Using lexical context to discover the noun category: Younger children have it easier

Philip A. Huebner & Jon A. Willits

Department of Psychology, University of Illinois at Urbana-Champaign

September, 2021

Abstract

Prior work has demonstrated that distributional dependencies between word or morpheme-like entities in artificial and naturalistic language can detect clusters of words which broadly conform to the categories of the adult language (Brent & Siskind, 2001; Mintz, 2002; Redington & Chater, 1998). In this work, we examine the hypothesis that the distributional statistics useful for the discovery of the noun category are more useful in speech to younger children compared to older children (approximately 1-3 vs. 3-6 years of age). First, using a novel method for quantifying the extent that nouns occur in mutually shared contexts, we demonstrate an advantage for speech to younger compared to older children. Second, we develop a theoretical framework for understanding why caregiver speech might be scaffolded in this way, and test its predictions against an array of information theoretic patterns computed on child-directed speech. Our account, based on entropy-maximization, and anchoring originally proposed by (Cameron-Faulkner et al., 2003), clarifies issues in incremental learning from non-stationary input - the problem faced by language learners - and paves the way towards integrating the scaffolded organisation of children's early language environment into computational models of acquisition.

I. Introduction

Language acquisition is often studied as if it were a static process: There is some built in knowledge or representational structure (whether that be a universal grammar or a neural network architecture with defined inputs and outputs), and a learning mechanism that operates over a relatively consistent input. With the right kind, or with the right amount of input, the system will acquire a set of representations that allow for linguistic behavior. Of course, no one actually argues that the learning process or the input is in fact static. This is usually adopted as a simplifying assumption while considering and modeling different learning and representational mechanisms. But adopting this assumption of a static process and static input (e.g. sampling language data from a stationary distribution) can lead to a serious mischaracterization of the learning problem that children face.

Language development unfolds over the course of many years and involves a multitude of factors that are changing over the course of development. During that time, many neurological changes take place, that shape children's computational capacities (Casey et al., 2000). For example, the brain of a newborn is vastly different from that of a six-year-old. Perceptual systems are changing, as visual acuity improves over the first two years of life (Graven & Browne, 2008). Auditory perception begins prenatally while the auditory system is still developing (Seebach et al., 1995), and the bodies of children are themselves developing and changing, in ways that affect learning and

representational capacities (Byrge et al., 2014).

In addition to these physical changes, as children's knowledge increases, the knowledge they acquire reshapes their perceptual and processing capacities. There is at this point a tremendous amount of evidence that children use prior knowledge to bootstrap further learning, and that each learning episode does not exist in isolation from all others (Estes et al., 2007; Fisher et al., 2010; Lany & Saffran, 2010). For example, Fernald and Marchman (2012) have argued that children with higher vocabularies process language more quickly at an early age. Closely associated is the "less-is-more" hypothesis proposed by E. L. Newport (1990) which claims that children's' developing cognitive abilities provides a crucial processing advantage for early language learning. For instance, young children's limited memory capacity is thought to break apart the speech stream into more useful smaller chunks. It has also been proposed that children's developing cognitive system predisposes them to regularize language to a greater extent than adult learners (Cochran et al., 1999; Hudson Kam & Newport, 2005; K. Smith et al., 2017).

One underappreciated way in which the language acquisition process is changing over the course of development is the way in which the input *itself* is changing. The fact that the visual and auditory systems are changing means that the nature of the data from which children are learning is not static or uniformly distributed across time, but instead is changing in structured ways. As infants bodies' develop (as they go from laying to sitting to crawling to walking, or gain the ability to hold objects), their perceptual input undergoes dramatic changes as well (Bertenthal et al., 1997; Jayaraman et al., 2017; Kretch et al., 2014). Further, qualitative aspects of language input, such as type token ratio and other measures of lexical diversity, are known to change as children encounter more data, and especially as they spend more time reading (Cunningham & Stanovich, 1998; Montag et al., 2015; Montag & MacDonald, 2015; Schwering et al., 2021).

In this paper, we focus on another important way in which input is changing during language acquisition: the lexical distributional signals that predict noun-membership in child-directed speech. We discuss in detail the unique learning problem that arises when conceptualizing the discovery of word classes as incremental learning from non-stationary input. Previous work has described several learning mechanisms by which children might use a distributional information to discover lexical classes (J. L. Elman, 1991; Freudenthal et al., 2013). We will show that any method that does not take into consideration the non-stationary aspect of children's language environment is potentially subject to a particular kind of learning trap, called fragmentation (Jakulin & Bratko, 2003). Fragmentation can occur when the distributional signals associated with a single category (like nouns) conflicts with the distributional signals associated with the presence of potentially multiple subcategories (nouns that refer to animals, foods, people, tools, etc.). The tug-of-war between distributional patterns generated by subordinate categories and other lexically specific idiosyncrasies pull apart ("fragment") the lexical representations of the superordinate category. In this paper, we provide an intuitive visual explanation of fragmentation, and develop quantitative tools for assessing fragmentation in speech to children, and how it varies over developmental time. Notably, our results demonstrate that the distributional signals that are diagnostic of noun-membership are less fragmented in speech to younger compared to older children. Our follow-up analysis indicate that reduced fragmentation in speech to younger children is made possible by anchor points, lexical contexts that are particularly suitable for abstracting over sub-category variation within the noun category.

I. Implications for Learning

The corpus analyses we report in this chapter were specifically developed to better our understanding of how incremental changes in qualitative aspects of language might influence the

hypotheses acquired by distributional learning models. Because we examine lexical distributions, and study how they change over time, our results have implications for models that learn 1) incrementally, and 2) by tracking distributional similarities between words. For example, many distributional models learn by predicting which words occur with which nearby words in large text corpora, and adjusting their parameters to better fit the target distribution. A particularly well-known model that learns in this way is the recurrent neural network of Elman (J. L. Elman, 1991) which outputs the probability that a word occurs after a sequence of words that has been provided to the model as input. In order to advance the field, we discuss ideas for developing distributional learning systems that can take advantage of the non-stationary nature of fragmentation in speech to children, and avoid its pitfalls. In particular, our work suggests that distributional learning systems can avoid fragmentation if they acquire representations via progressive differentiation (Keil, 1981; Rogers & McClelland, 2008) - first learning the larger, more superordinate categories, and then breaking those categories into progressively smaller subcategories. But the implications of our work also extend beyond neural networks and the computational learning literature; children are known to track distributional regularities in the speech they hear (Höhle et al., 2004; Shi & Melançon, 2010; Yuan et al., 2012) and use this information to predict upcoming words during language comprehension (Rabagliati et al., 2016) and when assigning novel labels to referents (Gelman & Taylor, 1984). To the degree that children base their linguistic hypotheses on distributional evidence - if only partially - our analyses can shed light on how such hypotheses are shaped and revised over time.

II. Background

I. Incremental Learning

In computational work, learning from one example at a time is often called "incremental" learning, and is distinguished from "batch" learning in which a large chunk of data is used to simultaneously update the parameters of a model. This distinction is especially important in language acquisition research because children do not wait patiently to build an inventory of distributional information before attempting to assign category labels; rather, they assign what they can when they can (Gleitman et al., 2005; Pinker, 2009; Seidenberg & MacDonald, 1999). Similarly, diary (Tomasello, 1992), priming (Kemp et al., 2005), and word learning studies (Mintz & Gleitman, 2002) indicate that young children form their knowledge of abstract categories, such as verbs, nouns, and adjectives, gradually and contingent on what information has been made available in the past (but see Valian, 1986). Apart from grammatical development, similar observations have been made in children's acquisition of concepts. For example, studies of children's early vocabulary have shown that children tend to learn about basic level words before superordinate or subordinate items (Mervis, 1983). Along the same lines, Clark (1973) found that children learn the meaning of words gradually by adding more features to their lexical entries. Both studies indicate that the linguistic representations learned by children are fashioned incrementally, and potentially via "progressive differentiation" - that is, integrating information about increasingly finer-grained semantic distinctions. Collectively, these findings suggest that children's early knowledge - based on very little language exposure - is incomplete and requires continuous revision as more information becomes available. While incremental learning is potentially useful for children to quickly acquire a foothold, this strategy risks susceptibility to falling into a learning trap, from which recovery can be costly. For instance, one could imagine that early exposure to input that is particularly skewed may make it much more difficult for a learner to overcome the initial bias of their early statistical analyses. The first few instances of language input could thus have a disproportionately large influence on the course of learning than subsequent in-

put. This is not a problem when the data is sampled from a stationary distribution (the parameters of the underlying process generating the data do not change over time), because initial samples will be equally representative of the data as all future samples. However, as our discussion above suggests, the language input to children (in all languages where it has been documented and analyzed) is non-stationary, and consequently, a learner's initial distributional analysis without inductive biases or strong priors may produce particularly skewed hypotheses. As we shall see, this can either help or hurt learning, depending on the hypotheses licensed by the initial exposure.

