-nominal contexts has immediate and long-lasting effects at the earliest stage of training, I predict that withholding of semantically redundant pre-nominal adjectives would support lexical semantic development of nominals, provided such an intervention occurred sufficiently early in development when experiential knowledge about adjectives is not yet on children's radar.

12.2.2 Frequency, Repetition, and Contextual Diversity

If my explanation for the age-order effect is correct, this would suggest that word frequency alone is not as important to the formation of atomic lexical semantic representations as other factors such as informativity (i.e. how predictive is a context where it about an upcoming target word?), and fragmentation (e.g. how coherent is a lexical category?). To support the claim that noun frequency is not implicated in the age order effect, I re-ran the simulations presented in Chapter 10 with a modified version of the same corpus such that nouns occur with comparable frequency in each corpus partition. Because in this condition, the age-order effect was still observed, we can rule out the hypothesis that age-related changes in noun frequency alone are responsible for learning more atomic lexical semantic representations in the RNN.

This finding is important because it suggests that quality, and not just the quantity of language input to children can shape the construction of form-based lexical semantic category knowledge. This calls into question the extent to which pure repetition, as opposed to the variety of contexts in which words are repeated, matters in lexical semantic development. For instance, early work on the frequency of linguistic expressions in children's input showed no or even negative correlations with children's syntactic development, suggesting that previous accounts based on drill (e.g. rehearsal and efficient storage) needed scrutinizing (Newport et al., 1977). This shift from effects of the statistics of words themselves (e.g. frequency) to statistics of the contexts in which they occur, is mirrored by a trend in cognitive psychology towards context effects on memory of words and other stimuli (Dennis & Humphreys, 2001). However, while many agree that context effects in language acquisition are real, much debate remains about the different roles that frequency and context diversity play in predicting language growth. Work in syntax acquisition shows each may play a slightly different role; for instance, frequency alone predicts aspects of the acquisition of the German relative clause (Brandt et al., 2008), and verb argument structure (Kidd et al., 2010). On the other hand, context diversity predicts the variety of complex clausal constructions a child will produce (Huttenlocher et al., 2010). For a deeper discussion of repetition and variability in input to children, including cross-linguistic analyses, see (Lester et al., 2021).

The findings in this thesis shed some additional light on this issue: To support the productive re-use of linguistic units, SPIN theory predicts that input which promotes a combinatorial decomposition into syntactic or other kind of constituent structure should produce lexical representations that can be more readily re-used in novel constructions. If, instead, language input consists of repeated expressions that do not clearly demarcate constituency boundaries, an RNN-like distributional learning system would chunk those components, giving rise to 'unanalyzed wholes'. The account developed in this thesis suggests that chunking is a result of repeated exposure to identical expressions, and atomicity is a result of target words occurring in entropy-maximizing contexts. Thus, to the extent that context diversity is driven by entropy-maximizing contexts (Chapter 8), context diversity should aid the formation of atomic representations, and, in turn, the productive re-use of previously observed linguistic units in syntactically well-formed positions. The key insight is that contextual diversity alone says little about the way in which a statistical analysis may facilitate productive re-use of linguistic units; the relationship is mediated by the degree to which a context is entropy-maximizing.

With all this attention to context and away from frequency, what is the role of repetition in the absence of context diversity? There is an independent role for repetition, from the perspective of distributional learning, and especially in connectionist systems prone to catastrophic forgetting. While repetition may not directly support grammatical competence in such networks, repetition can be useful when each occurrence of an expression occurs far apart in the training data, so that a network can revisit and strengthen previously learned knowledge.

12.2.3 An Early Bias for Compound Cues

When language acquisition researchers make predictions about how infants might parse language data, adult intuition about word meanings, and deeply entrenched knowledge about how language works, can often lead researchers astray. The same intuitions can yield incorrect predictions about what a computational model like the RNN will learn from a given sentence or corpus. For instance, when the RNN is provident redundant information, we might be tempted to think that the network will isolate redundant predictors (i.e. linguistic units) and ignore those we know are less relevant, just like we do. However, this view is incompatible with an information-theoretic perspective which does not consider the presence of two cues as constituting more information when each makes identical predictions. Without adult intuitions and in the absence of other kinds of (social) cues, the best option for infants is to rely on informativity (e.g. conditional entropy), which does not proscribe how identical or highly correlated cues should be parsed. If the RNN accurately describes aspects of children's construction of form-based lexical semantic category knowledge, future research should be able to detect an early bias towards compound cues in children. Eventually such a bias should disappear as children converge on adult knowledge about causal relations between referents and their perceptual properties.

As purely distributional knowledge becomes less important, children no longer need to rely exclusively on predictive relations in language, where — mixtures of category-relevant and category-irrelevant — compound cues abound.

12.2.4 Compound Nouns

The work presented herein makes predictions concerning the effect of exposure to noun compounds on the ability to perform the distributionally-mediated extension of category-associated features (DECAF). Consider, for example, presenting the word *bicycle* to a child who has not yet learned the meaning of this word. As explained in detail in Chapter 1, the child may use DECAF to infer semantic features for *bicycle* without previously having had perceptual experiences with bicycles. Importantly, DECAF is in part the result of distributional knowledge, such as knowing that the distributional similarity between *bicycle* and VEHICLE words is greater than other words that do not belong to the VEHICLE category. Consider, that the novel word *bicycle* is presented in the sentence (a) and that the child has previously heard sentences like (b) and (c).

(a) Do you want to ride $[_{NP}$ the bicycle] to town?

(b) Do you want to ride $[_{NP}$ the truck] to town?

(c) Do you want to ride $[_{NP}$ the fire truck] to town?

In this situation, the child might exploit the distributional similarity between *bicycle* and *truck* in sentences (a) and (b) to infer that a *bicycle* is likely a member of the VEHICLE category. Distributional evidence, therefore, would warrant the extension of semantic features common to members of VEHICLE to the novel word *bicycle*. Importantly, the findings in this thesis predict that the success of this kind of induction would depend on the extent to which the child can draw upon atomic lexical representations, as opposed to chunk-level representations. In this case, if the child had been primarily exposed to sentences like (c) where the category-relevant known word *truck* is embedded inside a compound noun, DECAF should be impaired. The reason is that the form-based lexical semantic representation for *truck* would be less similar to *bicycle*, given that some of the distributional semantic properties of *truck* would be trapped inside the representation for *fire*. In contrast, a child who had been primarily exposed to sentences like (b) would likely have constructed a more atomic representation for *truck*, and one that would be more similar to *bicycle* than a child who had been primarily exposed to (c). This prediction can be tested empirically using standard psycholinguistic methods.

12.2.5 The Utterance-final and Utterance-initial Positions

Many previous studies of child-directed input have examined how the placement of a target word in either utterance-initial or utterance-final position may support acquisition of the target word, and grammar. Among some of the findings are that caregivers preferentially introduce new words in utterance-final position (Fernald & Mazzie, 1991), and that infants segment words from the edges of utterances more readily than from the middle (Seidl & Johnson, 2006). Further, there is computational evidence for the facilitatory role of utterance-boundaries in distinguishing between grammatical classes (Freudenthal et al., 2013; Freudenthal et al., 2016). A classic finding is that the extent to which children hear auxiliaries in utterance-initial position is linked with future ability to use auxiliaries (Newport et al., 1977). Whereas absolute frequency of auxiliaries in a child's input is not predictive of later competency, there is a strong correlation with a mother's tendency to use auxiliaries in utterance-initial position (e.g. 'C*an you kiss your elbow?*').

However, most studies have not considered the potential effect that the utterance-final and utterance-initial position might have on lexical semantic development. If such studies were to be conducted, SPIN theory predicts that nouns which occur in these positions would be represented more atomically, and, this would, in effect, facilitate the discovery of form-based semantic category clusters, and, in turn, DECAF. Why would utterance boundaries promote lexical atomicity? The reason is that prediction error tends to be highest at such boundaries, acting as a natural bottleneck that squashes potentially maladaptive chunk-level statistical dependencies. Put differently, the words that precede a target word in utterance-initial position or follow a target word in utterance-final position are less likely to enter into a redundancy relationship with the target word. In order for this to work, utterance boundaries — perceptible in the acoustic signal by the presence of brief periods of silence — must be treated as linguistic units in distributional analysis. Computational evidence for the benefit of including such units was previously obtained by Freudenthal et al. (2013) who observed that inclusion of utterance boundaries as features in a classifier improved part-of-speech classification. This finding is consistent with the work presented herein, namely that high prediction error is needed to discover lexical categories in a model trained on next-word prediction. Sensitivity to these silent units would reduce lexically-specific, category-irrelevant predictive shortcuts useful for prediction error minimization, but detrimental to lexical atomicity. This account provides an alternative to established wisdom that the utterance-final and utterance-initial position are psychologically privileged, and instead suggests that their facilitatory effect on language growth might in part be due to the high prediction error that must be generated at utterance boundaries. More work is needed to understand whether children preferentially listen to utterance beginnings and endings, whether facilitation can be explained in terms of the special status of utterance-boundaries from an information-theoretic perspective, or both.

Chapter 13

Limitations and Future Directions

In this chapter, I discuss several limitations of the current work, and future directions that would extend and test claims made in this thesis. I begin by pointing out limitations of the modeling work, Semantic Property Inheritance (SPIN) theory, and AO-CHILDES. I end the chapter by proposing future directions, including extensions to the basic RNN architecture, the potential for other models to acquire more atomic lexical organization, and the need for behavioral experiments to connect the work herein more closely to the psycholinguistic literature.

13.1 Limitations of the Model and Simulations

13.1.1 Extension of Semantic Category-Related Features

There are several limitations of the current work. One is that I did not actually model the distributional extension of category-associated features (DECAF). The premise of this thesis is that a distributional system that has learned atomic lexical representations for nouns is better suited for inducing category-associated semantic features for novel nouns than a system that has captured chunk-level distributional regularities. By exploiting the distributional similarity between known nouns (e.g. *bird*, *cat*) and a novel noun (e.g. *gorilla*), a learner can infer semantic features useful for diagnosing semantic category membership, such as is-ANIMAL or is-VEHICLE. While atomic lexical semantic representations provide better estimates of semantic category membership (by definition), additional work is needed to empirically verify that more atomic representations actually correlate with better DECAF outcomes. This would require a system that goes beyond the RNN where (i) perceptual and conceptual semantic features are learned and stored, and (ii) extended to novel words when sufficient distributional evidence that warrants doing so becomes available. For an example, see M. Jones and Recchia (2010).

13.1.2 Lexical Polysemy

Many words in English and in other languages are polysemous, expressing a family of distinct but sometimes related meanings. As an example of related meanings, *chicken* can refer to a kind of animal or meat; as an example of unrelated meanings, *bank* can refer to a financial institution or a feature of a river. Yet within developmental science, word learning is typically studied as if children need to learn one meaning per word. The same invalid assumption applies in computational work when researchers dedicate exactly one vector representation for each word in a model's vocabulary. By defining lexical semantic representations as static entries in the RNN's input-to-hidden-weights, I have committed the same error. Because the input-to-hidden weights of the RNN contain exactly one vector for each orthographically identical token, these vector representations converge on mixtures of information about potentially multiple different meanings expressed by the same word. This is potentially problematic because vectors that encode mixed information are more likely to be some distance away from the positions in the semantic space of the individual meanings. Put another way, the vector for a word that is semantically similar to one of the target meanings, would be farther away from a vector that encodes a composite meaning compared to a vector that is dedicated to the target meaning only. The paradigmatic similarities produced by a model without dedicated vectors are therefore skewed, and are likely less effective for performing DECAF. Whether this intuitive explanation holds in the real world requires empirical testing. In this work, I have avoided these concerns by focusing on nouns that are for the most part monosemous — words that have a single, or otherwise strongly dominant, meaning.

It should be noted that the RNN does in fact differentiate the different meanings of polysemous words. But it does this only at the hidden layer, where contextually sensitive semantic information accumulates while processing entire sentences. However, as explained in Chapter 1, the contextualized representations at the hidden layer are not suitable representations of lexical items because they capture all sorts of other information not relevant to a single item. This means that polysemy is not necessarily problematic for the RNN; the problem is that disambiguation requires context, and lexical representations are context-less, so to speak. This points to a deep issue at the core of this work: I have been extracting knowledge from a location in a model that was not specifically created for this purpose. The model has no way of differentiating different meanings of the same word at the input-to-hidden weights, and yet, that is precisely where I am extracting lexical semantic representations. What can be done?

One way that children can solve the polysemy problem is by exploiting phonological information. This has been shown to work well for ambi-categorical words, which have different meanings depending on their part-of-speech context (Conwell & Morgan, 2012). Being able to identify the part-of-speech of a word in this way would help the RNN correctly align the semantic properties that are specific to a part-of-speech class or word sense to the representations dedicated to that class or sense. Of course, this would require inputting unsegmented language in the form of acoustic signal to the RNN, which would raise many questions about how to segment the input, identify distinct word meanings, and recruit dedicated lexical representations for each. Such efforts are already underway in the speech recognition community (Räsänen & Rasilo, 2015).

13.1.3 Input is Pre-Segmented and Localist

One of the implicit assumptions of this work is that the input to children's distributional learning system has already been segmented; that is, the lexical units have already been isolated from the speech stream prior to being input to the distributional learning system. This assumption is not a theoretical commitment, but rather a simplification adopted to examine the feasibility of modeling lexical semantic representation learning in the RNN under idealized conditions. It is not clear whether a more dynamic interaction with a segmentation system would affect the atomicity of learned representations. At first blush, it appears that a more variable segmentation might destabilize learning in the distributional system: For instance, on some occasions, the distributional semantic properties of *fire truck* might be assigned to a single representation for the compound as a whole, and on other occasions, the word segmentation system might separate the compound noun such that semantic properties of the whole are encoded as an interaction among the parts.

The latter situation is maladaptive, for reasons discussed at length in Chapter 6.

Although the input to the RNNs trained in this work is pre-segmented, I have conducted several investigations into the atomicity of learned representations under conditions in which the input has been segmented differently. In total, I have examined three segmentation strategies: Units were identified by (i) splitting transcribed text on white-space, (ii) tokenizing transcribed text using the Python library *spacy*, or (iii) tokenizing using the Byte-Level Byte-Pair Encoding strategy implemented in the Python *tokenizer* library. The age-order effect was observed in each condition, which demonstrates that it is robust against different segmentation strategies. In a related experiment conducted with Andrew Flores, I that the RNN learned more atomic lexical semantic noun representations when the input to the RNN was morphologically parsed.[1]

How infants learn to segment the speech stream is a difficult research question in its own right. Emerging evidence indicates that infants leverage a variety of information sources to succeed at this task (McCauley & Christiansen, 2019; Monaghan et al., 2007). Future work is needed to bridge the output of models directly trained on acoustic input with distributional semantic models trained on pre-segmented and localist input.

13.1.4 Shortcut Learning Bias

The dynamics of learning statistical associations in the RNN is potentially at odds with the way that humans learn complex rules. Despite having been exposed to stimuli that promote the acquisition of a simpler rule, people eventually arrive at a more complex rule that replaces the simpler rule (Kruschke, 1996; Nosofsky et al., 1994). In particular, Nosofsky et al. (1994) reported that knowledge of the exclusive-or rule (XOR) emerges during training despite being given feedback in a classification task in which a simpler one-dimensional rule was sufficient for successful classification. Interestingly, during early training, people applied the more expedient one-dimensional rule, but later learned the more complex XOR rule, seemingly spontaneously, as evidenced by sudden successful categorization of stimuli where the more expedient rule fails. These results suggest that people can spontaneously acquire a more complex rule, and even after having previously learned a simpler and more expedient rule, provided enough learning opportunity. Taken at surface value, this ability appears to contrast with learning in the RNN, where previously acquired 'shortcuts' tend to stick around in the long-term, and impede acquisition of a more complex rule or heuristic. Unlike people, who spontaneously transition to a more complex rule, the strong tendency to rely on previously learned shortcuts, suggests human learners are potentially equipped with additional capabilities that go beyond the slow error-based discovery of implicit regularities in connectionist systems (but see Anderson et al., 2019b).

The distinction between fast insight-based and slow incremental learning is often termed the procedural-declarative contrast, and has been called upon to explain differences in child language acquisition (Ullman, 2004; for review, see Hamrick et al., 2018). Further research is needed to clearly establish (i) when people spontaneously arrive at a more complex rule, (ii) whether acquisition is indeed spontaneous as opposed to gradual, and (iii) whether similar insight-based learning is necessary to explain aspects of distributional learning in language, the domain in which the RNN appears to be most viable as a cognitive model.

[1] Morphological parsing included stripping infinitival and past tense endings from verbs, and plural, possessive, and diminutive endings from nouns.