Provided we take seriously the non-stationary aspect of children's language environment, what might be the consequences for the learner? In other words, how might shifts in the complexity or co-occurrence relationships between words impact what representations are required? Would learning outcomes differ between systems trained on data that is ordered differently, for example? Only a handful of computational works have addressed this question (J. L. Elman, 1993; Freudenthal et al., 2016). For example, J. L. Elman (1993) showed that a recurrent network trained on syntactically simplified pseudo-English sentences first was better able to model syntactically more complex dependencies involving embedded relative clauses compared to a model trained on fully complex input right away (but see Rohde and Plaut, 1999). Freudenthal et al. (2016) exposed their model to input by gradually expanding access to contexts in a developmentally plausible way, and found that this results in the development of a noun category before a verb category in line with child data. While other work has explored the consequences of incremental as opposed to batch-based learning systems (Alishahi & Chrupala, 2012) - an important topic by itself - much of this work does not combine incremental learning with non-stationary input, a crucial next step for building more veridical models of acquisition. In order to support research on this front, we require tools and conceptual frameworks that facilitate working with and quantifying aspects of non-stationary language data.

I.1 Developmental Changes in Input Quantity and Quality

An obvious consequence of the incremental nature of language acquisition is that more language data becomes available with time. It is important to distinguish the quantity from the quality of the input, as these two factors may change independently from one another. While quantity must increase with time, the quality of the input, such as syntactic complexity, exposure to novel accents or pronunciations, and lexical diversity among many others (see next section), may increase, or decrease as a function of time, or not change at all. Because distributional analysis relies on limited samples of language data, large amounts of data are needed. Inferences about, say, word segmentation, or the assignment of words in a sentence to parts-of-speech categories, requires large quantities of information about transition probabilities between specific pairs of syllables or words. Given the large number of such pairs in natural languages, language experience is critical for distributional learners to succeed at such tasks. More data is not only helpful for sampling novel co-occurrences, but also for distinguishing which co-occurrences are co-incidental or linguistically relevant.

But children are not just accumulating larger quantities of input with time; the qualitative nature of the input also changes. Many infants begin the process of acquisition in a simplified language environment, and are not fully exposed to the statistics of adult language until many years later, during young adulthood. Evidence for such incremental changes in the language environment of children comes from studies that compare speech directed to children to speech between adults. For example, when talking to children, caregivers tend to employ larger pitch contours, lengthened vowels (Fernald & Kuhl, 1987) and less complex constructions (Broen, 1972; Furrow et al., 1979; E. Newport et al., 1977; Snow & Fergu-

son, 1977). Further, when speaking to children, caregivers are more likely to restrict the range of conversational topics, limit lexical diversity and grammatical constructions (Huttenlocher et al., 2010; Lieven, 1994; Snow & Ferguson, 1977), and make longer pauses at utterance boundaries (Gallaway & Richards, 1994). For reviews, see (Pine, 1994; Richards, 1994). Each are thought to bestow distinct learning advantages on the language learner, such as more accurate speech segmentation, and word recognition (Golinkoff & Alioto, 1995) [1]. Given the supporting role of child-directed speech in facilitating various aspects of language acquisition, we considered the possibility that distributional signals in the language environment of younger learners could, by extension, better support the discovery of lexical classes compared to speech to older children. If so, our findings would lend support to a growing list demonstrating the importance of early language input for supporting future learning outcomes.

Prior corpus studies have revealed that lexical diversity and mean utterance length are both strongly age-related (Broen, 1972; Foushee et al., 2016; Hayes & Ahrens, 1988; Kirchhoff & Schimmel, 2005; Phillips, 1973). But few researchers have examined how such changes in language complexity may affect the distributional cues for distinguishing between part-of-speech classes. Hills et al. (2010) have shown that child-directed language amplifies the associative structure and contextual consistency of the earliest learned words (Hills et al., 2010). While Hills and colleagues did not examine lexical categories directly, their findings provide the theoretical motivation for our work: If it is true that the association between words and their contexts is more consistent in the early language environment, then this should facilitate learning more general categories first. Preliminary computational support for this idea was provided by Borovsky and Elman (2006), who found that simpler grammatical constructions aid early category formation in the simple recurrent network. However, Borovsky and Elman (2006) employed artificial language which was under tight experimental control, and therefore left unanswered whether a similar benefit would apply to input that children actually receive.

At the earliest stages of language learning, when the quantity of input is limited, a learner's initial hypotheses and proto-categories may be particularly sensitive to qualitative aspects of the input. As a learner is exposed to more language data and can therefore increasingly rely on statistical power for making inferences, sensitivity to qualitative aspects of the data are likely to become less important. However, little is known about the persistence of early acquired linguistic abstractions (e.g. how much data is needed to refute inaccurate hypotheses?) and the role of such biases in the formation of novel abstractions (e.g. how might inaccurate hypotheses influence how distributional information is used to group words into word classes?). Further, it is not unreasonable to assume there exist qualitative aspects in language data that are particularly helpful for avoiding inaccurate hypotheses during early acquisition. But many of these questions remain unanswered and are in many cases never asked.

II. Distributional Learning of Word Classes

An important milestone in language development is the acquisition of word classes. Predicting that a novel word in a sentence is, for example, a noun, or refers to an animal, enables language learners to make many inferences about the kinds of events and relations that will also occur. While semantic categories are essential for organizing our understanding of and making inferences about the world, grammatical categories, such as part-of-speech (e.g nouns, verbs), enable the systematic re-use of

[1] Despite numerous benefits on language acquisition, cross-linguistic examination revealed that some children appear to learn language just as well when their primary caregivers do not employ child-directed adjustments (Lieven, 1994; Schieffelin & Ochs, 1986).

words to express and understand novel ideas. In this chapter, we are concerned with the distributional signals in speech to children that cue the presence of a noun category, and specifically, explore how the quality of these lexical cues changes across the developmental period during which children acquire language.

Many studies have established that infants and children exploit lexical co-occurrence distributions in their input to support the formation of part-of-speech representations. Because of the predictable structure of natural languages, words that belong to the same part-of-speech category tend to occur in similar lexical contexts, whereas members of different categories tend to occur in different contexts (Braine, 1963; Firth, 1961; Harris, 1954). These distributions over contexts can be exploited to infer the category membership of previously heard or novel words. Numerous corpus and computational studies have shown that there is sufficient distributional information in child-directed language for the induction of part-of-speech categories (Mintz, 2003; Redington & Chater, 1998), and that cognitively plausible algorithms can capitalize on this information to infer category membership (Alishahi & Chrupala, 2012; Freudenthal et al., 2013; Kodner, 2018). Further, distributional models previously trained on artificial input by J. L. Elman (1991) scale to corpora with millions of words of naturalistic and noisy speech to children (Huebner & Willits, 2018).