13.1.5 Semantic Categories that Cross Part-of-speech Boundaries

The choice of using the RNN as a model of children's distributional learning system requires certain assumptions and/or commitments about the kind of knowledge structures that can emerge. For example, the lexical semantic knowledge that emerges in the RNN is heavily constrained by part-of-speech boundaries. But, in fact, there are many other kinds of semantic category structures which children might use to organize their knowledge of word meanings — not just *within* nouns or verbs, but also *across* grammatical categories. Lexical semantic categories that exist within POS boundaries are useful because they point to substitutability relations within linguistic contexts, i.e phrases, and sentences, and are therefore useful to support children's distributionally-mediated extension of category-associated features (DECAF). In contrast, across-POS semantic categories are useful because they point to relations among words in the same or similar conceptual domain (i.e. semantic field). Lexical semantic categories whose members are not within the same grammatical category (e.g. *school* and *learning*; a noun, and verb, respectively) are of interest because they form an intermediate link between knowledge about how words function in a language and conceptual knowledge. My focus in this thesis on semantic distinctions within the noun category is clearly a small and incomplete window on the organization of children's semantic knowledge. More research is needed to establish whether across-POS — often termed thematic, or syntagmatic — semantic relations play an important role in children's lexical semantic development, and if so, how the RNN might be extended with (or replaced by) additional machinery that operates on these distributional semantic relations.

While a bias for paradigmatic similarity is useful for modeling DECAF, syntagmatic similarity is also potentially useful to the developing child. One way out of this dilemma, is to use the next word prediction machinery to derive syntagmatic similarity judgments, instead of deriving those from the similarity between learned lexical representations. More work is needed to better understand how paradigmatic and syntagmatic similarity trade-off during early lexical semantic development, and how each is computed; are separate systems responsible for each, or can the same system such as the RNN be used to obtain both?

13.2 Limitations of SPIN theory

13.2.1 Limited Analysis of Lexical Context

The experiments that form the foundation of SPIN theory have not exhaustively examined all the possible ways in which lexical context may influence learning in the RNN. While I have examined both left and right contexts, my analyses were limited to immediately adjacent context words. Prior work has shown that words are predictable at much longer distances, including distances of at least dozens of words (Bullinaria & Levy, 2007).[2] I have not examined non-adjacent context words in my corpus analyses because the RNN is initially most sensitive to adjacent dependencies, and only gradually expands its temporal window in which learning can take place (T. A. Chang & Bergen, 2022; Ravfogel et al., 2019). Similarly, it is plausible that children also begin their analysis of linguistic co-occurrences by selectively attending to adjacent context words. Given that the distributional learning capacities of young children may not extend beyond adjacent relations, the analysis of adjacent co-occurrence relations is of most theoretical importance.

Another reason for examining the effect of only adjacent dependencies is the following: By strategically manipulating only the redundancy between a target word and an adjacent neighbor (e.g. Chapter 5), the

[2] My own work showed that the average correlation length for probe words in AO-CHILDES is about 8-9 tokens.

results represent the best-case scenario for the RNN. Because the primary goal of this work is to establish the feasibility of the RNN for learning lexical semantic representations, failure to learn atomic lexical semantic representations under these most felicitous conditions would provide clear evidence that atomicity would also likely not emerge under less felicitous conditions. The reason is simple: redundancy, and therefore fragmentation, can only get worse with additional context. When an adjacent context word provides redundant information, it is impossible for another non-adjacent context words to mitigate this redundancy — existing redundancy can only be strengthened. This is backed up by evidence from analyses of natural language statistics. For example, Bullinaria and Levy (2007) showed that the kinds of subcategory-specific co-occurrences that would encourage fragmentation of the noun category are more likely to occur at greater distances from the target word. This idea, combined with the fact that sentence length and combinatorial diversity, factors that tend to contribute to fragmentation, are greater in speech to older children (Foushee et al., 2016; Hayes & Ahrens, 1988; Kirchhoff & Schimmel, 2005), suggests that my analysis is underestimating the extent to which speech to younger children helps protect against noun category fragmentation.

Furthermore, SPIN theory only considers how left-contexts can potentially negatively influence atomicity of lexical representations for target words; however, in work published elsewhere, I have shown that right-contexts, too, may negatively impact atomicity (P. A. Huebner & Willits, 2021a). More specifically, when category-relevant information in the right context of target words is separated by intervening items, the intervening items may enter into a redundancy relationship with both the target word and informative right-context, and, in turn, impede atomicity regardless of whether the left-context is predictive or not. The farther away a semantically informative signal in the right-context, broadly defined, is, the more opportunity for items in the intervening span to provide redundant information. This means that a thorough understanding of the construction of form-based lexical semantic representations in the RNN is not complete until such non-adjacent dependencies, and their potential to fragment statistics associated with grammatical categories, have been examined. For instance, there is preliminary evidence that nouns are increasingly followed by modifying clauses across developmental time. For instance, Szubert et al. (2021) found that, in the Adam corpus, post-nominal adjectival clauses and other kinds of relative clauses that are typically wedged between a noun and the verb with which it participates, become more frequent across developmental time. From this we may preliminary conclude that the potential of such intervening clauses to participate in maladaptive across-target relationships increases in language input to English-learning children. This warrants a more detailed study of redundancy between multiple items in the right context of nouns and across longer distances; I leave such analyses for future work.

13.2.2 Interplay between Left and Right Contexts

Finally, what is missing from SPIN theory is an account of how lexical atomicity is influenced in the presence of redundant left context *and* in the absence of semantically uninformative right-contexts. The counterbalancing requirement (requirement 2 of SPIN theory) states that redundant information provided by left-contexts is only problematic when the target word is followed by a semantically informative right-context. This means that left contexts are not always detrimental to atomicity: When left-contexts co-occur with semantically vacuous right-context, left-contexts cannot inherit category-relevant semantic properties of the target word, because none are given. This is a good thing, but I have not shown how frequently this occurs in child-directed language, or whether this situation is specifically exploited by caregiver input to younger relative to older children. In fact, while I have only proposed one strategy for improving lexical atomicity (by counterbalancing left-contexts), an alternative strategy would be to only permit the occurrence of semantically informative left-contexts in the absence of semantically informative right-contexts. Toward that end, I have conducted some preliminary analyses that suggest this is indeed another way in which nouns are protected from fragmentation in input directed to younger children: Preliminary analyses of AO-CHILDES (not shown) demonstrate that nouns are followed by utterance-boundary markers much more frequently in input to younger as opposed to older children. Because utterance-boundary markers (e.g. period) provide virtually no semantic information about the nouns that precede them, there is less potential for pre-nominals to enter into a maladaptive redundancy relationship with post-nominal contexts in input to younger children. Therefore, even with equally fragmenting pre-nominal contexts in input to both age groups, the frequent occurrence of semantically vacuous punctuation in the right contexts of nouns in input to younger children would further immunize the RNN against encoding across-target statistical relationships which impede atomicity. More work is needed to confirm this and to elucidate the role that utterance-final nouns play in preserving lexical atomicity.

13.3 Limitations of the Corpus

Next, I discuss several limitations of working with AO-CHILDES. In particular, I consider (i) that AO-CHILDES is an aggregate collection of transcribed speech to a large number of children, the need to validate SPIN theory (ii) outside of the child-directed language domain, (iii) and in languages other than English, and (i) limitations of corpus research for drawing causal inferences.

13.3.1 Individual Differences

The CHILDES database (MacWhinney, 2000), which is the source of all data in the AO-CHILDES corpus, contains transcribed speech to several dozens of children, and was produced by many caregivers including mothers, fathers, family members, and experimenters. As such, the corpus is an aggregate picture of child-directed input, and is therefore not representative of any particular child. This has two consequences: First, aggregation makes it difficult to study individual differences in developmental trends in caregiver input; some caregivers may not adapt their speech to younger children as much as other caregivers do. Second, aggregation by collapsing transcripts of speech to children with similar ages into the same age bin, potentially masks the effect that age-ordered presentation of input might have on the lexical semantic development of individual children. On this view, it is possible that the age-order effect observed in Chapter 10 is severely underestimating the potential for a staged training regime to promote the formation of atomic lexical semantic representations for nouns in the RNN. To illustrate, a comparably sized corpus of transcribed speech to an individual child might be characterized by an even stronger longitudinal trend in fragmentation of the noun category (i.e. lower fragmentation when the child is younger), and therefore would promote an even greater atomicity bias during early training compared to using an aggregated corpus like AO-CHILDES.

There would be additional advantages of having large longitudinal corpora of input to individual children. If available, it would be possible to examine the influence of the input an individual child was exposed to, such as fragmentation of the noun category, and use these measures to predict learning outcomes, such as vocabulary size, and other markers of healthy lexical semantic development. One prediction that can be derived from the work presented herein is that the amount of fragmentation of the contexts in which an individual child has heard nouns would be negatively correlated with behavioral measures of that same child's understanding of the meaning of nouns, and the number of nouns he or she knows.

13.3.2 Languages other than English

An important follow-up question is whether an age-order effect would result when training on age-ordered language input to children learning languages other than English. Cross-linguistic validation is needed before we can draw strong conclusions about potential universal properties or objectives underlying caregiver input directed to children. In the meantime, we must consider the possibility that the age-order effect is little more than an artefact of English. There are several reasons for this: First, not all languages place nominal modifiers pre-nominally (Dye, 2017). In fact, across languages, post-nominal placement is significantly more frequent than pre-nominal placement (Culbertson et al., 2012). In such languages, it is less likely that left-contexts of nouns provide redundant information about upcoming right-contexts. All else being equal, there would be no a priori reason to predict that, in such languages, we would find asymmetries in the longitudinal structure of input to children that could increase atomicity of learned representations for nouns. In other words, because pre-nominal contexts are already better 'counterbalanced' in such languages, it would matter less whether an RNN were trained on input to younger or older children first — the amount of counterbalancing would likely be comparable.

Different languages likely make different trade-offs concerning learnability and efficiency of processing. For instance, cross-linguistic differences in word-order can be mapped onto distinct paradigms from associative learning (Osgood, 1949). In the 'divergent' paradigm, the sequential presentation of stimuli is designed to keep entropy constant; in the 'convergent' paradigm, entropy of predicting upcoming stimuli is much more variable. For example, the front-loading of nominal modifiers in English is an example of a divergent information structure, because a pre-nominal modifier tends to reduce prediction error at the noun. In contrast, a post nominal modifier would not be able to lower the next word prediction error at the noun when processing left-to-right. To understand the potential psychological consequences of learning in one or the other paradigm, Dye (2017) examined how the information structure of word sequences (i.e divergent vs. divergent) influenced outcomes in an implicit word learning task. The authors found that in the divergent paradigm, word sequences were processed more efficiently and were better recalled; however, the convergent paradigm facilitated semantic learning. Importantly, these two paradigms are mutually exclusive; the information structures adopted by different languages therefore optimize learnability or efficiency of processing, but not both. Many other characteristics of a language, such as phonology and morphological complexity, likely contribute to balance the overall trade-off.

Further, recent work has revealed unexpected cross-linguistic differences in longitudinal trends related to statistical properties of input to children. For example, Lester et al. (2021) found that in morphologically complex languages, such as Turkish, repetition tends to increase in child-directed input across developmental time, rather than decrease as it does in English (Chapter 2). What this shows is that (i) we cannot readily extrapolate findings from English to other languages, and that (ii) longitudinal trends identified in one corpus of child-directed input in one language may be inverted in another.

Cross-linguistic validation was not undertaken here for two reasons: First, the CHILDES database is only sparsely populated with data from non-English languages. The resulting non-English corpora would therefore be much smaller compared to English, and contain fewer data to fully cover a wide age range. Second, the study of atomicity would require construction of language-specific target semantic category structures, which in turn would necessitate native speakers of non-English languages for conducting norming and validation studies. For now, the answer will have to wait until we have sufficiently large and longitudinally broad corpora of cross-linguistic input to children.

13.3.3 Causal Inferences

Correlation does not imply causation. This is especially important when interpreting longitudinal analyses of the AO-CHILDES corpus (or any other corpus). The corpus by itself does not give us insight concerning why caregivers talk the way they do to younger compared to older children, or what compels them to adapt the way they talk as children grow older. While it is tempting to draw causal conclusions from observations of corpus data alone, we must refrain without additional experimental validation. The fact that caregivers talk differently to children, and especially to younger children, does not by itself show that caregiver's linguistic adaptations either influence acquisition, nor that they are necessary for acquisition. To make such claims would require evidence from intervention studies that manipulate how caregivers talk to their children. Such studies have not been conducted, presumably because they would be practically infeasible (it would be difficult to reliably restrict how caregivers talk to their children over the course of many years), and, further, such studies could be considered unethical or excessively intrusive. While a great deal of sophisticated correlation studies have been conducted over the years, one cannot be certain of cause-effect relationships, no matter how sophisticated the methodology is (Newport et al., 1977).

It is possible that changes in the statistical properties of language input across developmental time — the shift from child-directed to adult-directed language — does not occur for pedagogic reasons (i.e. optimizing language teaching), but for reasons related to socio-communicative and pragmatic pressures (Dong et al., 2021). In other words, we should not assume that caregivers talk to their children to teach them about language; rather, caregivers talk to their children in order to communicate with them — to transmit meaning in a format that the child is likely to comprehend. In fact, there are many reasons why caregiver speech might differ from adult-direct speech that are unrelated to language teaching: For instance, simplification may arise due to (i) children's preferences for certain kinds of input, (ii) their limited ability to engage with more sophisticated input, and (iii) the range of conversational topics that are relevant to children and their caregivers (the emphasis on the so called 'here and now'). My corpus analyses make little contact with questions about the psychology and ecology of caregivers. Questions about the constraints on producers, and the child-caregiver dyad are interesting directions to pursue in the future. For example, to what extent are caregivers aware of their linguistic choices when speaking to children, and do they adjust the complexity of their speech in accordance with the linguistic competence of the child?

13.4 Future Directions

13.4.1 Integrating Extra-linguistic Knowledge

The RNNs trained in this work were not provided access to extra-linguistic input that would ground their knowledge of lexical co-occurrence data in the perceptual properties of the natural world. Numerous works in cognitive psychology have shown the importance of extra-linguistic information in lexical semantic development, such perceptual experience (Barsalou, 2008; S. S. Jones et al., 1991a), conceptual knowledge (Pinker, 1987), the ability to perform actions on real-world entities (Yu & Smith, 2012), and cognitive simulation (N. Chang et al., 2005). However, the goal of this thesis is not to show how the construction of form-based meaning classes interacts with other knowledge systems, but rather, whether the RNN can learn distributional lexical semantic representations that could support the distributionally-mediated extension of category-associated features (DECAF). Put differently, the work in this thesis is based on the premise that there is (i) a system that is responsible for collecting distributional information in children, and (ii) that the distributional knowledge gathered by this system is kept separate from perceptual and conceptual knowledge about word meanings. That is, the simulations in this work are not meant as approximations for concept development or acquisition of perceptual features of words. I consider the distributionally constructed semantic space that emerges in the RNN as a system of knowledge that is, largely, kept separate from experiential knowledge about word meanings. That said, the knowledge accumulated by the RNN interacts with meaning systems during word learning — for the purpose of migrating category-associated semantic features from known words to novel and distributional similar novel words (i.e. DECAF).

However, this approach leaves unanswered an important question concerning how separate children's distributional knowledge is from experiential knowledge about word meanings. In contrast to the simulations presented here, children's distributional systems may be intertwined with knowledge of grounded, embodied, experiential knowledge from vision, hearing, touch, taste, and smell. How well children are able to separate language-internal distributional statistics from other statistics is an empirical question. If it turns out that children's distributional system is tightly integrated with perception, the assumption that guides this work, namely that distributional information is kept separate, would need to be re-evaluated. Experiential knowledge may interact with and modify representations of words as children learn to link words to their meanings. In fact, several scholars argue for such a tight integration (J. L. Elman, 2011; M. Jones & Recchia, 2010; Riordan & Jones, 2011). On the flipside, it is likely that integration of extra-linguistic knowledge would not qualitatively change the results presented herein. There is no a priori reason why co-occurrence statistics of entities, as opposed to words, should be treated differently by the network. One possibility, however, is that, in such a network, syntactic factors may no longer be as influential for organizing learned representations compared to when distributional statistics is the only source of information available.

13.4.2 Non-Stationary Input

One of the primary insights of this thesis is that the way in which nouns pattern with other words changes in theoretically important ways in input directed to children between the age of 1 and 6 years. This suggest that similar developmental trends might be found in other information sources available to children. Lexical co-occurrence statistics are probably not the only input that is non-stationary — undergoing a distribution shift across developmental time. For example, visual and tactile input to young children is likely constrained to a small set of toys and people (e.g. family members but very few strangers), and is gradually expanded as they grow older. Further, the number of environments that a young child experiences (e.g. home, yard) is probably smaller and less varied than those experienced by older children (e.g. school, mall, neighborhood, etc.). Preliminary evidence for this idea comes from studies of children's head-cam recordings, which showed that the distribution of objects present in children's visual field changes over the first two years of life (Fausey et al., 2016). In sum, a theory of child lexical semantic development would be incomplete without considering how changes in extra-linguistic input channels influence how lexical semantic representations are learned.

13.4.3 Categories other than Nouns

The theory presented in this work has been developed to be broadly applicable to any grammatical category, not just nouns. The primary reason that nouns were used in this work is that (i) they can be straightforwardly separated into semantic categories, (ii) they are pre-dominant in children's early vocabulary, and (iii) their linguistic contexts are more straightforward to analyze than, say, verbs which project argument structure. In principle, the same benefit of age-ordered training could also apply to verbs or adjectives, because they, too, can be broken down into smaller semantic classes. For example, a verb can be semantically distinguished by whether the activity it refers to involves transference, motion, or contact (Levin, 1993). Similarly, an adjective can be semantically distinguished by whether the property it refers to involves color, texture, or size (Dixon, 2010). However, whether these semantic category boundaries can be discovered using distributional statistics, and whether age-ordered training is beneficial, would require empirical confirmation.[3] One possibility is that distributional statistics do tease apart semantic classes of verbs, but that these statistical differences would be strongly correlated with syntactic properties unique to different verb classes. If so, a staged training strategy may not make the representations learned by the RNN more atomic. The staged training strategy is useful for nouns because it emphasizes the relations between nouns *relative to other nouns*. This is only possible because nouns are syntactically homogeneous in English. The only impediment to atomicity is the redundant information provided by their left-contexts. For verbs, atomicity would likely also be impeded by additional factors, which are not addressed by SPIN theory. In effect, it is likely that semantic verb clusters would not be progressively differentiated from a top-level verb category, but independently of each other. If so, the order in which data is presented to the RNN would matter little in how these clusters relate in semantic space.

In general, verbs are semantically more complex than nouns (Gentner et al., 2001), and their interpretation is much more reliant on context (J. L. Elman, 2011; McRae et al., 1998). In fact, J. L. Elman (2011) argued that the enterprise of learning static lexical semantic representations for verbs is misguided. The key insight is that how people interpret verbs relies on a variety of contextual factors, and that storing this knowledge in a static lexicon is incompatible with the traditional notion of a lexicon as a dictionary (R. Jackendoff, 2002). The evidence in favor of this argument is that during online sentence processing, people's hypotheses about which of multiple verb senses is most likely, are based on the degree of thematic fit of previously observed verb arguments. To J. L. Elman (2011), such observations suggest a new way of thinking about lexical knowledge, and argues that words are 'cues to meaning' rather than inherently meaningful units that enter into a processor devoid of meaning. With this in mind, it is likely that the distributional structures that correlate with semantic category membership is qualitatively different between verbs and nouns: While learned representations for (concrete, common) nouns may approach high degrees of atomicity, the ubiquity of context effects on verb interpretation makes this prospect much less likely for verbs. More research is needed to support this possibility.

13.4.4 Extensions and Other Architectures

In order to learn lexical semantic representations that would be more useful for the induction of semantic category-associated features for novel words (Chapter 1), researchers working within the language modeling formalism are faced with two alternatives: (i) either let go of the notion of a static lexicon that houses atomic lexical knowledge, or (ii) develop inductive biases that promote such an organization. The findings presented in this thesis show that the latter approach continues to be useful; in particular, curriculum learning strategies, such as staged training regimes, for constraining the solution space of neural networks models, are methods that might support the convergence on more atomic representations and internal states without being explicitly told to do so. In this section, I explore such biases, in the form of architectural extensions to the RNN, but also other non-recurrent architectures and alternative learning systems.

While the work herein suggests that there are strong constraints on the conditions in which lexical

[3] Preliminary analyses of verbs in AO-CHILDES suggest that they are not less fragmented in language directed to younger children, contrary to the developmental pattern for nouns.

atomicity can emerge in the RNN, it would be pre-mature to conclude that the RNN is inadequate in principle. Since the simple RNN was first introduced by J. L. Elman (1990), a great number of extensions and augmentations have been introduced to deal with its shortcomings. One example, is the addition of gating units at the hidden layer that allow it to learn longer distance dependencies (Hochreiter & Schmidhuber, 1997). This architecture, the LSTM, while much improved for tracking longer distance dependencies, does virtually nothing to hamper entanglement; the findings reported in Chapter 10 clearly demonstrate that the LSTM is just as vulnerable to redundancy as is the simple RNN. Fortunately, there are many extensions that explicitly aim to learn more generalizable states (and therefore, more atomic lexical representations); such strategies typically include some form of self-attention (Ke et al., 2018), explicit or implicit linguistic supervision (Shen, Lin, et al., 2018; Shen, Tan, et al., 2018), units that explicitly encode categorical information such as semantic roles (F. Chang et al., 2006), the addition of external memory (Webb et al., 2020), pre-training (Calvo & Colunga, 2003), and regularization that constrains, in a principled manner, which kinds of mappings can be learned (Gordon et al., 2019). Below, I discuss some of these ideas, their potential for producing more atomic states, and alternative architectures and algorithms that attempt to do this explicitly.

Maturational Constraints on Processing

A promising future direction, first proposed by J. L. Elman (1993) and Newport (1990) is to build maturation constraints on language processing abilities into models of language acquisition. This idea is in alignment with findings presented in this thesis, which suggest the need to 'ignore' semantic information provided by left-contexts when it is redundant with information also provided by an upcoming target word. Such a maturation constraint on processing would be supported by behavioral evidence that children prior to the age of 3 years do not readily integrate semantic information provided by pre-nominal adjectives with nouns (Ninio, 2004; Sekerina & Trueswell, 2012; Thorpe et al., 2006). By not integrating adjectival meanings, children can focus on the association between the noun and upcoming linguistic material independently of the adjective. Children's propensity to naturally separate these two learning problems would approximate the staged learning regime proposed in Chapter 9.

Selective Attention

It is possible that constraints on processing are not due maturational constraints on brain and cognitive development, but arise regardless, because they are of use to the learner. Rather than conceiving the child as a passive listener with upper limits on the amount of information they can process, children might actively select dimensions of the input that are most beneficial in accordance with their developmental stage. As explained in Chapter 1, the ability to selectively attend to specific aspects of the input is an important pre-requisite for acquiring atomic lexical semantic representations. There is some support that children do selectively attend to linguistic input (Ervin-Tripp, 1973; Shipley et al., 1969; Slobin, 1973). This idea, sometimes called 'child as filter', is also in agreement with a proposal by Newport et al. (1977) about children's selective attention to linguistic material:

> The basic position for which we will now argue is that the child is biased to listen selectively to utterance-initial items and to items presented in referential obvious situations: the child acts as a filter through which the linguistic environment exerts its influence.

Important questions remain: What do humans do when faced with redundant predictors (in language or other domains)? What biases or strategies might they use to distinguish between compound cues (where multiple predictors work together) and independent cues? Under what circumstances (e.g age, modality, task), do people decide (explicitly or implicitly) whether two statistical cues are better treated as independent predictors or combined into a compound cue? Finally, how can the RNN be extended to incorporate selective attention? Preliminary work on this front has proven the feasibility of this approach (Barrett et al., 2018).

Explicitly Encoding Syntactic Dependency Relations

It is possible that more advanced processing capabilities would further support the formation of atomic lexical semantic representations. One promising research direction is to constrain the set of dependencies that the RNN is allowed to consider in a principled manner. Work by Shen, Lin, et al. (2018) has examined the capabilities of an RNN that also performs 'latent tree learning' as part of the language modeling task. The extended RNN model, called the Parsing-Reading-Predict Networks (PRPN), can simultaneously induce the syntactic structure from unannotated sentences and use the inferred structure to learn a better language model. This is accomplished by incorporating a neural parsing model into the RNN, and back-propagating the gradient from the language model loss into the parsing network. Follow up work shows that this model is robust, and "strikingly effective at latent tree learning" (Htut et al., 2018). In related work, Russin et al. (2020) have shown that the RNN can be made more robust against idiosyncratic dependencies by separating syntactic and semantic processing. Their approach proved useful for capturing compositional structure in a difficult sequence-to-sequence learning task requiring compositional generalization.[4]

Transformers

Another question concerns how well the current findings generalize to other models, including recently introduced language models based on the non-recurrent Transformer architecture (Vaswani et al., 2017). Given that the findings herein are robust to variation in the RNN architecture (simple RNN vs LSTM), it is possible that many of the principles identified in this work will also be applicable in other connectionist language models. However, analyses of how Transformer-based language models represent natural language data have shown a surprisingly robust capacity for encoding abstract knowledge without the need for specialized training strategies. For instance, (Ravfogel et al., 2021) showed that Transformers are able to represent relative clauses as an abstract unit, which is something that RNNs struggle to learn (Linzen & Baroni, 2021). Without work that compares these two architectures side-by-side it is difficult to draw conclusions regarding the apparent advantage of Transformers over RNNs in learning context-free, abstract representations. Does this have to do with the absence of recurrence in Transformer language models, and the fact that they are less constrained by the liner ordering of items in their input? Linear order is not encoded as part of an architectural constraint but using positional encoding vectors, which are randomly initialized. This means that Transformer language models optionally learn positional information but are not constrained by it. It is possible that the replacement of recurrence, which enforces linear order, with self-attention ameliorates many of the concerns regarding lexical atomicity raised in this work, and that specialized constraints on training, such as the presentation order of the data, may be less important for Transformer-based compared to RNN-based language models. An alternative explanation, however, is that the apparent superior abilities

[4] Compositional generalization is the transfer of knowledge about semantic properties of simple expressions to situations not encountered in previous experience with language (see Kim and Linzen, 2020)

of Transformers are due to having been trained on much larger datasets compared to RNNs.[5].

Preliminary comparisons of a Transformer language model based on GPT-2 (Radford et al., 2019) and a comparably sized RNN trained on equally sized corpora showed that the Transformer learns more generalizable lexical semantic representation (Mao et al., 2022). Specifically, Shufan Mao[6] and I found that the miniature GPT-2 language model consistently outperformed the simple RNN in a semantic inference task that requires compositional generalization. To succeed in this task, a model must infer the plausibility of word combinations never before seen during training. This task requires a structured decomposition of the input sequences seen during training, and learning similarity relations between paradigmatically similar words. Follow-up analyses showed that the simple RNN did not learn the form-based semantic similarities implicit in the training data because the semantic signal that diagnoses semantic category membership always occurs in the left context of target words. This finding is an agreement with observations made in Chapter 5. It is likely that the reason for the improved performance of the Transformer is due to its ability to take advantage of semantic cues in the left-context when encoding lexical semantic representations.

Another promising direction is work by Zheng and Lapata (2021) who introduced a novel extension to Transformer language models purposefully designed to disentangle internal states. Briefly, the output of the Transformer decoder module is used to condition the encoder module at every time step; this enables the system to attend to specific relations while making predictions, and thereby reduces the pressure to combine information about multiple relations in the output of the decoder. Their approach, called 'DANGLE', performs well on a number of benchmarks of compositional generalization, even surpassing the state-of-the-art in some cases.

Graphical Models of Distributional Semantics

Alternatively, it is possible to encode distributional semantic information in a graphical data structure. An advantage of doing so is the ability to de-couple representation (structure) from processing (function): While distributional similarity is captured in the network topology, semantic inference can be computed using a spreading-activation procedure, or random-walk on the network topology (Mao et al., 2022; Rotaru et al., 2018). The advantage of separating structure and function is that the procedure used to acquire new knowledge and the procedures that use learned knowledge during semantic inference do not interfere with each other. In the RNN, processing, learning, and inference are all intertwined, and are based on the same task, next word-prediction. This means that the procedure used during semantic inference is strongly tied to the procedure used to acquire learned knowledge. In other words, because the RNN has been optimized on a single task, the farther away a target task is from next-word prediction, the less likely it is to succeed. In contrast, a 'task-free' acquisition procedure (e.g. the integration of nodes into a semantic network) allows graph-based models to remain flexible and uncommitted about how the knowledge encoded in the network is potentially used in the future.

In the same set of experiments that compared the Transformer to the RNN, mentioned in the previous section, Shufan Mao and I also examined the abilities of a novel graph-based model, which we call the Constituent Tree Network (CTN). We observed that both the Transformer and RNN achieved lower performance relative to the CTN, and concluded that compositional generalization does not readily emerge in connectionist

[5] The reason that Transformer-based language models are typically trained on orders of magnitude more language data is because self-attention is computationally more efficient than recurrence

[6] Shufan Mao is a colleague at the UIUC Learning and Language Lab.

models (Mao et al., 2022). We showed that the CTN achieved ceiling performance on all tasks, including a challenging compositional generalization task. There are many potential reasons why the connectionist models did not succeed in the most difficult generalization portion of our task. First, more sophisticated architectures may be needed to more strongly promote the emergence of similarity relations between themes. Second, it is possible that training on more naturalistic data would better promote generalization compared to our carefully balanced artificial dataset. On the other hand, these concerns do not apply to the CTN, which does not require any hyper-parameter tuning. Compositional generalization is built into the architecture of the CTN even prior to the start of training. The CTN succeeded in the experimental condition due to its representational substrate: its edges represent constituency relations, which explicitly constrain the spreading of activation in accordance with constituent structure. More generally, the CTN cleanly separates structure and function. The formation of the network structure — joining parse-trees — is completely independent of the spreading-activation algorithm used to compute relatedness. This sharp distinction between training and inference, and structure and function, is absent in many contemporary neural networks, where the task used during training (e.g. next-word prediction) constrains the kinds of tasks that can be used during inference.

Bayesian and Rule-based Alternatives

There are many other architectures and systems that are potentially more suitable candidates for modeling the distributional construction of lexical semantic representations from children's language input. In particular, Bayesian and rule-based approaches stand out. The advantage of Bayesian models is that they allow the experimenter to precisely tune the expectations of the model, in the forms of priors. Priors can be used to build in all sorts of expectations and inductive biases, and titrate precisely which hypotheses are most useful to rely on when faced with high uncertainty in the data (Tenenbaum et al., 2006). An additional advantage of Bayesian models is that they can track and update multiple hypotheses simultaneously. This enables researchers to examine the influence of a set of starting hypotheses prior to start of training. While Bayesian models do this by design, the The RNN does not explicitly track multiple distinct hypotheses; instead, it tends to average multiple competing hypothesis, which makes them difficult to disentangle and revise.

Work by Fremann and Lapata (2016) showed that an incremental Bayesian model trained to categorize lexical items in child-directed input arrives at a better clustering of lexical items when it is allowed to entertain multiple hypotheses during learning. Armed with multiple hypotheses, an incremental learner is less likely to be forced to re-organize previously acquired knowledge, as it is possible to discard unlikely hypotheses without loss of performance or the need for re-organization. It would be useful for the RNN to be able to react to counter-evidence not by having to re-reorganize existing knowledge whenever it is encountered, but by delaying re-organization until some confidence threshold is reached for how to update existing knowledge. This would be especially useful when faced with uncertainty created by redundancy in the data: Which of multiple signals is most diagnostic of category membership? Entertaining multiple hypotheses at once and the ability to explicitly select between competing hypotheses is a direction worth further research.

More broadly, in traditional symbolic systems, the meanings of words is strictly separated from its syntactic context. In such systems, where content is neatly separated from rules, maladaptive semantic property inheritance (Chapter 6) would not be an issue, as it is in the RNN. In such systems, lexical semantic representations do not modify each other as a result of sequential processing; instead, content is preserved by successive application of rules (Fodor & Pylyshyn, 1988). While such an architecture would have certain advantages over the RNN when generalizing beyond the training distribution, there is controversy surrounding the question whether rules can be learned at all, and whether children's learning can actually be accounted for by rule-learning, or other more heuristic approaches (Bowerman & Choi, 2001; J. L. Elman, 2011; Ervin-Tripp, 1973; Tomasello, 2005). The absence of traditionally envisioned rules in the RNN is not necessarily a limitation of the RNN. Instead, the RNN presents an opportunity to examine how well human cognitive abilities can be approximated by systems without built-in primitives.

I look forward to more research pushing the boundaries of what neural networks can do without domain-specific constraints on their representations (e.g. syntax) or on the computations operating on those representations (e.g. symbolic processing, context-freeness). While such constraints would circumvent many, but not all, of the shortcomings of the RNN reported here, it is not yet clear whether such constraints would best account for human learning and performance, or whether other, more heuristic, constraints (e.g. age-ordered training) may prove equally or even more successful in this regard.

13.4.5 Behavioral Verification

Ultimately, the goal of this research is to determine the nature of the distributional apparatus children use to support their lexical semantic development, and in particular, whether it resembles the RNN. To make progress, SPIN theory makes clear predictions about the nature of distributional learning in an RNN-like mechanism; if children use a similar mechanism, we should observe similar learning dynamics in children. One of the limitations of this thesis, then, is that no such behavioral studies were conducted. To researchers interested in conducting such studies, there are several points to consider. First, it is important that future studies clearly separate language internal distributional knowledge other kinds of knowledge. Because linguistic and non-linguist factors often *collectively* influence behavioral output, questions about the organization of linguistic units — independent of their relationship with real-world entities — will require careful analysis and control of participants' prior experiences. This will be difficult, and likely requires artificial language learning studies, in which children are both exposed to experimentally controlled sequences of non-sense words, and associations between words and perceptual features. The critical manipulation will be whether properties of the statistical input impact how children map non-sense words onto perceptual features. For an example, see (Lany & Saffran, 2010).