Importantly, behavioral evidence is mounting that children are in fact sensitive to and use this information in their acquisition of lexical category representations. First, studies of "syntactic bootstrapping" show that children make inferences about the transitivity of a novel verb based on the number of noun arguments with which it co-occurs (Yuan et al., 2012). Second, looking time studies show that 14 month-olds are sensitive to distributional violations in which nonsense words appear in linguistic contexts different from the context in which they appeared previously (Höhle et al., 2004; Shi & Melançon, 2010). Third, phonological competition effects on word learning can be eliminated by experimentally manipulating expectations about what class of words is likely to appear next in a sentence (Dautriche et al., 2015). Fourth, artificial-grammar learning studies, controlling for prior (possibly non-linguistic) experiences with words, show that distributional regularities imposed on nonsense words can be learned by both adults and children (Gomez, 2002; Lany & Saffran, 2010).

II.1 Anchoring

Corpus studies have shown that speech to English-speaking children consists of "anchor points", 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). Moreover, morphological markers such as *-ing* are also highly frequent units that are known to strongly differentiate between part-of-speech categories, such as between nouns and verbs (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. This view of acquisition opposes the central dogma of nativism which claims that the learner has available from birth knowledge about abstract syntactic categories, and that there is no need - in fact, that it is impossible - to construct such abstractions via experience alone (Chomsky, 2002). From the point of view of usage-based theories of language acquisition, anchor points represent a critical way-point for language learners between mastery of lexically-specific and fully abstract knowledge. Usage-based theorists hold that partially abstracted constructions, formed by exposure to individual language examples, guide learners towards abstract linguistic knowledge (Tomasello, 2001). While any variable slot construction, or any other type of distributional signal can be exploited for learning linguistic abstractions by the learner, anchor points are particularly useful due to their frequent occurrence and highly abstracted variable slot. Additionally, due to the small number of anchor points compared

to the number of possible contexts in which words can occur, they are particularly useful to a young learner with limited memory and experience. Because anchor points represent a very small fraction of possible grammatical contexts, a learner with limited experience is more likely to break into the regularities of their language by tracking anchor points than less frequent - possibly equally predictive - cues.

How might anchoring influence the course of learning word classes from distributional data? The most straightforward answer is that early discovery of highly predictive dependencies between lexical frames and variable slots could provide stronger demarcations between learned word clusters. Because anchor points are highly predictive relationships between lexical frames and groups of words that occur in those frames, the word clusters based on anchor points are more powerful and consistent cues to part-of-speech categories than clusters based on other distributional information. As such, anchor points are ideal cues for inferring category labels of novel items heard in idiosyncratic, ambiguous, or misleading (e.g. grammatical violation on part of the speaker) lexical contexts. Additionally, they could protect children from making inferences based on partial or incorrect perception or segmentation of the speech stream. For example, an infant with imperfect auditory perception may analyze the utterance *Where is your dog ?* as something that more closely resembles *Where did you dog ?* which would suggest, incorrectly of course, that *dog* should be clustered alongside other verb-like as opposed to noun-like forms. An error of this sort could delay the acquisition of the word *dog*, as a child might require many more experiences with the word to revise her initial interpretation. Had the same child heard *dog* in the context of an anchor point (e.g. *Look at the X*), this error could have been prevented, or perhaps, reversed more quickly. The high confidence provided by anchor points means that a learner 1) can be more selective in incorporating information from noisy examples, 2) may be less reliant on (or more robust against) noisy examples, and 3) can potentially recover from erroneous inferences more quickly.

Due to their high frequency in language use, distributional anchor points are more likely to be indicative of principled, or more broadly useful distinctions than other distributional cues. In fact, in the distributional learning literature, an important question is which co-occurrence dependencies are most useful for language acquisition? At the word level, a large number of possible relationships can be tracked, such as those between adjacent words or non-adjacent words, spanning phrases, clauses, or even across sentences. Is it better to track co-occurrences in the forward or backward direction, or in both directions simultaneously? While some work has shed light on some of these questions (Mintz, 2002, 2003; Redington & Chater, 1998), early exposure to anchor points could guide a learning system to discover these facts by itself, yielding representations that favor more robust distinctions between lexical categories.

II.2 Fragmentation

Whereas anchoring can provide useful constraints for learning more accurate linguistic abstractions from non-stationary data, a phenomenon, which we refer to as "fragmentation", can lead a learner astray. We 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)[2].

To illustrate how fragmentation might make it more difficult for distributional learners to arrive at useful representations, we re-

[2]While we adopt the term "fragmentation" from the learning literature verbatim, we 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 our work, because the discovery of fragmentation in the data (we will provide a technical definition in section III), does not require that a learner will actually be influenced by it. A learner may have built-in assumptions or inductive biases that reduce or eliminate the impact of fragmented data on the knowledge that is acquired.

turn to the noun category, the topic of this work. 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. For example, the recurrent neural network based language model used by Huebner and Willits (2018) was able to distinguish between 30 semantic sub-categories of nouns, such as labels for animals, toys, and planets. 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. 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 subcategory 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. sub-ordinate 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 we 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 speech 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.

II.3 A visual demonstration of fragmentation

To illustrate fragmentation, we constructed and visualized three hypothetical co-occurrence matrices that differ in fragmentation, shown in Figure 1. Before discussing the demonstration, some housekeeping. First, in all of our analyses rows correspond to nouns, and columns correspond to the contexts in which they occur. It follows that the co-occurrence matrices need not be symmetric, as there are likely to be a different number of contexts than there are nouns in a corpus. Further, because all our co-occurrence matrices are labeled by types rather than tokens, for simplicity we will refer to row and column labels as nouns and contexts rather than noun types and context types, respectively. Second, the black and white elements in the three co-occurrence matrices shown in Figure 1 indicate that either a noun co-occurs with a context (black) or that it does not (white).Third, Our use of binary matrices is for simplicity of demonstration, but our claims hold for co-occurrence values of any range.

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 fragmen-

tation continuum reveals a fundamental trade-off between super-ordinate and sub-ordinate category cues in lexical distributional data: If the co-occurrence patterns at a sub-ordinate or lexical level (idiosyncrasies associated with usage of individual words) are stronger than those at a super-ordinate 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 speech which prompted the question motivating our work: Does the formulaic nature of speech to younger children provide distributional patterns that are better suited for the discovery of the noun category than speech to older children?

III. Goals and Outline

Our primary goal in this chapter is to quantify how fragmented the lexical distributional patterns in speech to children actually are, and how fragmentation might change over the course of development. To answer these questions, we conducted a longitudinal analyses of child-directed speech by measuring fragmentation of the noun category in speech to children at two different ages. The outline of this paper is as follows. First, we discuss technical preliminaries relating to our novel method for quantifying fragmentation, corpus pre-processing, co-occurrence collection, and experimental design in section III. In section IV, we report basic descriptions of the data we have collected in each of our conditions and propose preliminary hypotheses. In section V, we then apply our novel technique to determine whether the noun category is less fragmented in speech to younger compared to older children. A positive result would imply that an incremental, and distributional learning system trained on age-ordered input will be more likely to discover the noun category than a system trained in random order. A negative result would imply that an incremental learner would not benefit by being exposed to age-ordered input, and might be better of trained in random order. In section VI, we used information-theoretic analyses to better understand the potential causes of age-related fragmentation observed in the previous experiment. Borrowing from the related ideas such as anchoring (Cameron-Faulkner et al., 2003) and slot entropy (Matthews & Bannard, 2010), we developed a testable hypothesis to explain why fragmentation is reduced in speech to younger compared to older children. Briefly, we reasoned that entropy-maximizing contexts - frequently occurring contexts which are shared by a large proportion of category members - can shield against fragmentation, and that the prevalence of entropy-maximizing contexts in speech to younger children would provide better protection than speech to younger children. We implemented our hypothesis in a set of quantitative simulations and tested its predictions using child-directed speech data. Lastly, we summarize and discuss the implications of our results in section VII, and end with concluding remarks in section VIII.

III. Tools for Analyzing Non-Stationary Input

.1 The AO-CHILDES Corpus

A study of the distributional patterns in the early language environment requires language data that accurately reflects what children actually hear. As a representative sample of naturalistic speech to children, we selected the CHILDES database, a large collection of transcripts of interactions with children (MacWhinney, 2014). It contains a mixture of transcriptions of structured in-lab activities (such as book-reading, mealtime, and playing with toys), free play in the lab, and in-home recordings.

To create the corpus used in all subsequent analyses, we first obtained all transcripts in the CHILDES database that involve children 0 to 6 years of age from American English speaking households and excluded those for which no

age information was available[3]. After removal of non-adult speech, we obtained 3,251 transcripts containing 272,250 unique word types, and 5,245,298 total word tokens. Considering that a typical American child receives approximately 6.5-11.0 million words per year (Hart & Risley, 2003), the corpus represents approximately 8–14% of lexical input of the average 6-year-old child.

The transcribed corpus was tokenized by splitting on spaces and contractions, and sentence-boundary punctuation (periods, exclamation marks, and question marks) was left in the corpus as individual tokens. This was intended to serve as a very crude way for representing the pauses and prosody that tend to accompany utterance boundaries. The resulting corpus is similar to that used by Huebner & Willits (Huebner & Willits, 2018) in their modeling studies, except that we did not perform any morphological parsing in order to leave intact as many naturalistic properties of the corpus as possible. Lastly, the transcripts were ordered by the age of the target child[4]. For simplicity we will refer to the resulting corpus as AO-CHILDES to indicate that the transcripts it contains are ordered by the age of the target child (AO is short for age-ordered).

While the CHILDES database is no doubt a useful resource to language researchers, it is important to mention several limitation. 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. Second, the CHILDES database is comprised of speech from many hundreds of speakers, and is thereby not ideal for drawing conclusions about the language environment of a single child. Further, due to the large number of different child-caregiver interactions available for a given age, the corpus is less likely to exhibit a consistent scaffolded organization because caregivers scaffold their speech based on the development of a specific child rather than universally by age - which would be required to achieve consistent scaffolding across children. But similar to the first issue, this should make age-related fragmentation more difficult to detect, and thus make an observed age-related effect more impressive. Third, due to transcription irregularity, the same word is often transcribed many different ways (e.g. *playdough*, *playdoh*, *play-dough*, *play-doh*). In our analyses different transcriptions of the same word were left as originally transcribed, and as such are treated as completely separate words in our analyses. Due to the textual as opposed to spoken representation, this means our corpus is nosier than the input that children actually receive (assuming perfect word recognition). Lastly, it is important to note that most of the unique properties of child- as opposed to adult-directed speech, such as prosody, gestures, or joint attention, are not captured by the text-based representation of AO-CHILDES. As is true of any corpus study, we are limited to a textual representation of speech, which potentially misses many important extra-linguistic factors that might further contribute to the scaffolding of children's language input.

I. Partitioning by Age

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. Due to the lack of data for children at the youngest ages, when speech is likely to be the most different (e.g. more anchoring), the results of our analysis would likely underestimate the scaffolded organization of children's language environment. Fur-

[3]Transcripts were obtained from childes-db.stanford.edu on Dec 1, 2017 and processed using code available at https://github.com/UIUCLearningLanguageLab/AOCHILDES

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

ther, the non-uniformity of the age distribution prevented us 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 required us to explicitly split by age, rather than by the number of tokens. To do so, we 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. We will refer to them as sub-corpus 1 and 2, and the age groups they represent as age group 1 and 2, respectively.

II. Selection of Nouns and Non-nouns

In order to collect co-occurrence data, we created a list of frequent nouns in AO-CHILDES. This list was created as follows: First, we part-of-speech tagged AO-CHILDES (using the Python package *spacy* v2.1), and collected all words that were tagged as a noun at least once. Next, we manually inspected the resulting list by removing words that cannot or are extremely unlikely to be used as nouns in child-directed speech. We also excluded plural nouns, proper nouns, interjections, number words, and gerunds. We further excluded words which did not occur at least 10 times in AO-CHILDES. The resulting lists contains 707 singular noun types. Importantly, when collecting co-occurrence data, we do not simply collect data for any occurrence of a word that is in our 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 our noun list. This 2-step procedure has the advantage of 1) using only words that are tagged as nouns in the sentences in which they actually occur, and 2) reducing false positives produced by the tagger.

Additionally, we created a list of non-nouns, which we used as "control" words. We 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. We did so by pairing each noun with a randomly selected word from AO-CHILDES that is approximately matched in frequency.

III. Collecting Co-occurrences

The foundation of all our analyses is the co-occurrence matrix (one for each condition), which we constructed as follows: First, we first collected all sliding windows of size 3 for both sub-corpora (i.e. a word and the immediately adjacently co-occurring words in both the forward and backward direction). Next, we separated the windows based on whether the center word met our criteria for noun or non-noun membership: If the word was in our non-noun "control" word list, it was used for the construction of a non-noun co-occurrence matrix. On the other hand, if the center the word was a noun, it was used to construct the noun co-occurrence matrix. In both cases, the center word (noun or non-noun) labeled the rows of the co-occurrence matrix, and contexts labeled the columns. For both nouns and non-nouns, we created two co-occurrence matrices, one in which the context is defined as the word preceding the center word (backward direction), and another in which the context is defined as the word following the center word (forward direction). We 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. 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, fragmenta-

tion in one or the other (or both) would persist in the combined condition. 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 age group 1), we 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 age group 1, we had to drop approximately 30K noun occurrences to equate the number of nouns in age group 1 with the maximum number of nouns in age group 2 (77,677). Similarly, we dropped about 3K non-noun occurrences in age group 2 to equate the number of non-nouns across the two age groups (104,394). We did so to remove any confound of frequency when comparing age groups.

We investigated the influence of additional variables, treating each as a factor in our experimental design. The full list of factors and factor levels are shown in Table 1. For example, we varied whether we collected original words or their lemmatized forms (rule-based removal of all inflectional morphemes). We included this factor because it is known that 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, atomic events. Additionally, we investigated the influence of including versus excluding punctuation, which can be considered textual markers of prosodic and temporal boundaries in fluent speech. One consequence of punctuation removal, is that when collecting co-occurrences between a 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. We 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, we also collected a "control" co-occurrence matrix for randomly-selected non-nouns matched in type and token frequency to our 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, there may be nothing special about changes in the distributional statistics of nouns in and of themselves.

IV. Quantifying Fragmentation

As we have discussed before, the ideal data for the distributional learner tasked with discovering the noun category is one without any fragmentation (an idealisation which does not occur in practice), where all members pattern identically with their contexts. We 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". Because virtually all work related to distributional analysis and word classes has focused on evaluating algorithms of category induction, there is a paucity of methods evaluating the data itself, and especially in terms of fragmentation. To rectify this, we developed our own measure for quantifying fragmentation.

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 correla-

tion 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 our goal is to quantify the degree to which a set of co-occurrence vectors instantiate a single, shared co-occurrence pattern, we require a multivariate tool. For example, multiple correlation is the correlation between one variable's observations (i.e. is the word a noun[5]?) 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 "fragmenting" lexically specific patterns.

IV.1 Singular Value Decomposition

To identify the best linear transformation, we opted for singular value decomposition of the noun co-occurrence matrix[6]. For simplicity, we 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 we are interested in the amount of variance explained by the first singular dimension, we can obtain our results directly using SVD, as opposed to a two-step procedure consisting of 1. the computation of the prototype vector, and 2. the computation of how much variance it accounts for using multiple correlation.

Our 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), we 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 proto-

[5]In our case, this variable would simply be a vector of ones, indicating that each word in our noun co-occurrence matrices is a noun.

[6]We could have used multiple regression, but it explicitly models error and intercept terms which we did not require.

type vector. Let fragmentation = f, then

$$f = \frac{(\sum_{i=1} s_i) - s_1}{\sum_{i=1} s_i} = 1 - \frac{s_i}{\sum_{i=1} s_i} \qquad (1)$$

where s_i is the i-th singular value.