Another potential obstacle for behavioral verification is that it is not clear at what age children would perform distributional extension of category-associate features (DECAF). There is a specific age range at which children would benefit most from DECAF, and behavioral work is needed to hone in on this range. There are two considerations that should narrow the search: First, children have to be old enough to know a good number of word meanings, and have a basic understanding of how word meanings are related taxonomically, so that they can extend semantic category-associated features to novel words. Second, children have to be young enough so that DECAF is not yet replaced by more sophisticated experiential knowledge that would likely prove more effective than language-internal distributional information alone.

Further, a key assumption of this thesis is that children are able to separate semantic features that are unique to a particular kind from those that diagnose membership in a larger semantic category (e.g. dog vs. ANIMAL). In order to perform DECAF, children would only extend those features for distributionally similar words that are related to the inferred semantic category of a novel word. In the absence of perceptual information about the novel word, there would be no information that would license the extension of more specific features about kinds. But, this does not mean that children might not attempt to do so, or that it would not be useful to do so. There are plenty of words that synonymous in meaning (e.g. *cat*, *kitty*, *kitten*), and extension of kind-related features (perceptual features about cats, rather than features diagnostic of ANIMAL) would be useful in such situations. But this raises the question how children might know when to extend features based on kind or category. Whether children are able to separate features in this manner, and how to model this extension is an important future direction.

Another potential line of follow-up concerns the potential for distributional similarity to pick out semantic category differences between nouns referring to abstract and/or less imageable concepts (e.g. *idea*, *heat*, *work*). In this work, I have primarily investigated the similarity structure that the RNN learns about common concrete nouns, such as members of ANIMAL and VEHICLE. But distributional information is not limited to these words. An initial distributionally-licensed categorization would like be useful to children by suggesting an initial ballpark meaning for abstract words. Language-internal distributional information is uniquely posed to facilitate children's discovery of the meanings of abstract words.

13.4.6 Additional Corpora

If the explanation of the age-order effect that is provided by SPIN theory is correct, then similar scaffolding effects due to staged training should be observable when training on other corpora that exhibit similar statistical properties as AO-CHILDES. Of course, there are not many corpora that exhibit longitudinal structure; instead, many existing corporate used in computational linguistics are purposefully stationary. For example, most Wikipedia corpora do not exhibit longitudinal structure because Wikipedia articles are inherently assorted.

That said, I have identified one exception: The Newsela corpus [7] is a collection of 1, 911 English and Spanish news articles, and for each article there exist 4 or 5 simplified versions, rewritten by professional annotators for children with different reading proficiency. I extracted only English text, and ordered the articles by simplification level, from high to low (mapping on to grades 2 through 12). At each level, I obtained approximately 1M words, roughly equal in size to AO-CHILDES. I refer to the resulting ordered corpus as AO-Newsela, to emphasize its longitudinal structure. To confirm that the resulting corpus exhibits longitudinal structure, I computed the Taylor exponent in the same manner as explained in Chapter 2. The results, shown in Chapter 2 Table 2.3, confirm that the corpus has longitudinal structure similar to AO-CHILDES.

In contrast to the AO-CHILDES corpus which covers input to children between the age of 1 to 6 years, the AO-Newsela corpus is directed at older children, and teenagers, between the ages of 6-18. Further, because it is written by professional writers for educational purpose rather than by caregivers with the goal of communicating with their children, the AO-Newsela corpus offers an opportunity to disentangle several factors not possible when studying only a single corpus such as AO-CHILDES. For instance, it would be useful to know whether the age-related changes in language statistics in child-directed input are due to special properties related to the caregiver-child milieu, or whether these properties can be reproduced, implicitly or explicitly, by professional writers, who have different teaching goals? In essence, are the age-related changes in AO-CHILDES highly specialized to the early language learning environment, or could they be accounted for by a more domain-general perspective — that language simplification produces highly similar results no matter who is doing it, and who the target audience is? Second, the distribution of topics is almost perfectly uncorrelated with simplification level (i.e. the same articles are discussed at each level), which presents an opportunity to examine weather a potential shift in the distribution of topics that children are exposed to is necessary to account for the age-order effect. If an age-order-like effect is observed when training

[7] Newsela. (2016). Newsela Article Corpus. Version: 2016-01-29, https://newsela.com/data. Newsela is also the name of the company, which produces reading materials for pre-college classroom use.

incrementally on AO-Newsela, this would be support for the notion that a topical shift is not necessary.

Preliminary analyses show that training incrementally on AO-Newsela (using the same RNN and hyper-parameters used to train on AO-CHILDES) does not reliably produce an age-order-like effect. There is a small, noticeable difference in balanced accuracy at the end of training, but this difference is not robust against differences in hyperparameters, and did not replicate with a different choice of probe words (700 nouns specifically chosen from frequent nouns in AO-Newsela). This result leaves many of the questions posed above unanswered. However, this result does provide preliminary support for the idea that when professional editors simplify written language, they likely do not make similar linguistic choices as caregivers communicating with their children. If so, this would suggest that adult intuitions about what 'simple' means in the context of language teaching differs depending on one's goal. The goal of the editors of the Newsela articles is to facilitate reading comprehension, which is not likely on the minds of caregivers of pre-school children. It is plausible that these two goals require choices that are at odds with or do not overlap with each other. More research on spoken and written language simplification is needed to understand such potential trade-offs, and how to optimize learning at different developmental stages.

13.4.7 Distinguishing Syntax and Semantics in the Distributional Realm

It appears as if the work presented herein is arriving at a distinction between syntactic and semantic lexical categories, and that the presence of the former is necessary to guide and constrain acquisition of the latter. More precisely, in my simulations of acquisition of lexical semantic knowledge from child-directed input, I emphasized the need to separate learning into two stages, one which promotes a "syntactic scaffolding" where semantically uninformative left-contexts that cue the NOUN category are encoded in one set of units, and a second stage in which semantically informative left-contexts are encoded in an orthogonal set of units.

Given this stage-like separation, the reader might wonder: Is the age-order effect a semantic or syntactic effect? Is the scaffolding provided by the structure of input to younger compared to older children really a fact about the syntax of caregiver language? My corpus analyses of AO-CHILDES suggest that both semantic and syntactic factors are likely involved. For instance, lexical diversity (a semantic factor), function word density (a syntactic factor), and mean utterance length (a mix of semantic and syntactic factors) all increase across age-ordered partitions of AO-CHILDES. The language acquisition literature also supports this view: While there are plenty of age-related changes in syntactic complexity as children grow older (Newport et al., 1977), more recent work has documented age-related changes in the semantic organization of speech to children and which influence what kind of information is encoded in distributional models (Jiang et al., 2020b). All of these factors may contribute in complicated and non-intuitive ways to the increase in age-related fragmentation of lexical categories. Moreover, I think that a principled distinction makes little sense in the realm of distributional modeling. Most distributional semantic models, including the RNN, treat co-occurrence relationships as statistical relationships, and do not actively attempt to recover underlying type distinctions.

Thus, questions about what whether the knowledge the RNN acquires neatly distinguishes syntactic from lexical semantic phenomena does not appear to lead to much insight. I think that a more important distinction for understanding the learning dynamics of the RNN is the difference between superordinate and subordinate category membership. Too much distributional evidence in favor of subordinate category distinctions obscure important superordinate category distinctions and disorganize the network's learned representational landscape. This distinction approximately maps onto the distinction between syntactic and semantic categories, because we typically think of semantic categories within a given syntactic category, but need perfectly capture this distinction. This is in agreement with evidence from neuroimaging work, which has shown time and again that “no brain region is selectively sensitive to only lexical semantic information or only syntactic information” (Fedorenko et al., 2012).

Chapter 14

Conclusion

Overall, the work presented in this thesis demonstrates that a distributional learning algorithm based on minimizing next-word prediction error, and trained on a representative dataset of transcribed speech to children in a cognitively plausible manner, can acquire lexical semantic representations that can be used to group common nouns into fine-grained semantic categories. This distributional semantic information would be useful for supporting children's inferences about novel word meanings in the absence of extra-linguistic referential information. It is unlikely that children simply ignore occurrences of novel words when they are unable to map them onto perceptually available semantic features. Instead, this thesis shows that children can leverage language-internal distributional cues to infer semantic features related to the semantic category of a novel word. This ability consists of (i) retrieving known words that are distributionally similar to the novel word, (ii) isolating shared (i.e. category-associated) semantic features of the retrieved words, and (iii) extending category-associated features to the novel word. I have termed this the distributionally-mediated extension of category-associated features (DECAF).

Moreover, this thesis provides novel insights into how the structure of co-occurrence data — and the order in which it is presented — influences how lexical semantic information is encoded in the recurrent neural network (RNN) language model. Despite the success of the RNN in this regard, this thesis also demonstrates the pitfalls of encoding what many would consider sparse and context-independent knowledge in a connectionist architecture: The absence of principled, top-down (linguistic) constraints on the organization of lexical semantic representations makes the simple RNN vulnerable to encoding idiosyncratic lexical dependencies that are not relevant for diagnosing semantic category membership (e.g. a washing machine is an APPLIANCE, whether it is white or not). Once such dependencies are encoded in the RNN, they are difficult to unlearn, and may negatively impact the distributional construction of semantic category clusters in the long-term.

One of the obstacles of learning lexical semantic representations that are useful for performing DECAF is the tendency of the RNN to overlook the atomic nature of lexical units, and to acquire chunk-level knowledge — the integration of partial distributional semantic information across two or more sequentially occurring lexical units — instead. One negative consequence of this tendency is that the resulting lexical semantic representation are highly inter-connected with the information already present in the hidden layer. While this is, no doubt, integral to the network's success in next-word prediction, this tight integration between lexical representations and processing dynamics produces word-level knowledge that is strongly tied to the specific arrangement of sequences in which those words occurred.

That said, whether the word-level knowledge learned by the RNN can be used in downstream lexical tasks has (to the best of my knowledge) not previously been investigated. In particular, I asked whether lexical representations that emerge over the course of predicting next-words in child-directed language encode sufficient distributional semantic information to help children infer the semantic category of novel words. The results of computational simulations in Chapters 6 and 10 demonstrate that this is the case. Although some semantic information is lost to the highly contextual processing dynamics at the hidden layer of the RNN, most of the category-relevant distributional semantic information eventually finds its way to the bottom of the network, where it can be accessed in a cognitively plausible manner.

In contrast to many other works in distributional semantic modeling, my primary objective is to demonstrate the feasibility of statistical learning for *extending*, rather than *constructing* semantic knowledge. While I agree with many previous scholars that language-internal distributional information can itself be considered a direct source of word meaning (Dingemanse et al., 2015; Harris, 1954; Lupyan & Lewis, 2019), the possibility that the same statistical information could be used as a tool to extend previously acquired word meanings has received considerably less attention (but see Borovsky and Elman, 2006; Lany and Saffran, 2010), and given the findings presented in this thesis, warrants further study.

A third contribution of this thesis is the investigation of the learning dynamics of the RNN across training. If the appropriate scaffolding is not provided to the RNN during early training (see Chapter 9), the RNN may be guided towards sub-optimal processing dynamics that negatively impact learning and generalization, as evidenced in a downstream semantic categorization task (Chapter 10). One way to illustrate the importance of skeletal structure in the RNN is by analogy with the construction of a toy puzzle: Because there are likely some areas in the puzzle where pieces have fewer degrees of freedom, a good strategy would be to complete such sections as early as possible. For instance, pieces on the outer border of the puzzle are most constrained in the orientation in which they can be assembled, and their assembly allows the player to construct a bird's eye view of the problem at hand. With the outer border constructed, further assembly can proceed in a more organized fashion, moving inwards from the border. This strategy enables the player to capitalize on existing patterns to accelerate progress in more difficult sections.

I have argued that a distributional learning system such as the RNN can benefit from a similar strategy. At the start of training, a randomly initialized RNN does not posses linguistic knowledge useful for filtering and/or organizing the information it receives. The network is using trial-and-error to fit individual pieces of a high-dimensional puzzle. The RNN may later discover counter-evidence and dismantle part of the patterns it has already assembled. This undoing and/or re-organization of previously acquired knowledge can have long-term negative consequences. I showed that this can, in part, be avoided by promoting the formation of superordinate category structure in the RNN — not unlike the border pieces of a toy puzzle. By training on data that promotes the formation of the noun category first, it is possible to build high-level expectations into the network for how to organize subsequent information. These expectations can reduce the likelihood of future mismatches between existing knowledge and novel information, which might otherwise over-write existing knowledge. In sum, encoding superordinate category structure early during training can provide a useful landmark for navigating the complex world of distributional statistics in natural language. It is likely that this principle extends to other domains where self-supervised learning algorithms are used.

What makes the age-order effect a fascinating topic of inquiry is that the RNN — an artificial learning system —- benefits from training on input in the order in which children are exposed to language across developmental time. This link between the RNN and children's longitudinally structured language environment suggests — albeit speculatively — that the distributional learning capabilities in children and the RNN may be governed by similar principles. Moreover, the age-order effect is a reminder of the important role that early experiences have on subsequent learning experiences. While this is well known in the developmental literature, it is not often discussed by neural network researchers, who tend to train their models non-incrementally and on randomly order data. However, as my corpus analyses revealed, language directed to children is not stationary, and researchers using neural networks to model language acquisition should take note.

Further, this thesis draws attention to limitations of next-word prediction in the absence of top-down constraints provided by linguistic and/or experimental knowledge. Without yet knowing which lexical associations might be relevant (i.e meaningful) in the long-term, this raises the age-old chicken-vs-egg problem: How does learning work when we don't yet know what statistical associations are actually diagnostic of the knowledge we wish to acquire? The response given by proponents of distributional learning is 'incidental learning' — everything goes. Information that eventually turns out not to be useful is simply discarded when a sufficient confidence level is reached to warrant such a move. While this argument is tempting, the findings presented herein show that the story is more complicated. It *matters* when (and that) irrelevant associations are learned, because every observation has the potential to influence the long-term dynamics and preferences of the RNN.[1] Fortunately, for proponents of connectionism, and others interested in the role of connectionist networks as cognitive models of aspects of language learning, the current work offers a straightforward response this criticism: If the order in which data is presented to a distributional learner matters, then the burden is on caregivers, the developing brain/mind, and the environment to play their part in constraining the statistical associations on offer to language learning children. A similar conclusion has been drawn by McClelland and Rogers (2003) who have proposed a connectionist and emergentist theory of semantic cognition, where the role of the input, as opposed to innate conceptual filters, plays a central role. The work in this thesis is broadly aligned with such a view, namely that the input does much of the work of differentiating what statistical associations are relevant and which are not.

It if turns out that distributional learning in children is dissimilar to how recurrent neural networks learn, the theory and insights presented in this thesis would still be relevant to research in machine learning. At its core, this thesis is an attempt to unravel the 'black box' of the RNN and to discover simple principles useful to anyone who wishes to understand artificial neural networks better. My own motivation is not only scientific in nature, but also to drive innovation in engineering and practical application, such as modeling of discrete time-series data, and natural language processing. In sum, I hope that the findings presented herein will be useful not only to scientists working in language acquisition research, but also to machine learning researchers interested in better understanding the fascinating and complex learning dynamics of artificial neural networks.

[1] This is presumably also the case for most connectionist architectures not explicitly designed to address this shortcoming.

Appendix A

Target Semantic Category Structure

The set of probe words used to examine the RNN's learned lexical semantic knowledge were obtained from a list of the most frequent common nouns in English child-directed input, retrieved from the CHILDES database (MacWhinney, 2000; Sanchez et al., 2019). The procedure used during collection and semantic categorization was previously reported by P. A. Huebner and Willits (2018). Briefly, the authors (1) collected all word forms in AO-CHILDES which could be nouns (even if, in practice they appear more often in verb form, such as *jump*), (2) choosing the subset of those that refer to a concrete object, and (3) choosing the subset of those that unambiguously belong to a semantic category which contains at least six other words, according to a set of human raters. For example, *apple*, *orange*, and *banana* were included because they belonged to a large category of words with at least six members. Note, that probes in the NUMBER category were excluded in all analyses reported in this thesis. The reason is that number words are more often used as determiners rather than nouns. The full list of probe words and their semantic category label are shown below, in Tables A.1, A.2, and A.3.

It is important to examine how probe words are distributed over transcripts belonging to different age bins. If the distribution is not close to uniform, this would be an important consideration when interpreting analyses of how semantic category structure might vary with age. The distribution is shown in Figure A.1. Splitting based on concreteness rather than by semantic category was done to prevent cluttering of the figure. For example, categories like MAMMAL, INSECT, FURNITURE, MEAT were classified as concrete, and categories like WEATHER and DAY, were classified as abstract based on the intuition of the authors. Importantly, in correspondence with greater noun density in input to younger children (green), probe words occur most frequently in early age bins, peaking between days 400-600. Interestingly, the ratio of abstract (blue) versus concrete (orange) probe words steadily increases with age.