Given the novelty of our technique, we preemptively provide some notes of caution for how not to interpret and use our method. First, our measure of fragmentation was not developed for the evaluation of a distributional learning system; it requires full access to the raw data, and is therefore a method for evaluating the data itself, not what a system can learn from it. Second, our measure does not quantify the extent to which members of one lexical category can be discriminated or are distinct from words that belong to another category. Fragmentation is defined separately for each category because it quantifies how much the members of a single category are distinct from each other and not from members of another category. Lastly, fragmentation shares little with common metrics such as accuracy (e.g. the percentage of words classified as nouns that are actually nouns), or completeness (e.g. the percentage of words that are actually nouns that are classified as such). That said, our measure of fragmentation should strongly predict the classification success of distributional learning systems.

V. Information Theoretic Tools

In addition to the multivariate analysis provided by our measure of fragmentation, we also evaluated bivariate relationships between pairs of individual lexical distributions. To do so, we adopted standard quantities from information theory.

We can quantify the statistical strength between two random discrete variables (such as two lexical distributions) by computing how their entropy - the amount of uncertainty about the outcome of a variable - changes when the outcome of one variable is known. For example, the entropy of randomly retrieving a word from a bucket of nouns, which we will refer to as the variable X, is the amount of uncertainty we have about the exact noun that will be chosen. But nouns usually occur in lexical contexts which help us predict them, and this means that our predictions are often better than random. With this in mind, let us refer to observations of the lexical context (in the forward or backward direction, it does not matter) as Y. The outcome of Y should provide information that reduces our uncertainty about the outcome of X. The uncertainty about the outcome of X that remains is referred to as the conditional entropy of X given Y, written as $H(X|Y)$. In other words, by conditioning X on Y, we are effectively measuring the amount of uncertainty left over after accounting for our knowledge of Y. If the outcomes of Y strongly predict the outcomes of X, the conditional entropy would be small, and vice versa. The conditional entropy is given by

$$H(X|Y) = \sum p(x,y) \log_2 \frac{1}{p(x|y)} \qquad (2)$$

where $p(x,y)$ is the joint probability, the probability of observing outcome x and y together, and $p(x|y)$ is the conditional probability, the probability of observing outcome x given that y has already occurred. This formula makes it clear that the conditional entropy is simply a weighted average of the log-transformed inverse of the conditional probabilities. Due to the inverse (which converts probabilities into the number of equally likely outcomes), the higher the conditional probabilities, the lower the resulting conditional entropy. We hope this helps readers who are familiar with conditional probability make the leap to entropy based measures.

The relation between fragmentation and conditional entropy is as follows. Because the conditional entropy is a bivariate measure (there are only two variables, X and one Y), and fragmentation is a multivariate measure (each column is a distinct variable), the two are only related in the case when there are no multivariate correlations - that is, when nouns do not systematically occur in mutually

shared contexts. This can occur if fragmentation is driven purely by lexical, rather than sub-category specificity. Under such circumstances, fragmentation and conditional entropy would be inversely associated. To elaborate, if fragmentation is driven entirely at the lexical level, higher fragmentation would make it easier to predict observations of one variable (X or Y) from the other, and this would reduce the conditional entropy. On the other hand, if fragmentation is low (i.e when nouns are more likely to occur in the same contexts as other nouns), predictability is sacrificed, this would increase the conditional entropy. A similar relationship has been demonstrated empirically by Cassani et al. (2018), who found a negative correlation between the average conditional probability - inversely related to conditional entropy - and lexical classification accuracy. Their conclusion that "categorization works best for words which are [...] hard to predict given the co-occurring contexts" is closely associated with our claim that high predictability below a superordinate category (high fragmentation) is disadvantageous for a distributional learner tasked with the discovery of the superordinate category. The novelty of our work lies in demonstrating how this predictability changes over developmental time.

Estimates of conditional entropy, and other measures based on entropy, are influenced by the number of possible outcomes of X, and Y. The number of possible outcomes (i.e. the number of word types) of X and Y corresponds directly to the number of rows and columns in our co-occurrence matrices, respectively. For example, an increase in the number of columns in our co-occurrence matrices, which would increase the number of types in Y, would produce - everything else being equal - smaller estimates for $H(X|Y)$, while an increase in the number of rows would not influence estimation. Similarly, an increase in the number of rows would produce larger estimates for $H(Y|X)$, while an increase in the number of columns would not influence estimation. Because we plan to compare conditional entropies across our two sub-corpora, we must take the necessary steps to reduce any bias that would result from comparing two unequally sized co-occurrence matrices. Towards this end, we randomly permuted the rows and columns of each co-occurrence matrix, calculated the mean of the conditional entropies across all permutations, and subtracted from this the original estimation on the unshuffled data.

IV. Simple Statistics of Nouns in Caregiver Speech

In this section, we carefully examine the shapes and sums of the co-occurrence matrices generated in each of our conditions. We report these quantities because they are important methodological and theoretical considerations for interpreting subsequent demonstrations, and because they provide an overview of the data which we will re-visit in the remainder of this chapter. As a reminder, we collected co-occurrence data varying in 6 factors: 1) the age of the target child, 2) context direction (forward vs. backward), 3) lemmatization (True vs. False), 4) punctuation (intact vs. removed), 5) normalization (dividing each element by its column sum vs. none), and 6) whether co-occurrences were collected for nouns ("experimental" words) or a frequency-matched set of non-nouns ("control" words).

The number of rows that make up our co-occurrence matrices are equivalent to the number of noun types or non-noun types, depending on whether we are collecting the lexical context of nouns or non-nouns. Similarly, the number of columns is equivalent to the number of context word types. These values are shown in Table 2. In the table, each pair of rows contrasts age group 1 and age group 2 (shown in column 1), for a particular combination of context direction, lemmatization, and punctuation (whose values are shown in columns 2-4). For simplicity, in the tables below we did not include normalization as a factor because it does not influence the collection of co-occurrence data. Normalization, however, will play a crucial role In the analysis of fragmentation in the next section.

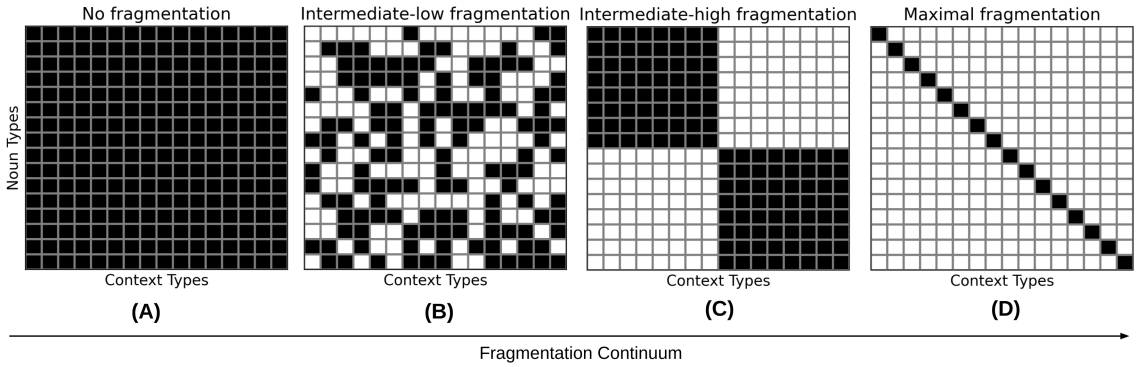

Figure 1: A visual demonstration of the fragmentation continuum, with hypothetical co-occurrence matrices (nouns in rows, contexts in columns) that exhibit either no fragmentation (A), intermediate-low fragmentation (B), intermediate-high fragmentation (C), or maximal fragmentation (D).

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 1: Factors influencing the construction of co-occurrence matrices.

Condition				Nouns		Non-Nouns	
age	direction	lemma	punctuation	rows	columns	rows	columns
group 1	backward	False	intact	653	1,998	688	3,494
group 2	backward	False	intact	679	2,571	662	3,788
group 1	backward	False	removed	653	2,732	688	4,668
group 2	backward	False	removed	679	3,066	662	4,957
group 1	backward	True	intact	653	1,857	713	2,917
group 2	backward	True	intact	680	2,339	691	3,106
group 1	backward	True	intact	653	2,454	713	3,900
group 2	backward	True	intact	680	2,737	691	4,047
group 1	forward	False	intact	653	2,150	688	3,925
group 2	forward	False	intact	679	2,727	662	4,359
group 1	forward	False	removed	653	3,514	688	4,721
group 2	forward	False	removed	679	3,790	662	4,907
group 1	forward	True	intact	653	1,767	713	3,128
group 2	forward	True	intact	680	2,151	691	3,402
group 1	forward	True	removed	653	2,971	713	3,848
group 2	forward	True	removed	680	3,082	691	3,880

Table 2: Number of word types captured by co-occurrence matrix in each condition.

The quantities in the table indicate several facts worth noting. First, the total number of co-occurrences collected (equivalent to the sum of a co-occurrence matrix) was purposefully limited in order to enable fair comparisons across age groups. As noted in the previous section, this was done by counting the number of co-occurrences obtained from each sub-corpus, and then using the smaller of the two counts to limit the number of co-occurrences that is collected for the other (larger) sub-corpus. This control measure ensured that the number of co-occurrences were equal across age groups, and resulted in 77,677 total co-occurrences of nouns with their contexts, and 104,394 total co-occurrences of non-nouns with their contexts.

Second, the shape of each co-occurrence matrix varied freely across conditions. We thought it best not to artificially constrain which co-occurrences should be collected in order to capture the statistics of nouns and non-nouns in each sub-corpus with the least amount of intervention as possible. Because we did not limit our word lists or the set of possible context words such that each must occur at least once in each sub-corpus, it was possible that the resulting co-occurrence matrices varied in the number of types that were captured. For example, it is in principle possible that our procedure for collecting co-occurrences would have returned perfectly disjoint sets of co-occurrences, meaning that neither any noun (or non-noun) or context word needed to be present in both co-occurrence matrices. In practice, however, we observed a large - though not perfect - amount of type overlap. Variation in shapes must be considered when interpreting the results of comparisons across matrices with different shapes. More importantly, the number of rows and columns can be used as a first approximation of the lexical diversity of each sub-corpus. For example, Table 2 shows there are consistently fewer noun types in age group 1 compared to age group 2 (speech to younger vs. older children). This can be explained in terms of reduced lexical diversity - noun diversity in particular - in speech to younger children. This is a common finding in corpus analyses of child-directed speech. We also noted a consistent increase in the number of contexts in which nouns occur from age group 1 to age group 2. For example, there are roughly 600 more context types (1998 vs 2571; row 1 and 2) in age group 2 when co-occurrences of non-lemmatized word forms are collected in the backward direction. This means that, as children grow older, they encounter nouns in an increasingly more diverse set of lexical contexts. This is a first indicator that the distributional pattern of nouns has the potential to be more fragmented in speech to older children. Fragmentation is in part driven by lexical specificity - the tendency of specific words to pair with specific other words - and greater lexical diversity is a prerequisite for greater lexical specificity.

Finally, we wish to draw attention to the influence that the removal of punctuation has on the number of context types in which nouns occur. For example, Table 2 shows the removal of punctuation adds approximately 700 non-lemmatized context types when co-occurrences are collected in the backwards direction (compare rows 1 and 3) and approximately 1,300 non-lemmatized context types (compare rows 9 and 11) when co-occurrences are collected in the backwards direction . This trend holds across conditions, and suggests utterance boundaries may play a critical role in the discovery of the noun category in child-directed speech. Punctuation also markedly increases - but to a lesser extent - the number of contexts of randomly selected non-nouns. This is evidence that punctuation symbols are more prevalent in the context of nouns relative to non-nouns, a finding which we will revisit in the following sections.

To summarize: First, our demonstration points to a substantial increase in lexical diversity at age group 2 compared to age group 1, consistent with prior findings that speech to younger children is less lexically diverse (Broen, 1972; Foushee et al., 2016; Hayes & Ahrens, 1988; Kirchhoff & Schimmel, 2005; Phillips, 1973). Second, nouns occur with a smaller set of contexts than do non-nouns,

and this difference between nouns and non-nouns is greatly magnified at age group 1 compared to age group 2, and is also magnified when punctuation symbols are considered a unit in the analysis of lexical context. The consequences of these differences will be further explored in the next sections.

V. Age-Related Fragmentation of Nouns

Next, we investigated how fragmented the noun category is in speech to younger children compared to older children. 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), we 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, we predicted that the pattern of increasing fragmentation with age is specific to nouns, and therefore would not extend to our control word list (non-nouns).

As described in a previous section, fragmentation is obtained using SVD by computing the amount of variance in the word co-occurrence matrix that is not explained by the first singular value of the co-occurrence matrix. The results of our fragmentation experiment is shown in Figure 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. We are primarily interested in a potential effect of age group, and whether such an effect is influenced by any of the other manipulations.

I. Results

We 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, we 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 (std). Intuitively, 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 co-occurrence pattern that best fits the category "noun". 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, we 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 we 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 (notably, punctuation would often occur in the forward direction if it were not removed). This interaction between forward 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 speech 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.