SPORT	TIME	WEATHER	MONTH	SHAPE	SPACE	INSTRUMENT	PLANT	DRINK
baseball	afternoon	cloud	april	circle	earth	accordion	acorn	beer
basketball	daytime	fog	august	cone	jupiter	banjo	bush	cider
game	hour	ice	december	cube	mars	bell	cactus	cocacola
golf	midnight	lightning	february	curve	mercury	clarinet	chestnut	cocoa
soccer	minute	rain	january	diamond	moon	drum	daisy	coffee
softball	morning	rainbow	july	line	neptune	flute	flower	coke
volleyball	night	snow	june	loop	planet	guitar	grass	cola
	noon	storm	march	octagon	pluto	harmonica	lily	drink
	second	thunder	month	rectangle	saturn	instrument	oak	juice
	sunset	tornado	november	shape	space	music	pine	koolaid
		weather	october	square	star	orchestra	plant	lemonade
		wind	september	trapezoid	sun	piano	seaweed	milk
				triangle	uranus	trumpet	sunflower	pepsi
					venus	tuba	tree	pop
					world	violin	tulip	soda
						xylophone	violet	tea
								wine

Table A.1: Target semantic category structure used to evaluate contextualized and non-contextualized lexical semantic representations learned by RNNs trained on AO-CHILDES.

MEAT	INSECT	DESSERT	ELECTRIC	VEGGIE	BATH	TOOL	BIRD	FURNITURE
bacon	ant	brownie	alarm	asparagus	bathtub	axe	bird	bed
beef	bee	cake	battery	broccoli	brush	broom	bluebird	bench
bologna	bug	candy	cd	cabbage	comb	drill	cardinal	blanket
burger	butterfly	chocolate	calculator	carrot	kleenex	hammer	chick	bookcase
drumstick	caterpillar	cookie	camera	cauliflower	lipstick	iron	cock	bookshelf
fish	cricket	cream	clock	celery	lotion	ladder	crow	cabinet
flounder	dragonfly	cupcake	computer	cucumber	makeup	lawnmower	cuckoo	candle
ham	flea	custard	dvd	lettuce	pee	mop	dove	carpet
hamburger	fly	dessert	earphone	mushroom	poo	needle	duck	chair
meat	grasshopper	donut	headphones	olive	poop	pail	duckling	couch
pepperoni	insect	fudge	microphone	onion	potty	paint	eagle	crib
pork	ladybug	lollipop	microscope	parsley	shampoo	paintbrush	flamingo	cushion
roast	mosquito	m&m	phone	pea	shit	pen	goose	desk
salami	scorpion	marshmallow	radio	pepper	shower	pencil	hawk	drawer
salmon	snail	oreo	register	pickle	soap	plier	hen	dresser
sausage	spider	pie	stereo	potato	sponge	plow	ostrich	dryer
steak	wasp	popsicle	tv	pumpkin	tissue	rake	owl	furniture
tuna	worm	pudding	telephone	salad	toilet	scissor	parrot	grill
		sweet	telescope	spinach	toothbrush	screw	peacock	lamp
		tapioca	television	turnip	toothpaste	screwdriver	penguin	pillow
		treat	video	vegetable	towel	shovel	pigeon	quilt
				zucchini	tub	tape	rooster	seat
						tool	seagull	shelf
						umbrella	sparrow	sofa
						vacuum	stork	stool
						wheelbarrow	swan	table
						wrench	vulture	washer
							woodpecker	wastebasket

Table A.2: (continued from table above)

FRUIT	KITCHEN	HOUSEHOLD	TOY	FAMILY	VEHICLE	CLOTHING	BODY	MAMMAL
apple	blender	backyard	ball	aunt	airplane	apron	ankle	armadillo
apricot	bowl	basement	balloon	babe	ambulance	bathrobe	arm	baboon
avocado	colander	bathroom	bat	boy	ark	belt	ass	badger
banana	cup	bedroom	block	brother	bicycle	bib	beard	bear
berry	dish	ceiling	book	child	bike	blouse	belly	beaver
blueberry	dishwasher	cellar	chess	cousin	boat	bonnet	blood	buck
cantaloupe	drain	chimney	comic	dad	bulldozer	boot	body	buffalo
cherry	fork	closet	crayon	daughter	bus	bra	bone	bull
coconut	freezer	counter	die	family	cab	buckle	bottom	bunny
cranberry	fridge	curtain	doll	father	caboose	cap	braid	camel
fruit	glass	den	gijoe	girl	canoe	cape	brain	cat
grape	knife	door	gift	gran	car	clothes	breast	cheetah
grapefruit	microwave	driveway	glitter	granddad	carriage	clothing	bruise	chimpanzee
lemon	mixer	fence	kite	grandfather	cart	coat	butt	chipmunk
lime	napkin	floor	lego	grandma	dumptruck	diaper	cheek	collie
mango	oven	garage	marble	grandmother	helicopter	dress	chest	cow
melon	pan	hallway	playdoh	grandpa	jeep	glove	chin	coyote
orange	pitcher	kitchen	puppet	infant	jet	gown	ear	deer
peach	plate	lawn	puzzle	kid	motorcycle	hat	elbow	dingo
pear	refrigerator	nursery	racket	ma	pickup	helmet	eye	dog
pineapple	saucer	patio	rattle	mama	plane	hood	eyeball	dolphin
plum	silverware	porch	seesaw	mom	rocket	jacket	eyebrow	donkey
raisin	spatula	roof	skate	mother	scooter	mitten	face	elephant
raspberry	spoon	room	skateboard	nephew	ship	nightgown	finger	fox
strawberry	stove	sandbox	sled	newborn	shuttle	outfit	fingernail	giraffe
tangerine	teapot	shed	sport	niece	spaceship	pajamas	foot	goat
tomato	teaspoon	stair	sticker	pa	stroller	pant	forehead	gorilla
watermelon	toaster	step	teddy	papa	submarine	purse	hair	groundhog
	whisk	study	tennis	parent	subway	robe	hand	hamster
		wall	toy	pet	taxi	sandal	head	hare
		window	tricycle	sis	tractor	scarf	heart	hedgehog
		yard	yoyo	sister	trailer	shirt	hip	hippo
				son	train	shoe	jaw	horse
				stepmother	truck	shoelace	knee	hyena
					van	short	lap	jaguar
					wagon	skirt	leg	kangaroo
						slack	lip	kitten
						slipper	liver	koala
						sneaker	memory	lamb
						sock	mind	leopard
						suit	mood	lion
						sweater	mouth	mammal
						sweatshirt	muscle	mammoth
						tie	mustache	mole
						underwear	neck	monkey
						uniform	nipple	moose
						vest	nose	mouse
							penis	mule
							ponytail	opossum
							ribs	orangutan
							shoulder	otter
							skin	ox
							skull	panda
							stomach	panther
							throat	pig
							thumb	pony
							toe	porcupine
							tongue	pup
							tooth	rabbit
							tummy	raccoon
							vagina	rat
							waist	reindeer
							weener	rhino
							wrist	seal
								sheep
								skunk
								squirrel
								steer
								tiger
								walrus
								weasel
								whale
								wolf
								zebra

Table A.3: (continued from table above)

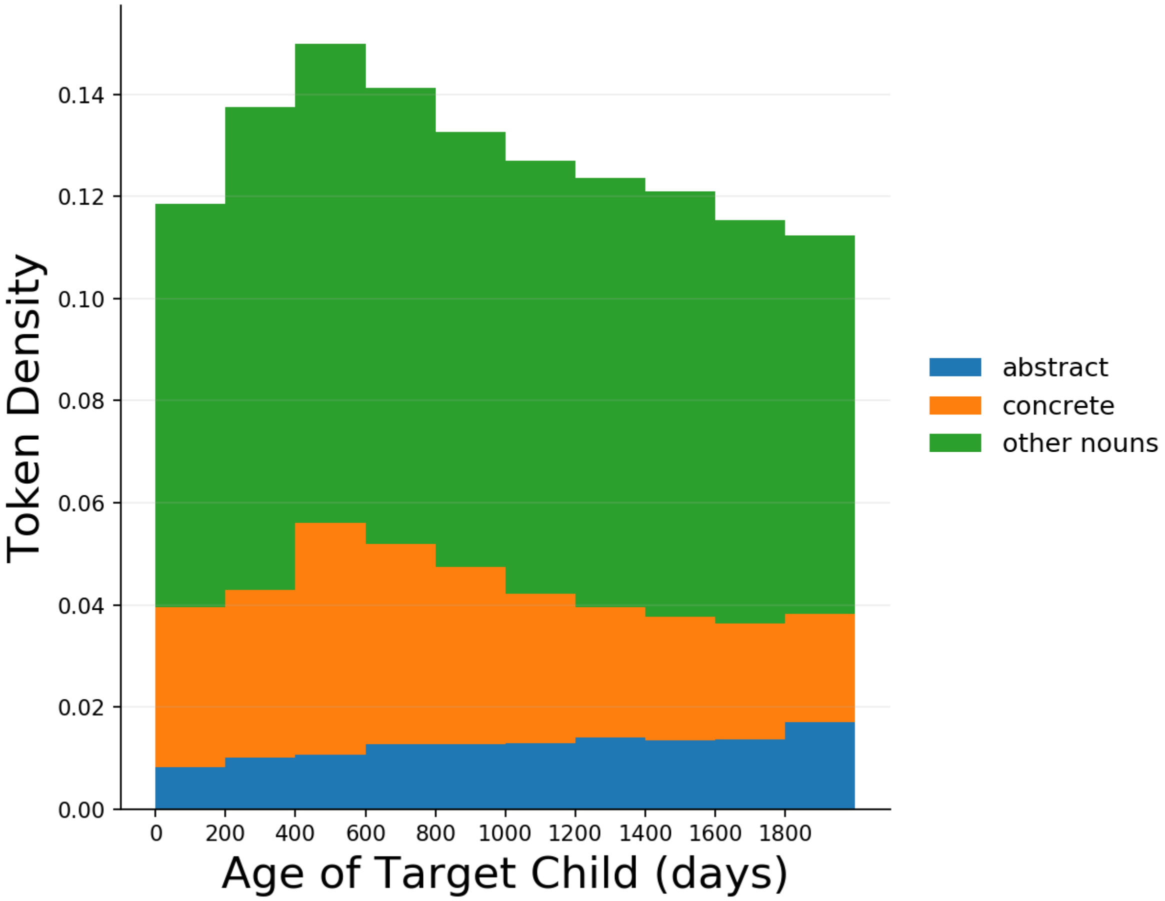


Figure A.1: The distribution of probe words (orange and blue) and all other nouns (green) over consecutive age bins in AO-CHILDES. Token density is the total frequency of words in a given age bin normalized by the total number of words in the same bin. Probe words are marked as abstract or concrete. Interestingly, only concrete probe word density peaks at 400-600 days. This distinction is for reference only, and is not used in this thesis.

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *Linguistic Review*, *23*(3), 275–290.

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition*, *164*, 116–143.

Achille, A., Rovere, M., & Soatto, S. (2018). Critical learning periods in deep networks. *International Conference on Learning Representations*.

Aitchison, L., & Lengyel, M. (2017). With or without you: Predictive coding and bayesian inference in the brain. *Current Opinion in Neurobiology*, *46*, 219–227.

Ali, A., Ahmad, N., de Groot, E., van Gerven, M. A. J., & Kietzmann, T. C. (2021). Predictive coding is a consequence of energy efficiency in recurrent neural networks. *bioRxiv Preprint*.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*(4), 583–609.

Altmann, G. T. (1998). Ambiguity in sentence processing. *Trends in Cognitive Sciences*, *2*(4), 146–152.

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*(2), 239–273.

Amenta, S., Günther, F., & Marelli, M. (2020). A (distributional) semantic perspective on the processing of morphologically complex words. *The Mental Lexicon*, *15*(1), 62–78.

Amos, R. M., Seeber, K. G., & Pickering, M. J. (2022). Prediction during simultaneous interpreting: Evidence from the visual-world paradigm. *Cognition*, *220*.

Anderson, N. D., & Dell, G. S. (2018). The role of consolidation in learning context-dependent phonotactic patterns in speech and digital sequence production. *Proceedings of the National Academy of Sciences*, *115*(14), 3617–3622.

Anderson, N. D., Holmes, E. W., Dell, G. S., & Middleton, E. L. (2019a). Reversal shift in phonotactic learning during language production: Evidence for incremental learning. *Journal of Memory and Language*, *106*, 135–149.

Anderson, N. D., Holmes, E. W., Dell, G. S., & Middleton, E. L. (2019b). Reversal shift in phonotactic learning during language production: Evidence for incremental learning. *Journal of Memory and Language*, *106*, 135–149.

Andhale, N., & Bewoor, L. (2016). An overview of text summarization techniques. *2016 international conference on computing communication control and automation (ICCUBEA)*, 1–7.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological review*, *116*(3), 463–498.

Arbib, M. A., & Érdi, P. (2000). Précis of neural organization: Structure, function, and dynamics. *Behavioral and Brain Sciences*, *23*(4), 513–533.

Arnon, I., & Christiansen, M. H. (2017). The role of multiword building blocks in explaining L1-L2 differences. *Topics in Cognitive Science*, *9*(3), 621–636.

Arnon, I., & Clark, E. V. (2011). Why brush your teeth is better than teeth – children's word production is facilitated in familiar Sentence-Frames. *Langue Learning and Development*, *7*(2), 107–129.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324.

Asr, F. T., Willits, J., & Jones, M. N. (2016). Comparing predictive and co-occurrence based models of lexical semantics trained on child-directed speech. *CogSci*.

Au, T. K.-F., & Markman, E. M. (1987). Acquiring word meanings via linguistic contrast. *Cognitive Develoment*, *2*(3), 217–236.

Baillargoen, R. (1993). The object concept revisited: New directions in the investigations of infants' physical knowledge. In C. Granrud (Ed.), *Visual perception and cognition in infancy*. L. Erlbaum Associates.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*(3), 241–248.

Baroni, M. (2020). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, *375*(1791), 20190307.

Barrett, M., Bingel, J., Hollenstein, N., Rei, M., & Søgaard, A. (2018). Sequence classification with human attention. *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 302–312.

Barsalou, L. W. et al. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*(4), 577–660.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, *37*(6), 1554–1563.

Bender, E. M., & Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. *Proceedings of ACL*.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, *3*(Feb), 1137–1155.

Bengio, Y., & Frasconi, P. (1993). Credit assignment through time: Alternatives to backpropagation. *Advances in neural information processing systems*, *6*.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. *Proceedings of the 26th Annual International Conference on Machine Learning*, 41–48.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.

Berman, R. A. (1988). Word class distinctions in developing grammars. *Categories and processes in language acquisition*, 45–72.

Berwick, R. C. (2014). Learning word meanings from examples. *Semantic Structures*, 89–124.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by english-speaking adults and infants. *Journal of experimental psychology: human perception and performance*, *14*(3), 345.

Bliss, D. P., Sun, J. J., & D'Esposito, M. (2017). Serial dependence is absent at the time of perception but increases in visual working memory. *Scientific reports*, *7*(1), 1–13.

Bloom, L., Tinker, E., & Margulis, C. (1993). The words children learn: Evidence against a noun bias in early vocabularies. *Cognitive development*, *8*(4), 431–450.

Bloom, P., & Kelemen, D. (1995). Syntactic cues in the acquisition of collective nouns. *Cognition*, *56*(1), 1–30.

Bloom, P., & Markson, L. (1998). Capacities underlying word learning. *Trends in Cognitive Sciences*, *2*(2), 67–73.

Boland, J. E., & Tanenhaus, M. K. (1991). The role of lexical representations in sentence processing. *Advances in psychology* (pp. 331–366). Elsevier.

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annu. Rev. Appl. Linguist.*

Booth, A. E., & Waxman, S. R. (2009). A horse of a different color: Specifying with precision infants' mappings of novel nouns and adjectives. *Child Dev.*, *80*(1), 15–22.

Borovsky, A., & Elman, J. (2006). Language input and semantic categories: A relation between cognition and early word learning. *Journal of child language*, *33*(4), 759–790.

Bower, G. H. (1970). Organizational factors in memory. *Cognitive psychology*, *1*(1), 18–46.

Bowerman, M. (1973). Structural relationships in children's utterances: Syntactic or semantic? In T. E. Moore (Ed.), *Cognitive development and acquisition of language* (pp. 197–213). Academic Press.

Bowerman, M., & Choi, S. (2001). 16 shaping meanings for language: Universal and language-specific in the acquisition of spatial. *Language acquisition and conceptual development*, *3*, 475.

Brandt, S., Diessel, H., & Tomasello, M. (2008). The acquisition of german relative clauses: A case study. *Journal of Child Language*, *35*(2), 325–348.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–44.

Broen, P. A. (1972). *The verbal environment of the language-learning child. asha monographs, no. 17.* Eric.

Brooks, R., & Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental science*, *8*(6), 535–543.

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.

Brown, R., & Berko, J. (1960). Word association and the acquisition of grammar. *Child Dev.*, *31*, 1–14.

Brown, R. W. (1957). Linguistic determinism and the part of speech. *J. Abnorm. Psychol.*, *55*(1), 1–5.

Brown, R. (2013). *A first language.* Harvard University Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are Few-Shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 1877–1901). Curran Associates, Inc.

Bruner, J. (1984). Vygotsky's zone of proximal development: The hidden agenda. *New directions for child development.*

Brusini, P., Seminck, O., Amsili, P., & Christophe, A. (2021). The acquisition of noun and verb categories by bootstrapping from a few known words: A computational model. *Frontiers in Psychology.*

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behav. Res. Methods*, *39*(3), 510–526.

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behav. Res. Methods*, *44*(3), 890–907.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, *25*(2-3), 211–257.

Burgess, C., & Lund, K. (1997). Representing abstract words and emotional connotation in a high-dimensional memory space. *Proceedings of the Cognitive Science Society*, 61–66.

Cai, X., Huang, J., Bian, Y., & Church, K. (2020). Isotropy in the contextual embedding space: Clusters and manifolds. *International Conference on Learning Representations*.

Callanan, M. A. (1985). How parents label objects for young children: The role of input in the acquisition of category hierarchies. *Child Development*, 508–523.

Calvo, F., & Colunga, E. (2003). The statistical brain: Reply to marcus' the algebraic mind. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *25*.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27*(6), 843–873.

Carey, S., & Spelke, E. (1994). Domain-specific knowledge and conceptual change. *Mapping the mind: Domain specificity in cognition and culture* (pp. 169–200).

Carta, A., Sperduti, A., & Bacciu, D. (2020). Incremental training of a recurrent neural network exploiting a multi-scale dynamic memory. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 677–693.

Cartwright, T. A., & Brent, M. R. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition*, *63*(2), 121–170.

Cassani, G., Grimm, R., Daelemans, W., & Gillis, S. (2018). Lexical category acquisition is facilitated by uncertainty in distributional co-occurrences. *PLoS One*, *13*(12), e0209449.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4960–4964.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234–272.

Chang, N., Feldman, J., & Narayanan, S. (2005). Structured Connectionist Models Of Language, Cognition And Action. *Modeling language, cognition and action* (pp. 57–67). World Scientific.

Chang, T. A., & Bergen, B. K. (2022). Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, *10*, 1–16.

Chen, H., Zheng, G., & Ji, Y. (2020). Generating hierarchical explanations on text classification via feature interaction detection. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton; Co.

Chomsky, N. et al. (1976). *Reflections on language*. Temple Smith London.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*(2-3), 221–268.

Christiansen, M. H., & Chater, N. (1999a). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*(2), 157–205.

Christiansen, M. H., & Chater, N. (1999b). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*(2), 157–205.

Chrupała, G. (2012). Hierarchical clustering of word class distributions. *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, 100–104.

Clark, A. (1994). Representational trajectories in connectionist learning. *Minds Mach.*, *4*(3), 317–332.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181–204.

Clark, E. V. (1995). *The lexicon in acquisition*. Cambridge University Press.

Clark, E. V., & Garnica, O. K. (1974). Is he coming or going? on the acquisition of deictic verbs. *Journal of verbal learning and verbal behavior*, *13*(5), 559–572.

Clark, E. V., Gelman, S. A., & Lane, N. M. (1985). Compound nouns and category structure in young children. *Child Development*, *56*(1), 84–94.

Clark, E. V., & MacWhinney, B. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition*, 1–33.

Clark, E. V. (2017a). Chapter 16 - semantic categories in acquisition. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science (second edition)* (pp. 397–421). Elsevier.

Clark, E. V. (2017b). Chapter 16 - semantic categories in acquisition. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science (second edition)* (pp. 397–421). Elsevier.

Clark, H. H. (1996). *Using language*. Cambridge university press.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural computation*, *1*(3), 372–381.

Coffey, J. R., Shafto, C. L., Geren, J. C., & Snedeker, J. (2022). The effects of maternal input on language in the absence of genetic confounds: Vocabulary development in internationally adopted children. *Child development*, *93*(1), 237–253.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, *8*(2), 240–247.

Conwell, E., & Morgan, J. L. (2012). Is it a noun or is it a verb? resolving the ambicategoricality problem. *Langue Learning and Development*, *8*(2), 87–112.

Cueva, C. J., & Wei, X.-X. (2018). Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. *International Conference on Learning Representations*.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329.

Davies, C., Lingwood, J., Ivanova, B., & Arunachalam, S. (2021). Three-year-olds' comprehension of contrastive and descriptive adjectives: Evidence for contrastive inference. *Cognition*, *212*(104707).

De Mulder, W., Bethard, S., & Moens, M.-F. (2015). A survey on the application of recurrent neural networks to statistical language modeling. *Computation Speech Language*, *30*(1), 61–98.

De Saussure, F. (1989). *Cours de linguistique générale* (Vol. 1). Otto Harrassowitz Verlag.

Dell, G. S., & Chang, F. (2014). The p-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634).

Dell, G. S., Kelley, A. C., Hwang, S., & Bian, Y. (2021). The adaptable speaker: A theory of implicit learning in language production. *Psychological review*, *128*(3), 446.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological review*, *108*(2), 452–478.

des Combes, R. T., Pezeshki, M., Shabanian, S., Courville, A., & Bengio, Y. (2018). Convergence properties of deep neural networks on separable data. *Unpublished manuscript*.

Dienes, Z., Altmann, G. T., & Gao, S.-J. (1999). Mapping across domains without feedback: A neural network model of transfer of implicit knowledge. *Cognitive Science*, *23*(1), 53–82.

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2017). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism: Language and Cognition*, *20*(5), 917–930.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, *19*(10), 603–615.

Dixon, R. M. (2010). *Where have all the adjectives gone?: And other essays in semantics and syntax* (Vol. 107). Walter de Gruyter.

Dong, S., Gu, Y., & Vigliocco, G. (2021). The impact of child-directed language on children's lexical development. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Doumas, L., & Hummel, J. E. (2005). Approaches to modeling human mental representations: What works, what doesn't and why. *The Cambridge handbook of thinking and reasoning, ed. KJ Holyoak & RG Morrison*, 73–94.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1–43.

Dye, M. (2017). *Bridging levels of analysis: Learning, information theory, and the lexicon* (Doctoral dissertation).

Dye, M., Jones, M. N., Yarlett, D., & Ramscar, M. (2017). Refining the distributional hypothesis: A role for time and context in semantic representation. *CogSci*.

Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. A. (2016). Recurrent neural network grammars. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 199–209.

Edgell, S. E., & Morrissey, J. M. (1987). Delayed exposure to additional relevant information in nonmetric multiple-cue probability learning. *Organ. Behav. Hum. Decis. Processes*, *40*(1), 22–38.

Eisler, Z., Bartos, I., & Kertész, J. (2008). Fluctuation scaling in complex systems: Taylor's law and beyond. *Advances in Physics*, *57*(1), 89–142.

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

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Mach. Learn.*, *7*(2), 195–225.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

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

Elman, J. L. (2011). Lexical knowledge without a lexicon? *Ment. Lex.*, *6*(1), 1–33.

Elman, J. L., & McRae, K. (2019). A model of event knowledge. *Psychological Review*, *126*(2), 252–291.

Erk, K., & Padó, S. (2008). A structured vector space model for word meaning in context. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 897–906.

Ervin-Tripp, S. (1973). Some strategies for the first two years. *Cognitive development and acquisition of language* (pp. 261–286). Elsevier.

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

Falandays, J. B., Nguyen, B., & Spivey, M. J. (2021). Is prediction nothing more than multi-scale pattern completion of the future? *Brain Research*, *1768*.

Farrell, M., Recanatesi, S., Moore, T., Lajoie, G., & Shea-Brown, E. (2019). Recurrent neural networks learn robust representations by dynamically balancing compression and expansion. *bioRxiv*.

Fausey, C. M., Jayaraman, S., & Smith, L. B. (2016). From faces to hands: Changing visual input in the first two years. *Cognition*, *152*, 101–107.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*(6), 1017–1063.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505.

Federmeier, K. D. (2022). Connecting and considering: Electrophysiology provides insights into comprehension. *Psychophysiology*, *59*(1).

Fedorenko, E., Nieto-Castañon, A., & Kanwisher, N. (2012). Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia*, *50*(4), 499–513.

Feijoo, S., Muñoz, C., & Serrat, E. (2015). Morphosyntactic cues to noun categorization in english child-directed speech. *Language Communications*, *45*, 1–11.

Ferguson, B., Graf, E., & Waxman, S. R. (2014). Infants use known verbs to learn novel nouns: Evidence from 15-and 19-month-olds. *Cognition*, *131*(1), 139–146.

Ferguson, B., Graf, E., & Waxman, S. R. (2018). When veps cry: Two-year-olds efficiently learn novel words from linguistic contexts alone. *Language Learning and Development*, *14*(1), 1–12.

Fernald, A., & Morikawa, H. (1993). Common themes and cultural variations in japanese and american mothers' speech to infants. *Child Dev.*, *64*(3), 637–656.

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant behavior and development*, *10*(3), 279–293.

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental psychology*, *27*(2), 209.

Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Dev. Psychol.*, *42*(1), 98–116.

Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, *9*(3), 228–231.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of child language*, *16*(3), 477–501.

Fernald, A., Thorpe, K., & Marchman, V. A. (2010). Blue car, red car: Developing efficiency in online interpretation of adjective–noun phrases. *Cognitive Psychology*, *60*(3), 190–217.

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language. *Developmental psycholinguistics: On-line methods in children's language processing*, *44*, 97–135.

Finn, A. S., & Kam, C. L. H. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition*, *108*(2), 477–499.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neurosci.*, *17*(5), 738–743.

Fisher, C., Gertner, Y., Scott, R. M., & Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(2), 143–149.

Fitz, H., & Chang, F. (2017). Meaningful questions: The acquisition of auxiliary inversion in a connectionist model of sentence production. *Cognition*, *166*, 225–250.

Flege, J. E., & Davidian, R. D. (1984). Transfer and developmental processes in adult foreign language speech production. *Applied psycholinguistics*, *5*(4), 323–347.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Ford, M., Bresnan, J., & Kaplan, R. M. (1982). The mental representation of grammatical relations.

Fornaciai, M., & Park, J. (2020). Attractive serial dependence between memorized stimuli. *Cognition*, *200*.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of verbal learning and verbal behavior*, *12*(6), 627–635.

Fourtassi, A. (2020). Word co-occurrence in Child-Directed speech predicts children's free word associations. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 49–53.

Foushee, R., Griffiths, T., & Srinivasan, M. (2016). Lexical complexity of Child-Directed and overheard speech: Implications for learning. *CogSci*.

Frermann, L., & Lapata, M. (2016). Incremental bayesian category learning from natural language. *Cognitive Science*, *40*(6), 1333–1381.

Freudenthal, D., Pine, J., Jones, G., et al. (2013). Frequent frames, flexible frames and the Noun-Verb asymmetry. *Proceedings of the Annual*.

Freudenthal, D., Pine, J. M., & Gobet, F. (2006). Modeling the development of children's use of optional infinitives in dutch and english using mosaic. *Cognitive Science*, *30*(2), 277–310.

Freudenthal, D., Pine, J. M., Jones, G., & Gobet, F. (2016). Developmentally plausible learning of word categories from distributional statistics. *CogSci*.

Friedrich, M., & Friederici, A. D. (2005). Semantic sentence processing reflected in the event-related potentials of one-and two-year-old children. *Neuroreport*, *16*(16), 1801–1804.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, *360*(1456), 815–836.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

Furrow, D., Nelson, K., & Benedict, H. (1979). Mothers' speech to children and syntactic development: Some simple relationships. *Journal of child language*, *6*(3), 423–442.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 32–42.

Gallaway, C., & Richards, B. J. (1994). *Input and interaction in language acquisition*. Cambridge University Press.

Gelderloos, L., Chrupała, G., & Alishahi, A. (2020). Learning to understand child-directed and adult-directed speech. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Gelman, S. A., & Taylor, M. (1984). How two-year-old children interpret proper and common names for unfamiliar objects. *Child Dev.*, *55*(4), 1535–1540.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Center for the Study of Reading Technical Report; no. 257.*

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Dev.*, *59*(1), 47–59.

Gentner, D., Boroditsky, L., Bowerman, M., & Levinson, S. (2001). Individuation, relativity, and early word. *Language, culture and cognition*, *3*, 215–256.

Gershman, S., & Tenenbaum, J. B. (2015). Phrase similarity in humans and machines. *CogSci.*

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135–176.

Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, *1*(1), 3–55.

Gleitman, L. R., Newport, E. L., & Gleitman, H. (1984). The current status of the motherese hypothesis. *Journal of child language*, *11*(1), 43–79.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, *43*(3), 379–401.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *J. Exp. Psychol. Gen.*, *117*(3), 227–247.

Gluck, M. A., & Rumelhart, D. E. (2013). *Neuroscience and connectionist theory.* Psychology Press.

Gold, E. M. (1967). Language identification in the limit. *Information and control*, *10*(5), 447–474.

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*(2), 125–157.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54.

Golinkoff, R. M. (1975). Semantic development in infants: The concepts of agent and recipient. *Merrill-Palmer Quarterly of Behavior and Development*, *21*(3), 181–193.

Golinkoff, R. M., & Alioto, A. (1995). Infant-directed speech facilitates lexical learning in adults hearing chinese: Implications for language acquisition. *Journal of Child Language*, *22*(3), 703–726.

Golinkoff, R. M., Hirsh-Pasek, K., Bloom, L., Smith, L. B., Woodward, A. L., Akhtar, N., Tomasello, M., & Hollich, G. (2000). *Becoming a word learner: A debate on lexical acquisition.* Oxford University Press.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–135.

Gomez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436.

Gómez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, *7*(2), 183–206.

Goodman, N. (1972). Seven strictures on similarity. *Problems and projects.* Bobs-Merril.

Gopnik, A., & Meltzoff, A. (1987). The development of categorization in the second year and its relation to other cognitive and linguistic developments. *Child development*, 1523–1531.

Gordon, J., Lopez-Paz, D., Baroni, M., & Bouchacourt, D. (2019). Permutation equivariant models for compositional generalization in language. *International Conference on Learning Representations.*

Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access ii. infant data. *Journal of Memory and Language*, *51*(4), 548–567.

Granger, R. (1977). Foul-up. *5-th International Joint Conference on Artificial Intelligence. Cambridge, Massachusetts.*

Graves, A., Bellemare, M. G., Menick, J., Munos, R., & Kavukcuoglu, K. (2017). Automated curriculum learning for neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (pp. 1311–1320). Pmlr.

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE international conference on acoustics, speech and signal processing*, 6645–6649.

Graves, W. W., Binder, J. R., Desai, R. H., Conant, L. L., & Seidenberg, M. S. (2010). Neural correlates of implicit and explicit combinatorial semantic processing. *Neuroimage*, *53*(2), 638–646.

Griffin, S. A., Case, R., & Siegler, R. S. (1994). *Rightstart: Providing the central conceptual prerequisites for first formal learning of arithmetic to students at risk for school failure.* The MIT Press.

Grisoni, L., Miller, T. M., & Pulvermüller, F. (2017). Neural correlates of semantic prediction and resolution in sentence processing. *Journal of Neuroscience*, *37*(18), 4848–4858.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1195–1205.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-Space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspect. Psychological Science*, *14*(6), 1006–1033.

Hamrick, P., Lum, J. A., & Ullman, M. T. (2018). Child first language and adult second language are both tied to general-purpose learning systems. *Proceedings of the National Academy of Sciences*, *115*(7), 1487–1492.

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

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

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children.* Paul H Brookes Publishing.

Hayes, D. P., & Ahrens, M. G. (1988). Vocabulary simplification for children: A special case of 'motherese'? *Journal of child language*, *15*(2), 395–410.

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable modified kneser-ney language model estimation. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 690–696.

Hill, F., Bordes, A., Chopra, S., & Weston, J. (2015). The goldilocks principle: Reading children's books with explicit memory representations. *arXiv Preprint.*

Hill, F., Santoro, A., Barrett, D. G. T., Morcos, A. S., & Lillicrap, T. (2019). Learning to make analogies by contrasting abstract relational structure. *arXiv Preprint.*

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language*, *63*(3), 259–273.

Hinton, G. E. (1990). Connectionist learning procedures. *Machine learning* (pp. 555–610). Elsevier.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hockema, S. A. (2006). Finding words in speech: An investigation of american english. *Language Learning and Development*, *2*(2), 119–146.

Hoff-Ginsberg, E. (1986). Function and structure in maternal speech: Their relation to the child's development of syntax. *Developmental Psychology*, *22*(2), 155–163.

Horst, J. S., Oakes, L. M., & Madole, K. L. (2005). What does it look like and what can it do? category structure influences how infants categorize. *Child Dev.*, *76*(3), 614–631.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570–1582.

Htut, P. M., Cho, K., & Bowman, S. (2018). Grammar induction with neural language models: An unusual replication. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 371–373.

Hubbard, R. J., Rommers, J., Jacobs, C. L., & Federmeier, K. D. (2019). Downstream behavioral and electrophysiological consequences of word prediction on recognition memory. *Front. Hum. Neurosci.*, *13*(291).

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language learning and development*, *1*(2), 151–195.

Huebner, P., & Willits, J. (2019). A two-process model of semantic development. *osf.io Preprint*.