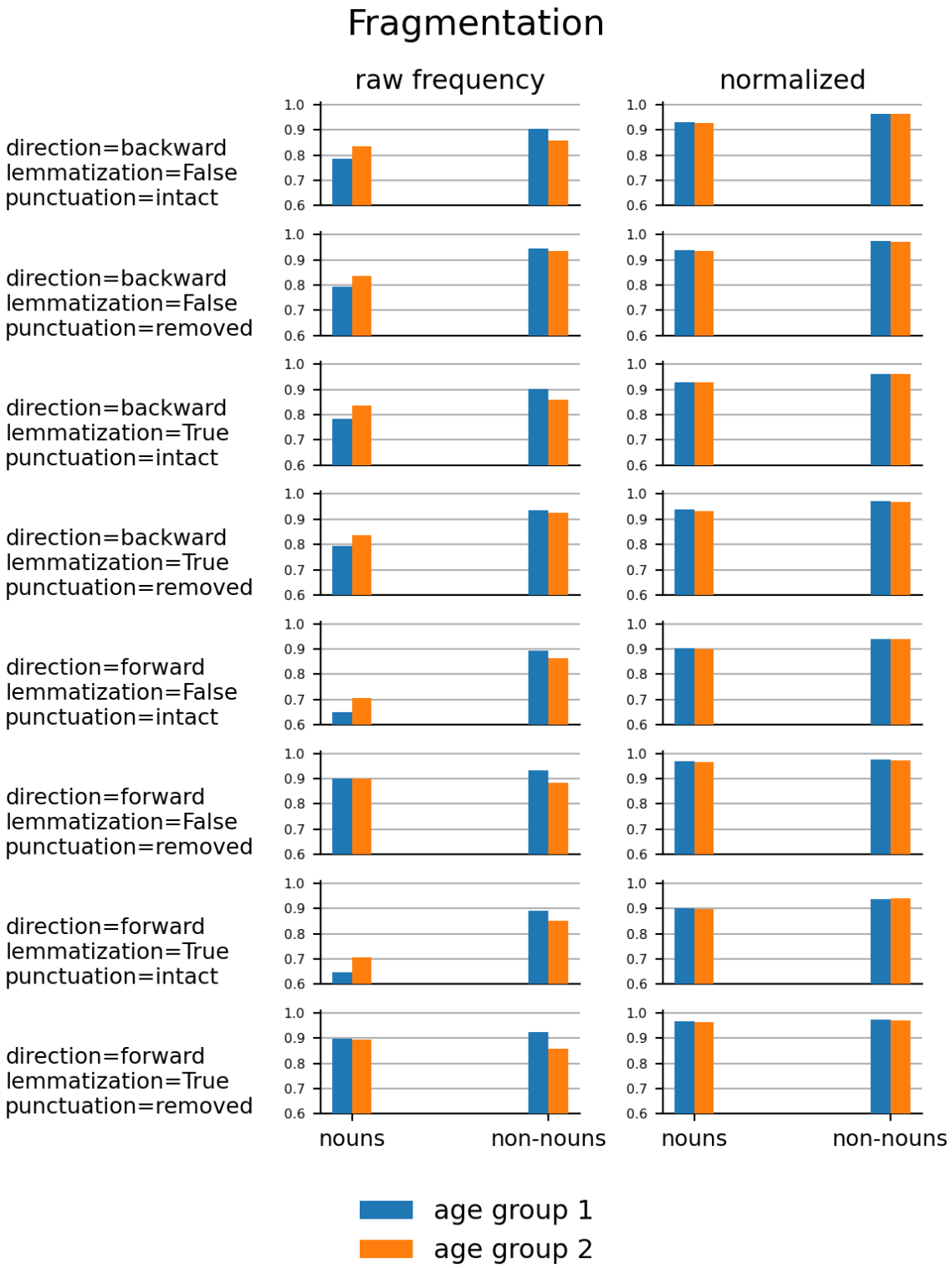

Figure 2: Fragmentation of the noun category (left panels), and a random set of non-nouns (right panels) for each condition. The factors context direction (forward vs. backward), lemmatization, and punctuation vary across rows, and age, word list, and normalization vary across rows. Fragmentation is 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.

Next, we focus on the 16 comparisons shown in the right panels of Figure 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) 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 we included normalization as a factor in our comparisons. Notably, whereas we have seen 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, we think it reveals an important clue related to the similar null-effect we observed in conditions in which punctuation had been removed. This will be discussed further below, and in Experiment 3.

II. Comments and Follow-up Questions

In a previous section, we established the intuition that the proportion of variance explained by the first singular dimension is conceptually equivalent to the proportion of variance accounted for by a "category-defining", shared co-occurrence pattern. Because subsequent singular dimensions are aligned with respect to the first dimension, subsequent dimensions can be thought of as deviations from this shared frequency pattern. We can think of these subsequent dimensions as a fragmentation of the underlying shared frequency pattern. Hence, one way to interpret our demonstration is that as children grow older, the nouns they hear occur in increasingly more divisive lexical contexts which obscure the fact that nouns belong to the same class of words.

Part of the reason why we included so many conditions 2x2x2x2x2x2=64) is because our measure of fragmentation produces a single value per condition, which precludes testing for statistically significant differences. Nonetheless, the fact that we observed the same trend across multiple conditions for our experimental noun word list, and a different but consistent trend for our non-noun word list, suggests that our results are likely not a chance fluctuation or an idiosyncrasy of our setup.

As predicted, we found that the noun category is less fragmented in speech to younger compared to older children, and that this effect is not due to a shift in global corpus statistics (the opposite trend was observed for non-nouns). This reinforces the notion that there is something special about nouns, not necessarily just in how they are conceptually processed by children (Gentner, 1982), but also in how their distributional statistics change across developmental time. Not knowing ahead of time that nouns exist, our demonstration suggests that a statistical learner tracking the lexical contexts in which words occur is more likely to discover the noun category in speech directed to 1-3 year-olds compared to 3-6 year-olds. This makes ecological sense in that incremental exposure - to superordinate category information before subordinate category information - temporally separates the conflicting distributional signals generated at different levels in the hierarchy of lexical classes: It is easier to learn that nouns are a distinct category before learning that nouns are sub-divided into smaller classes (e.g. animates vs in-animates, or birds vs. mammals).

Lastly, we think that the two null effects, in conditions in which punctuation was removed or the columns of the co-occurrence matrices were normalized, are worth further investigation. The null-effect due to normalization suggests that the age-related difference in fragmentation is driven by a small number of contexts that explain a disproportionately large amount of variance in the data. Interestingly, this pattern of results is precisely what one might predict based on a stronger presence of anchor points in speech to younger children.

In the next section, we explore this possibility, by diving deeper into the potential forces that drive the observed age-related increase in fragmentation.

VI. The Entropy-Maximization Hypothesis

I. Converging Evidence

In the previous section, we demonstrated reduced fragmentation of the noun category in speech to younger compared to older children. A natural follow-up question is: How are the lexical distributions changing over developmental time to produce this result? We have already identified two important clues, namely that the effect of age-related fragmentation requires the presence of punctuation symbols (at least in the forward direction), and the preservation of the relative differences in total variance accounted for by different contexts (i.e. no column normalization). We think these two observations are pointing towards the same explanation. Given that punctuation symbols are extremely frequent in the AO-CHILDES corpus, and much more frequent in the age group 1 portion of the corpus, their corresponding columns in the co-occurrence matrices must explain a lot more variance than other columns. Preserving this relationship may be required to observe an age-related increase in fragmentation - at least in the forward direction. It is less clear that punctuation alone can explain differences in fragmentation in the backward direction, because removing punctuation did not have the same effect on fragmentation when co-occurrence was counted in the backward direction as it did when it was counted in the forward direction. However, it is possible that a similar type of explanation exists for the backward co-occurrences as well, but involving anchor tokens other than punctuation. In our next demonstration, we are testing for the existence of such anchor tokens that can explain the difference in fragmentation in the backwards direction that punctuation does in the forward direction.

I.1 Conditional Entropy

To start, we will focus on developing a better understanding of how the distributions of co-occurrences in the forward direction are changing with age, and then extend our understanding to explain fragmentation more generally. We begin by noting the results of corpus studies of child-directed speech, which have demonstrated that nouns frequently occur in the utterance-final position (Brent & Siskind, 2001; Swingley & Humphrey, 2018). In the AO-CHILDES corpus, the boundaries of utterances are marked by punctuation, which in most cases is a period, and less frequently an exclamation or question mark. Because nouns frequently occur in utterance-final position, a large portion of co-occurrences in the forward direction involve nouns followed by punctuation symbols. When punctuation is removed, however, the right neighbor of a noun in utterance-final position is the first word in the subsequent sentence, and, importantly, this makes the forward-direction context distributions of nouns more diverse, and much less predictable [7].

This change in the distributional relationship between nouns and their neighbors can be quantified using conditional entropy. By itself, entropy measures the average difficulty of predicting the outcomes of a random variable. In our case, we are dealing with two random variables; underlying our observations of nouns and their context are two discrete random variables which we will refer to as X and Y, respectively [8]. We can compute either the entropy of contexts conditioned on nouns ($H(Y|X)$), or

[7] This may appear to some a minor point about the idiosyncrasies of textual data - after all, punctuation is not explicitly represented in speech. However, punctuation in text data can mark prosody, and/or prolonged periods of silence between utterances in fluent speech. To the extent that punctuation symbols correlate with the presence of these acoustic markers, children *do* have access to the distributional patterns represented in AO-CHILDES by punctuation symbols.

[8] In our co-occurrence matrices, we treat rows as samples drawn from an unknown discrete random variable X, and columns as samples drawn from a different unknown discrete random variable Y.

the entropy of nouns conditioned on contexts ($H(X|Y)$). In the former, we observe which noun has occurred and use this information to predict which context will occur, while in the latter, the relationship is reversed. To illustrate the difference, consider a noun followed by a period. Because periods frequently follow nouns in AO-CHILDES, this results in high predictability and therefore a small value of $H(Y|X)$. However, due to the semantically uninformative nature of punctuation, it is incredibly difficult to predict in the reverse direction, and this results in a large 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.