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). Babyberta: Learning more grammar with small-scale child-directed language. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 624–646.

Huebner, P. A., & Willits, J. A. (2018). Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*, *9*.

Huebner, P. A., & Willits, J. A. (2021a). Scaffolded input promotes atomic organization in the recurrent neural network language model. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 408–422.

Huebner, P. A., & Willits, J. A. (2021b). Using lexical context to discover the noun category: Younger children have it easier. *Psychology of learning and motivation*. Elsevier.

Huettig, F., Audring, J., & Jackendoff, R. (2022). A parallel architecture perspective on pre-activation and prediction in language processing. *Cognition*.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological review*, *110*(2), 220–264.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Dev. Psychol.*, *27*(2), 236–248.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive psychology*, *45*(3), 337–374.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive psychology*, *61*(4), 343–365.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures*. Psychology Press.

Izhikevich, E. M. (2007). *Dynamical systems in neuroscience*. MIT press.

Jackendoff, R. (2002). What's in the lexicon? *Storage and computation in the language faculty* (pp. 23–58). Springer.

Jackendoff, R., & Jackendoff, R. S. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press, USA.

Jacobs, C. L., Dell, G. S., Benjamin, A. S., & Bannard, C. (2016). Part and whole linguistic experience affect recognition memory for multiword sequences. *Journal of Memory and Language*, *87*, 38–58.

Jacobs, P. S., & Zernik, U. (1988). Acquiring lexical knowledge from text: A case study. *Aaai*, *88*, 739–744.

Jakulin, A., & Bratko, I. (2003). Quantifying and visualizing attribute interactions. *arXiv Preprint cs/0308002*.

Jiang, H., Frank, M. C., Kulkarni, V., & Fourtassi, A. (2020a). Exploring patterns of stability and change in caregivers' word usage across early childhood. *psyArxiv Preprint*.

Jiang, H., Frank, M. C., Kulkarni, V., & Fourtassi, A. (2020b). Exploring patterns of stability and change in caregivers' word usage across early childhood. *PsyArxiv Preprint*.

John, M. F. S., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial intelligence*, *46*(1-2), 217–257.

Johns, B., & Jones, M. (2011). Construction in semantic memory: Generating perceptual representations with global lexical similarity. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*.

Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychon. Bull. Rev.*, *23*(4), 1214–1220.

Jones, M., & Recchia, G. (2010). You can't wear a coat rack: A binding framework to avoid illusory feature migrations in perceptually grounded semantic models. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Can. J. Exp. Psychol.*, *66*(2), 115–124.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological review*, *114*(1), 1–37.

Jones, M. N., Willits, J., Dennis, S., & Jones, M. (2015). Models of semantic memory. *Oxford handbook of mathematical and computational psychology*, 232–254.

Jones, S. S., Smith, L. B., & Landau, B. (1991a). Object properties and knowledge in early lexical learning. *Child Dev.*, *62*(3), 499–516.

Jones, S. S., Smith, L. B., & Landau, B. (1991b). Object properties and knowledge in early lexical learning. *Child development*, *62*(3), 499–516.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. *arXiv Preprint arXiv:1602.02410*.

Kabbach, A., & Herbelot, A. (2020). Avoiding conflict: When speaker coordination does not require conceptual agreement. *Front Artif Intell*, *3*.

Kádár, Á., Chrupała, G., & Alishahi, A. (2017). Representation of linguistic form and function in recurrent neural networks. *Computation Linguist.*, *43*(4), 761–780.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.

Kaiser, D., Jacobs, A. M., & Cichy, R. M. (2022). Modelling brain representations of abstract concepts. *PLoS Computation Biol.*, *18*(2).

Kamin, L. J. (1967). Predictability, surprise, attention, and conditioning. *Symposium on Punishment*.

Katz, N., Baker, E., & Macnamara, J. (1974). What's in a name? a study of how children learn common and proper names. *Child Dev.*, *45*(2), 469–473.

Ke, N. R., Alias Parth Goyal, A. G., Bilaniuk, O., Binas, J., Mozer, M. C., Pal, C., & Bengio, Y. (2018). Sparse attentive backtracking: Temporal credit assignment through reminding. *Adv. Neural Inf. Processes Syst.*, *31*.

Keibel, J.-H. (2005). *Distributional patterns in german child-directed speech and their usefulness for acquiring lexical categories: A case study* (Doctoral dissertation). Universität Freiburg.

Keil, F. C., Smith, W. C., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. *Cognition*, *65*(2-3), 103–135.

Keil, F., & Sessar, K. (1979). *Semantic and conceptual development: An ontological perspective*. Harvard University Press.

Keil, F. C. (1981). Constraints on knowledge and cognitive development. *Psychological review*, *88*(3), 197–227.

Kersten, A. W., & Earles, J. L. (2001). Less really is more for adults learning a miniature artificial language. *Journal of Memory and Language*, *44*(2), 250–273.

Kidd, E., Lieven, E. V., & Tomasello, M. (2010). Lexical frequency and exemplar-based learning effects in language acquisition: Evidence from sentential complements. *Language Sciences*, *32*(1), 132–142.

Kim, N., & Linzen, T. (2020). COGS: A compositional generalization challenge based on semantic interpretation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9087–9105.

Kirchhoff, K., & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *The Journal of the Acoustical Society of America*, *117*(4), 2238–2246.

Kirov, C., & Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, *6*, 651–665.

Klibanoff, R. S., & Waxman, S. R. (2000). Basic level object categories support the acquisition of novel adjectives: Evidence from preschool-aged children. *Child development*, *71*(3), 649–659.

Kóbor, A., Horváth, K., Kardos, Z., Nemeth, D., & Janacsek, K. (2020). Perceiving structure in unstructured stimuli: Implicitly acquired prior knowledge impacts the processing of unpredictable transitional probabilities. *Cognition*, *205*.

Koenig, J. P., Mauner, G., & Bienvenue, B. (2003). Arguments for adjuncts. *Cognition*, *89*(2), 67–103.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Dev.*, *67*(6), 2797–2822.

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, *8*(2), 225–248.

Kukona, A., Fang, S.-Y., Aicher, K. A., Chen, H., & Magnuson, J. S. (2011). The time course of anticipatory constraint integration. *Cognition*, *119*(1), 23–42.

Kuncoro, A., Ballesteros, M., Kong, L., Dyer, C., Neubig, G., & Smith, N. A. (2017). What do recurrent neural network grammars learn about syntax? *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1249–1258.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, *31*(1), 32–59.

Kuperberg, G. R., Kreher, D. A., Sitnikova, T., Caplan, D. N., & Holcomb, P. J. (2007). The role of animacy and thematic relationships in processing active english sentences: Evidence from event-related potentials. *Brain Language*, *100*(3), 223–237.

Landau, B., Gleitman, L. R., & Landau, B. (2009). *Language and experience: Evidence from the blind child* (Vol. 8). Harvard University Press.

Landau, B., & Jackendoff, R. (1993). Whence and whither in spatial language and spatial cognition? *Behavioral and brain sciences*, *16*(2), 255–265.

Landau, B., Smith, L. B., & Jones, S. (1992). Syntactic context and the shape bias in children's and adults' lexical learning. *Journal of Memory and Language*, *31*(6), 807–825.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299–321.

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), 211–240.

Lany, J., & Saffran, J. R. (2013). Statistical learning mechanisms in infancy. *Comprehensive developmental neuroscience: Neural circuit development and function in the brain*, *3*, 231–248.

Lany, J., & Gómez, R. L. (2008). Twelve-month-old infants benefit from prior experience in statistical learning. *Psychological Science*, *19*(12), 1247–1252.

Lany, J., & Saffran, J. R. (2010). From statistics to meaning: Infants' acquisition of lexical categories. *Psychological Science*, *21*(2), 284–291.

Lany, J., & Saffran, J. R. (2011). Interactions between statistical and semantic information in infant language development. *Developmental science*, *14*(5), 1207–1219.

Lee, L. S.-Y. (2015). On the linear algebraic structure of distributed word representations. *arXiv Preprint*.

Lenci, A. (2011). Composing and updating verb argument expectations: A distributional semantic model. *Proceedings of the 2nd workshop on cognitive modeling and computational linguistics*, 58–66.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*(1), 151–171.

Lester, N. A., Moran, S., Küntay, A. C., Allen, S. E. M., Pfeiler, B., & Stoll, S. (2021). Detecting structured repetition in child-surrounding speech: Evidence from maximally diverse languages. *Cognition*, *221*.

Levelt, W. J. M. (1975). What became of LAD. *Ut Vidaeum: Contributions to an Understanding of Linguistics*, 171–190.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Lewandowsky, S., Kalish, M., & Griffiths, T. L. (2000). Competing strategies in categorization: Expediency and resistance to knowledge restructuring. *J. Exp. Psychol. Learn. Mem. Cogn.*, *26*(6), 1666–1684.

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Dev. Sci.*, *14*(6), 1323–1329.

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2016). Visualizing and understanding neural models in NLP. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 681–691.

Liberman, A., Fischer, J., & Whitney, D. (2014). Serial dependence in the perception of faces. *Current biology*, *24*(21), 2569–2574.

Lieven, E. V. (1994). Crosslinguistic and crosscultural aspects of language addressed to children.

Lillicrap, T. P., & Santoro, A. (2019). Backpropagation through time and the brain. *Curr. Opin. Neurobiol.*, *55*, 82–89.

Lin, C. C., & Ahrens, K. (2005). How many meanings does a word have? meaning estimation in chinese and english. *Language acquisition, change and emergence: Essays in evolutionary linguistics*, 437–464.

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*, 195–212.

Linzen, T., & Jaeger, T. F. (2016). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, *40*(6), 1382–1411.

Liu, D., Lamb, A. M., Kawaguchi, K., ALIAS PARTH GOYAL, A. G., Sun, C., Mozer, M. C., & Bengio, Y. (2021). Discrete-valued neural communication. *Advances in Neural Information Processing Systems*, *34*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv Preprint arXiv:1907.11692*.

Locke, J. (1847). *An essay concerning human understanding*. Kay & Troutman.

Lotto, A. J., Sato, M., & Diehl, R. L. (2004). Mapping the task for the second language learner: The case of japanese acquisition of/r/and/l. *From Sound to Sense: 50+ Years of Discoveries in Speech Communication*.

Lu, S., Zhu, Y., Zhang, W., Wang, J., & Yu, Y. (2018). Neural text generation: Past, present and beyond. *arXiv Preprint arXiv:1803.07133*.

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Methods Instrum. Computation*, *28*(2), 203–208.

Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *34*(10), 1319–1337.

MacDonald, M. C. (2013). Distributional information in language comprehension, production, and acquisition: Three puzzles and a moral. *The emergence of language* (pp. 195–214). Psychology Press.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on just and carpenter (1992) and waters and caplan (1996).

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological review*, *101*(4), 676.

Macnamara, J. (1982). *Names for things: A study of human learning*. MIT Press.

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk. transcription format and programs* (Vol. 1). Psychology Press.

MacWhinney, B., & Bates, E. (1989). *The crosslinguistic study of sentence processing*. Cambridge University Press.

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabi, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., et al. (2020). Earshot: A minimal neural network model of incremental human speech recognition. *Cognitive science*, *44*(4).

Mandler, J. M., & McDonough, L. (1993). Concept formation in infancy. *Cognitive Develoment*, *8*(3), 291–318.

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—but only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 843–847.

Mannering, W. M., & Jones, M. N. (2021). Catastrophic interference in predictive neural network models of distributional semantics. *Computational Brain & Behavior*, *4*(1), 18–33.

Mao, S., Huebner, P. A., & Willits, J. A. (2022). Compositional generalization in a graph-based distributional semantic model. *CogSci*.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, *37*(3), 243–282.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by Seven-Month-Old infants. *Science*, *283*(5398), 77–80.

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, *46*(1), 53–85.

Marcus, G. F., Fernandes, K. J., & Johnson, S. P. (2007). Infant rule learning facilitated by speech. *Psychological Science*, *18*(5), 387–391.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*(1), 57–77.

Marslen-Wilson, W., Brown, C. M., & Tyler, L. K. (1988). Lexical representations in spoken language comprehension. *Language and Cognitive Processes*, *3*(1), 1–16.

Martin, A. E. (2020). A compositional neural architecture for language. *Journal of Cognitive Neuroscience*, *32*(8), 1407–1427.

Martin, A. E., & Doumas, L. A. (2017). A mechanism for the cortical computation of hierarchical linguistic structure. *PLoS biology*, *15*(3).

Matthews, D., & Bannard, C. (2010). Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child-directed speech. *Cognitive science*, *34*(3), 465–488.

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (mtld)* (Doctoral dissertation). The University of Memphis.

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review*, *126*(1), 1–51.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends Cognitive Science*, *14*(8), 348–356.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature reviews neuroscience*, *4*(4), 310–322.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of learning and motivation* (pp. 109–165). Elsevier.

McDonald, S., & Lowe, W. (1998). Modelling functional priming and the associative boost. *Proceedings of the 20th annual conference of the Cognitive Science Society*, 667–680.

McDonald, S., & Ramscar, M. (2001). Testing the distributioanl hypothesis: The influence of context on judgements of semantic similarity. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *23*(23).

McGill, W. (1954). Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, *4*(4), 93–111.

McRae, K., Hare, M., Elman, J. L., & Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Mem. Cognit.*, *33*(7), 1174–1184.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014–9019.

Merity, S., Keskar, N. S., & Socher, R. (2017). Regularizing and optimizing lstm language models. *arXiv Preprint arXiv:1708.02182*.

Meyer-Lindenberg, A., Ziemann, U., Hajak, G., Cohen, L., & Berman, K. F. (2002). Transitions between dynamical states of differing stability in the human brain. *Proceedings of the National Academy of Sciences*, *99*(17), 10948–10953.

Mikolov, T., Kombrink, S., Burget, L., Černocký, J., & Khudanpur, S. (2011). Extensions of recurrent neural network language model. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5528–5531.

Mikolov, T. (2012). *Statistical language models based on neural networks* (Ph.D. thesis). Brno University of Technology, Faculty of Information Technology.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv Preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119.

Mintz, T., & Gleitman, L. (1998). Incremental language learning: Two and three year olds' acquisition of adjectives. *Proceedings of the 20th Annual Conference of the Cognitive Science Society*.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117.

Mintz, T. H., & Gleitman, L. R. (2002). Adjectives really do modify nouns: The incremental and restricted nature of early adjective acquisition. *Cognition*, *84*(3), 267–293.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive science*, *26*(4), 393–424.

Mirman, D., Graf Estes, K., & Magnuson, J. S. (2010). Computational modeling of statistical learning: Effects of transitional probability versus frequency and links to word learning. *Infancy*, *15*(5), 471–486.

Misyak, J. B., Christiansen, M. H., & Bruce Tomblin, J. (2010). Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science*, *2*(1), 138–153.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, *34*(8), 1388–1429.

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, *55*(4), 259–305.

Monaghan, P., & Mattock, K. (2012). Integrating constraints for learning word-referent mappings. *Cognition*, *123*(1), 133–143.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*(2), 165–178.

Mu, J., & Viswanath, P. (2018). All-but-the-Top: Simple and effective postprocessing for word representations. *International Conference on Learning Representations*.

Murphy, G. (2004). *The big book of concepts*. MIT press.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of child language*, *17*(2), 357–374.

Naigles, L. R., Hoff, E., Vear, D., Tomasello, M., Brandt, S., Waxman, S. R., Childers, J. B., & Collins, W. A. (2009). Flexibility in early verb use: Evidence from a multiple-n diary study. *Monographs of the Society for Research in Child Development*, i–144.

Needham, A., & Baillargeon, R. (1997). Object segregation in 8-month-old infants. *Cognition*, *62*(2), 121–149.

Nelson, D. G. K., Hirsh-Pasek, K., Jusczyk, P. W., & Cassidy, K. W. (1989). How the prosodic cues in motherese might assist language learning. *Journal of child Language*, *16*(1), 55–68.

Neuman, S. B., Newman, E. H., & Dwyer, J. (2011). Educational effects of a vocabulary intervention on preschoolers' word knowledge and conceptual development: A cluster-randomized trial. *Reading Research Quarterly*, *46*(3), 249–272.

Newport, E. L., Gleitman, H., & Gleitman, L. R. (1977). Mother i'd rather do it myself: The contribution of selected child listener variables'. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to children*. Cambridge University Press.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive science*, *14*(1), 11–28.

Ninio, A. (2004). Young children's difficulty with adjectives modifying nouns. *Journal of Child Language*, *31*(2), 255–285.

Noelle, D. C., & Zimdars, A. L. (1999). Methods for learning articulated attractors over internal representations. *Proceedings of the 21st Annual Conference of the Cognitive Science Society*, 480–485.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *J. Exp. Psychol. Gen.*, *115*(1), 39–57.

Olney, A. M., Dale, R., & D'Mello, S. K. (2012). The world within wikipedia: An ecology of mind. *Information*, *3*(2), 229–255.

Onnis, L., Christiansen, M. H., Chater, N., & Gómez, R. (2003). Reduction of uncertainty in human sequential learning: Evidence from artificial grammar learning. *Proceedings of the Annual meeting of the Cognitive Science Society*, *25*(25).

O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and hebbian learning. *Neural computation*, *13*(6), 1199–1241.

Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological review*, *56*(3), 132–143.

Pannitto, L., & Herbelot, A. (2022). Can recurrent neural networks validate usage-based theories of grammar acquisition? *Frontiers in Psychology*.

Patterson, K., Plaut, D. C., Mcclelland, J. L., Seidenberg, M. S., Behrmann, M., & Hodges, J. R. (1996). Connections and disconnections: A connectionist account of surface dyslexia. *Neural modeling of brain and cognitive disorders* (pp. 177–199). World Scientific.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child development*, *80*(3), 674–685.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cogn. Neuropsychol.*, *33*(3-4), 175–190.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*(1).

Perfetti, C. A. (1998). The limits of co-occurrence: Tools and theories in language research. *Discourse Processe*, *25*, 363–377.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends Cognitive Science*, *11*(3), 105–110.

Pine, J. M. (1994). The language of primary caregivers. (C. Gallaway & B. Richards, Eds.).

Pinker, S. (1984). *Language learnability and language development*. Harvard University Press.

Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 399–441). Lawrence Erlbaum Associates.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73–193.

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), 786–823.

Poletiek, F. H., Conway, C. M., Ellefson, M. R., Lai, J., Bocanegra, B. R., & Christiansen, M. H. (2018). Under what conditions can recursion be learned? effects of starting small in artificial grammar learning of center-embedded structure. *Cognitive Science*, *42*(8), 2855–2889.

Pustejovsky, J. (1998). *The generative lexicon*. MIT Press.

Pustejovsky, J., & Boguraev, B. (1993). Lexical knowledge representation and natural language processing. *Artif. Intell.*, *63*(1), 193–223.

Pylyshyn, Z. W. (2000). Situating vision in the world. *Trends in Cognitive Sciences*, *4*(5), 197–207.

Quine, W. V. O. (2013). *Word and object*. MIT press.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the n400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693–705.

Rabovsky, M., & McClelland, J. L. (2020). Quasi-compositional mapping from form to meaning: A neural network-based approach to capturing neural responses during human language comprehension. *Philosophical Transactions of the Royal Society B*, *375*(1791).

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

Rafferty, A., & Griffiths, T. (2010). Optimal language learning: The importance of starting representative. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*.

Ramirez-Esparza, N., Garcia-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental science*, *17*(6), 880–891.

Ramscar, M. (2001). The influence of semantics on past-tense inflection. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *23*(23).

Ramscar, M. et al. (2013a). Suffixing, prefixing, and the functional order of regularities in meaningful strings. *Psihologija*, *46*(4), 377–396.

Ramscar, M., Dye, M., & Klein, J. (2013b). Children value informativity over logic in word learning. *Psychological Science*, *24*(6), 1017–1023.

Ramscar, M., & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends Cognitive Science*, *11*(7), 274–279.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79–87.

Rapp, R. (2002). The computation of word associations: Comparing syntagmatic and paradigmatic approaches. *COLING 2002: The 19th International Conference on Computational Linguistics*.

Räsänen, O., & Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological Review*, *122*(4), 792–829.

Ravfogel, S., Goldberg, Y., & Linzen, T. (2019). Studying the inductive biases of RNNs with synthetic variations of natural languages. *Proceedings of the 2019 Conference of the North American Chapter* *of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3532–3542.

Ravfogel, S., Prasad, G., Linzen, T., & Goldberg, Y. (2021). Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 194–209.

Recchia, G., Jones, M., Sahlgren, M., & Kanerva, P. (2010). Encoding sequential information in vector space models of semantics: Comparing holographic reduced representation and random permutation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*(32).

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive science*, *22*(4), 425–469.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive psychology*, *66*(1), 30–54.

Richards, B. J. (1994). Child-directed speech and influences on language acquisition: Methodology and interpretation. *Input and interaction in language acquisition*. Cambridge University Press.

Riordan, B. (2007). *Comparing semantic space models using child -directed speech* (M. Gasser & M. N. Jones, Eds.; Doctoral dissertation). Indiana University.

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303–345.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of verbal learning and verbal behavior*, *14*(6), 665–681.

Roark, C. L., Plaut, D. C., & Holt, L. L. (2022). A neural network model of the effect of prior experience with regularities on subsequent category learning. *Cognition*, *222*.

Rodriguez, P., Wiles, J., & Elman, J. L. (1999). A recurrent neural network that learns to count. *Connection Science*, *11*(1), 5–40.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Rohde, D. L., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*(1), 67–109.

Rohde, D. L. T. (2002). *A connectionist model of sentence comprehension and production* (D. C. Plaut, Ed.; Doctoral dissertation). Carnegie Mellon University.

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.

Rommers, J., Meyer, A. S., Praamstra, P., & Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. *Neuropsychologia*, *51*(3), 437–447.

Rosenfeld, R., Touretzky, D. S., & Group, B. (1987). Connectionist models as neural abstractions. *Behavioral and Brain Sciences*, *10*(2), 181–182.

Rotaru, A. S., Vigliocco, G., & Frank, S. L. (2018). Modeling the structure and dynamics of semantic processing. *Cognitive science*, *42*(8), 2890–2917.

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Dev.*, *83*(5), 1762–1774.

Rubino, R. B., & Pine, J. M. (1998). Subject–verb agreement in brazilian portuguese: What low error rates hide. *Journal of child language*, *25*(1), 35–59.

Ruge, G. (1992). Experiments on linguistically-based term associations. *Information Processing & Management*, *28*(3), 317–332.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Russin, J. L., Jo, J., O'Reilly, R. C., & Bengio, Y. (2020). Systematicity in a recurrent neural network by factorizing syntax and semantics. *CogSci*.

Sadeghi, Z., McClelland, J. L., & Hoffman, P. (2015a). You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia*, *76*, 52–61.

Sadeghi, Z., McClelland, J. L., & Hoffman, P. (2015b). You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia*, *76*, 52–61.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621.

Sahlgren, M. (2006). *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces* (Doctoral dissertation). Institutionen för lingvistik.

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). Childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior research methods*, *51*(4), 1928–1941.

Sandhofer, C. M., Smith, L. B., & Luo, J. (2000). Counting nouns and verbs in the input: Differential frequencies, different kinds of learning? *Journal of Child Language*, *27*(3), 561–585.

Saphra, N., & Lopez, A. (2019). Understanding learning dynamics of language models with SVCCA. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3257–3267.

Saphra, N., & Lopez, A. (2020). LSTMs compose—and Learn—Bottom-up. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2797–2809.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human perception and performance*, *3*(1).

Scarselli, F., & Tsoi, A. C. (1998). Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results. *Neural networks*, *11*(1), 15–37.

Schäfer, A. M., & Zimmermann, H. G. (2006). Recurrent neural networks are universal approximators. *International Conference on Artificial Neural Networks*, 632–640.

Schieffelin, B. B., & Ochs, E. (1986). Language socialization. *Annual review of anthropology*, *15*(1), 163–191.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings Natl. Acad. Sci. U. S. A.*, *118*(45).

Schwenk, H., & Gauvain, J.-L. (2005). Training neural network language models on very large corpora. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 201–208.

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *J. Exp. Psychol. Learn. Mem. Cogn.*, *23*(3), 681–696.

Seidenberg, M. S., & MacDonald, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive science*, *23*(4), 569–588.

Seidl, A., & Johnson, E. K. (2006). Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental science*, *9*(6), 565–573.

Sekerina, I. A., & Trueswell, J. C. (2012). Interactive processing of contrastive expressions by russian children. *First Language*, *32*(1-2), 63–87.

Sennrich, R., Haddow, B., & Birch, A. (2016). Edinburgh neural machine translation systems for wmt 16. *arXiv Preprint arXiv:1606.02891*.

Servan-Schreiber, D., Cleeremans, A., & McClelland, J. L. (1991). Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning*, *7*(2-3), 161–193.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., & Netrapalli, P. (2020). The pitfalls of simplicity bias in neural networks. *Adv. Neural Inf. Processes Syst.*, *33*, 9573–9585.

Shanks, D. R. (1991). Categorization by a connectionist network. *J. Exp. Psychol. Learn. Mem. Cogn.*, *17*(3), 433–443.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Shen, Y., Lin, Z., Huang, C.-w., & Courville, A. (2018). Neural language modeling by jointly learning syntax and lexicon. *International Conference on Learning Representations*.

Shen, Y., Tan, S., Sordoni, A., & Courville, A. (2018). Ordered neurons: Integrating tree structures into recurrent neural networks. *arXiv Preprint*.

Shipley, E. F., Smith, C. S., & Gleitman, L. R. (1969). A study in the acquisition of language: Free responses to commands. *Language*, 322–342.

Siegelmann, H. T., & Sontag, E. D. (1992). On the computational power of neural nets. *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*.

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy*, *14*(6), 654–666.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1-2), 39–91.

Slobin, D. I. (1973). Cognitive prerequisites for the development of grammar. *Studies of child language development* (pp. 175–208). Holt, Rinehart, & Winston.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*(3), 214–241.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Smith, L. B., & Thelen, E. E. (1993). *A dynamic systems approach to development: Applications*. The MIT Press.

Snedeker, J., Brent, M., & Gleitman, L. (2001). The changing character of the mental lexicon: An information-based account of early word learning. *Unpublished manuscript*.

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Dev.*, *43*(2), 549–565.

Snow, C. E., & Ferguson, C. A. (1977). *Talking to children*. Cambridge University Press.

Spencer, J. P., & Schöner, G. (2003). Bridging the representational gap in the dynamic systems approach to development. *Developmental Science*, *6*(4), 392–412.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Wiley-Blackwell.

Spivey, M. (2008). *The continuity of mind*. Oxford University Press.

St Clair, M. C., Monaghan, P., & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*, *116*(3), 341–360.

St. Clair, M. C., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science*, *33*(7), 1317–1329.

Staub, A., White, S. J., Drieghe, D., Hollway, E. C., & Rayner, K. (2010). Distributional effects of word frequency on eye fixation durations. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1280–1293.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis* (pp. 439–460). Psychology Press.

Stowe, L. A. (1989). Thematic structures and sentence comprehension. *Linguistic structure in language processing* (pp. 319–357). Springer.

Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and brain sciences*, *30*(3), 299–313.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc.

Syrett, K., & Lidz, J. (2010). 30-month-olds use the distribution and meaning of adverbs to interpret novel adjectives. *Language Learning and Development*, *6*(4), 258–282.

Szubert, I., Abend, O., Schneider, N., Gibbon, S., Goldwater, S., & Steedman, M. (2021). Cross-linguistically consistent semantic and syntactic annotation of child-directed speech. *arXiv Preprint arXiv:2109.10952*.

Tabor, W., Juliano, C., & Tanenhaus, M. K. (1997a). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language & Cognitive Processes*.

Tabor, W., Juliano, C., & Tanenhaus, M. K. (1997b). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language & Cognitive Processes*.

Tabor, W. (2002). The value of symbolic computation. *Ecological Psychology*, *14*(1-2), 21–51.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*(4), 355–370.

Tanaka-Ishii, K., & Kobayashi, T. (2018). Taylor's law for linguistic sequences and random walk models. *Journal of Physics Communications*, *2*(11).

Tanenhaus, M. K., Carlson, G., & Trueswell, J. C. (1989). The role of thematic structures in interpretation and parsing. *Language and Cognitive Processes*, *4*(3-4), Si211–si234.

Tardif, T., Shatz, M., & Naigles, L. (1997). Caregiver speech and children's use of nouns versus verbs: A comparison of english, italian, and mandarin. *Journal of Child Language*, *24*(3), 535–565.

Taylor, M., & Gelman, S. A. (1989). Incorporating new words into the lexicon: Preliminary evidence for language hierarchies in two-year-old children. *Child Dev.*, *60*(3), 625–636.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*(1), 53–71.

Thorpe, K., Baumgartner, H., & Fernald, A. (2006). Children's developing ability to interpret adjective-noun combinations. *Proceedings of the 30th Annual Boston University Conference on Language Development*, 631–642.

Thothathiri, M., & Braiuca, M. C. (2021). Distributional learning in english: The effect of verb-specific biases and verb-general semantic mappings on sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(1), 113–128.

Tomasello, M. (1999). Having intentions, understanding intentions, and understanding communicative intentions. Lawrence Erlbaum Associates.

Tomasello, M. (2005). *Constructing a language: A Usage-Based theory of language acquisition*. Harvard University Press.

Trainor, L. J., & Desjardins, R. N. (2002). Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. *Psychonomic Bulletin & Review*, *9*(2), 335–340.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, *33*(3), 285–318.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res.*, *37*, 141–188.

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*(1-2), 231–270.

Unger, L., & Fisher, A. V. (2021). The emergence of richly organized semantic knowledge from simple statistics: A synthetic review. *Developmental Review*, *60*.

Unger, L., Vales, C., & Fisher, A. V. (2020). The role of co-occurrence statistics in developing semantic knowledge. *Cognitive Science*, *44*(9).

Valian, V. (2014). Arguing about innateness. *Journal of child language*, *41*(S1), 78–92.

Valian, V., Prasada, S., & Scarpa, J. (2006). Direct object predictability: Effects on young children's imitation of sentences. *Journal of Child Language*, *33*(2), 247–269.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(11).

Van Rooij, I. (2008). The tractable cognition thesis. *Cognitive science*, *32*(6), 939–984.

Vankov, I. I., & Bowers, J. S. (2020). Training neural networks to encode symbols enables combinatorial generalization. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *375*(1791).

van Schijndel, M., Mueller, A., & Linzen, T. (2019). Quantity doesn't buy quality syntax with neural language models. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5831–5837.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. U., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus,

S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems.* Curran Associates, Inc.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *J. Exp. Psychol. Gen.*, *121*(2), 222–236.

Wang, L., Wlotko, E., Alexander, E., Schoot, L., Kim, M., Warnke, L., & Kuperberg, G. R. (2020). Neural evidence for the prediction of animacy features during language comprehension: Evidence from meg and eeg representational similarity analysis. *Journal of Neuroscience*, *40*(16), 3278–3291.

Warrington, E. K. (1975). The selective impairment of semantic memory. *The Quarterly journal of experimental psychology*, *27*(4), 635–657.

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. *Trans. Assoc. Computation Linguist.*, *7*, 625–641.

Wasow, T. (1973). The innateness hypothesis and grammatical relations. *Synthese*, *26*(1), 38–56.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302.

Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, *13*(6), 258–263.

Waxman, S. R., & Klibanoff, R. S. (2000). The role of comparison in the extension of novel adjectives. *Dev. Psychol.*, *36*(5), 571–581.

Webb, T. W., Sinha, I., & Cohen, J. D. (2020). Emergent symbols through binding in external memory. *arXiv Preprint.*

Werbos, P. J. (1990). Backpropagation through time: What it does and how to do it. *Proceedings IEEE Inst. Electr. Electron. Eng.*, *78*(10), 1550–1560.

Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits* (tech. rep.).

Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 694–709.

Willits, J. A., D'Mello, S. K., Duran, N. D., & Olney, A. (2007). Distributional statistics and thematic role relationships. *Proceedings of the annual meeting of the cognitive science society*, *29*.

Willits, J. A., Seidenberg, M. S., & Saffran, J. R. (2014). Distributional structure in language: Contributions to noun–verb difficulty differences in infant word recognition. *Cognition*, *132*(3), 429–436.

Wilson, R. C., Shenhav, A., Straccia, M., & Cohen, J. D. (2019). The eighty five percent rule for optimal learning. *Nature Communications*, *10*(1).

Wojcik, E. H., & Saffran, J. R. (2015). Toddlers encode similarities among novel words from meaningful sentences. *Cognition*, *138*, 10–20.

Xie, X., & Seung, H. S. (2003). Equivalence of backpropagation and contrastive hebbian learning in a layered network. *Neural Computation*, *15*(2), 441–454.

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neurosci.*, *22*(2), 297–306.

Yavaş, M., & Someillan, M. (2005). Patterns of acquisition of/s/-clusters in spanish-english bilinguals. *Journal of Multilingual Communication Disorders*, *3*(1), 50–55.

Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychon. Bull. Rev.*, *23*(4), 1015–1027.

Younger, B. (1990). Infant categorization: Memory for category-level and specific item information. *J. Exp. Child Psychol.*, *50*(1), 131–155.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–262.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, *47*(1), 1–29.

Zhang, X., Kumar, G., Khayrallah, H., Murray, K., Gwinnup, J., Martindale, M. J., McNamee, P., Duh, K., & Carpuat, M. (2018). An empirical exploration of curriculum learning for neural machine translation. *arXiv Preprint arXiv:1811.00739*.

Zheng, H., & Lapata, M. (2021). Disentangled sequence to sequence learning for compositional generalization. *arXiv Preprint*.

Zhu, H., & Clark, A. (2022). Distributional lattices as a model for discovering syntactic categories in child-directed speech. *Journal of Psycholinguistic Research*, 1–15.