I.2 Entropy-maximizing Contexts

Punctuation symbols present an intuitive opportunity for explaining how conditional entropy should change as structural factors underlying lexical context distributions evolve over developmental time; but they are likely not the only contexts to be affected. There are potentially many other highly frequent noun contexts that behave similarly, but are less frequent. We will refer to all such contexts as entropy-maximizing contexts. These are contexts that occur relatively indiscriminately with many members of a category. By doing so, these contexts drive up $H(X|Y)$, the difficulty of predicting which member is likely to occur given that we have already observed the entropy-maximizing context. We can think of many context words that are likely to have this property with respect to the noun category: Forward co-occurrences such as *and*, *to*, *with*, and backward co-occurrences such as *the*, *a*, *that*, and *my*. These contexts are all relatively semantically uninformative, and can therefore occur with most nouns with relatively similar likelihood.

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. However, we 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).

The reason we think entropy-maximizing contexts are important for explaining the reduced fragmentation in age group 1 are twofold: First, entropy-maximizing contexts - just like anchor points - are likely over-represented in speech to younger compared to older children. We can infer that punctuation symbols more frequently follow nouns in speech to younger compared to older children based on studies which have demonstrated that nouns are more likely to occur in utterance-final positions, and we have confirmed this in an offline analysis. Second, the absence of age-related fragmentation of nouns observed when the columns of the co-occurrence matrix were normalized is predicted by the notion of entropy-maximizing contexts: 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. Collectively, these two lines of evidence point towards an explanation of the age-related increase in fragmentation of nouns in terms of a disproportionate number of entropy-maximizing contexts in speech to younger children.

I.3 Predictions

Our hypothesis essentially requires that over development, there is a distribution shift from

nouns occurring in entropy-maximizing contexts (e.g. anchor points like punctuation symbols and other contexts 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, such as *delicious X* , *fluffy X*). This shift can be empirically tested using our data. To test this hypothesis, we first translate this general notion into precise quantitative predictions about the pattern of conditional probabilities we expect to observe under the assumption that our hypothesis is correct. We obtained our predictions from a set of simulations in which artificially generated co-occurrence matrices, modeled after those we have been using in our previous experiments, were compared under different assumptions about the number and importance of entropy-maximizing contexts. Specifically, we compared the conditional entropy $H(X|Y)$ across the conditions described below.

The main comparison was between two matrices intended to mimic the average shape and sum of co-occurrence matrices obtained for age groups 1 and 2. They differed from the latter matrices only in that their values were drawn from a pseudo-Zipfian distribution [9]. While the randomly generated matrix intended to simulate age group 2 was left as-is, we randomly replaced a subset of columns in the matrix intended to mimic age group 1 with columns that represent our notion of entropy-maximizing contexts. Rather than sampling from the same pseudo-Zipfian distribution we used to populate the columns corresponding to non-entropy-maximizing contexts, we populated the columns of entropy-maximizing contexts using a random uniform distribution which maximizes entropy. More importantly, in order for entropy maximization to take place, we sampled more frequently from these contexts relative to other contexts. As a result, the columns corresponding to entropy-maximizing contexts were more densely packed, and therefore more likely to result in higher entropy estimates. We did not de-bias our entropy estimates in these simulations, because our use of marginal distributions to generate co-occurrence matrices is inherently unbiased [10].

The results of our simulations are shown in Figure 3. As the number of entropy-maximizing contexts is reduced (from right to left across the x-axis), the conditional entropy $H(X|Y)$ (the leftover uncertainty when predicting a noun and knowing its context) of the simulated age group 1 eventually rises above the conditional entropy of the simulated age group 2. The point at which the crossover occurs depends on the fraction of co-occurrences that involve entropy-maximizing contexts: When fraction=2, one half of all co-occurrences involve entropy-maximizing contexts, and when fraction=4, one quarter of all co-occurrences involve entropy-maximizing contexts, and so on. Naturally, the smaller the fraction, the larger the effect of entropy-maximizing contexts. To summarize, our 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 prominent role in the distributional patterns of nouns to younger children, we expect that $H(X|Y)$ will be greater for sub-corpus 1 compared to sub-corpus 2 of CHILDES. If, on the other hand, we find the opposite pattern, we would have to either revise our understanding of entropy-maximizing contexts, or conclude that their role in combating fragmentation in speech to younger children is unwarranted. Finally, we do not have specific predictions for how $H(Y|X)$ should vary between age groups,

[9] 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.

[10] Because rows and columns were populated randomly - drawing from independent marginal distributions - the co-occurrence matrices in our simulations do not exhibit any joint structure. Put differently, there are no "real" interactions between rows and columns other than those that result due to chance. Both our simulations and the de-biased estimates of conditional entropy used for actual co-occurrence data are therefore relative to chance, and are thus comparable.

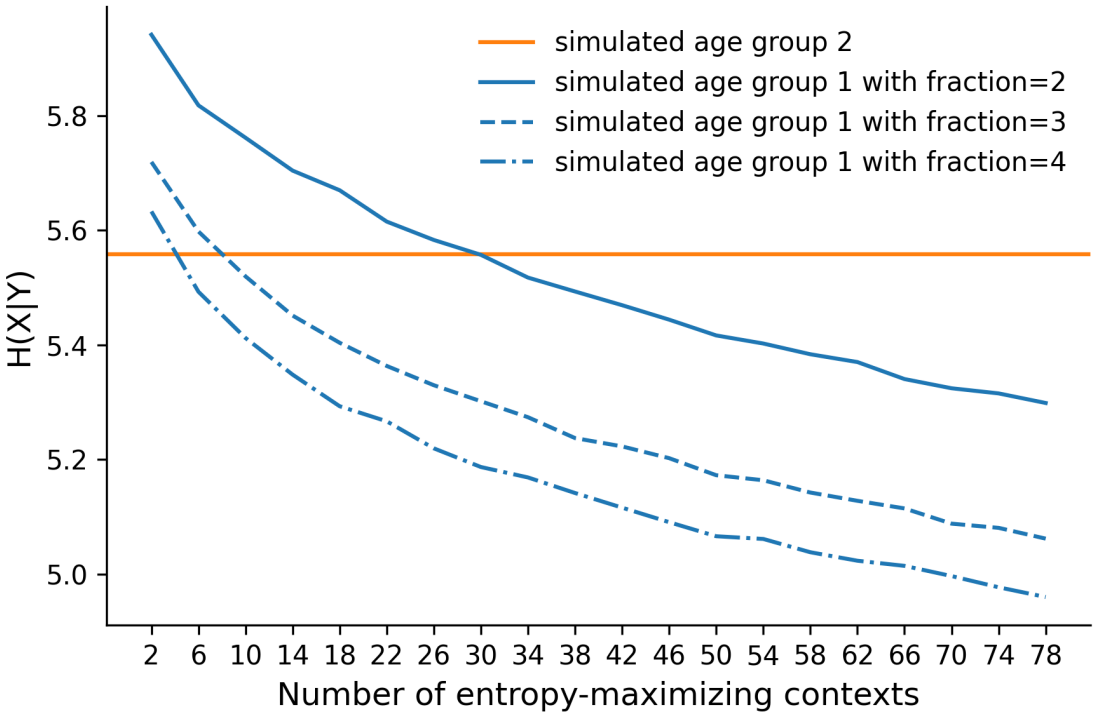

Figure 3: The relationship between $H(X|Y)$ and the number of "entropy-maximizing" columns for artificially generated co-occurrence matrices. $H(X|Y)$ is the conditional entropy of pseudo-nouns (X) given the contexts (Y) in which they occur. The shapes and sums of simulated co-occurrence matrices were modeled after AO-CHILDES data (650 rows and 2000 columns for age group 1; 680 rows and 2600 columns for age group 2), which allowed us to control for effects of matrix shape on entropy estimates. Simulated data was sampled either from a pseudo-Zipfian distribution (for non entropy maximizing contexts) or a random uniform distribution (for entropy-maximizing contexts). For each simulation of age group 1, we varied the fraction of co-occurrences that involve an entropy-maximizing context. Each fraction corresponds to a different line style in the figure.

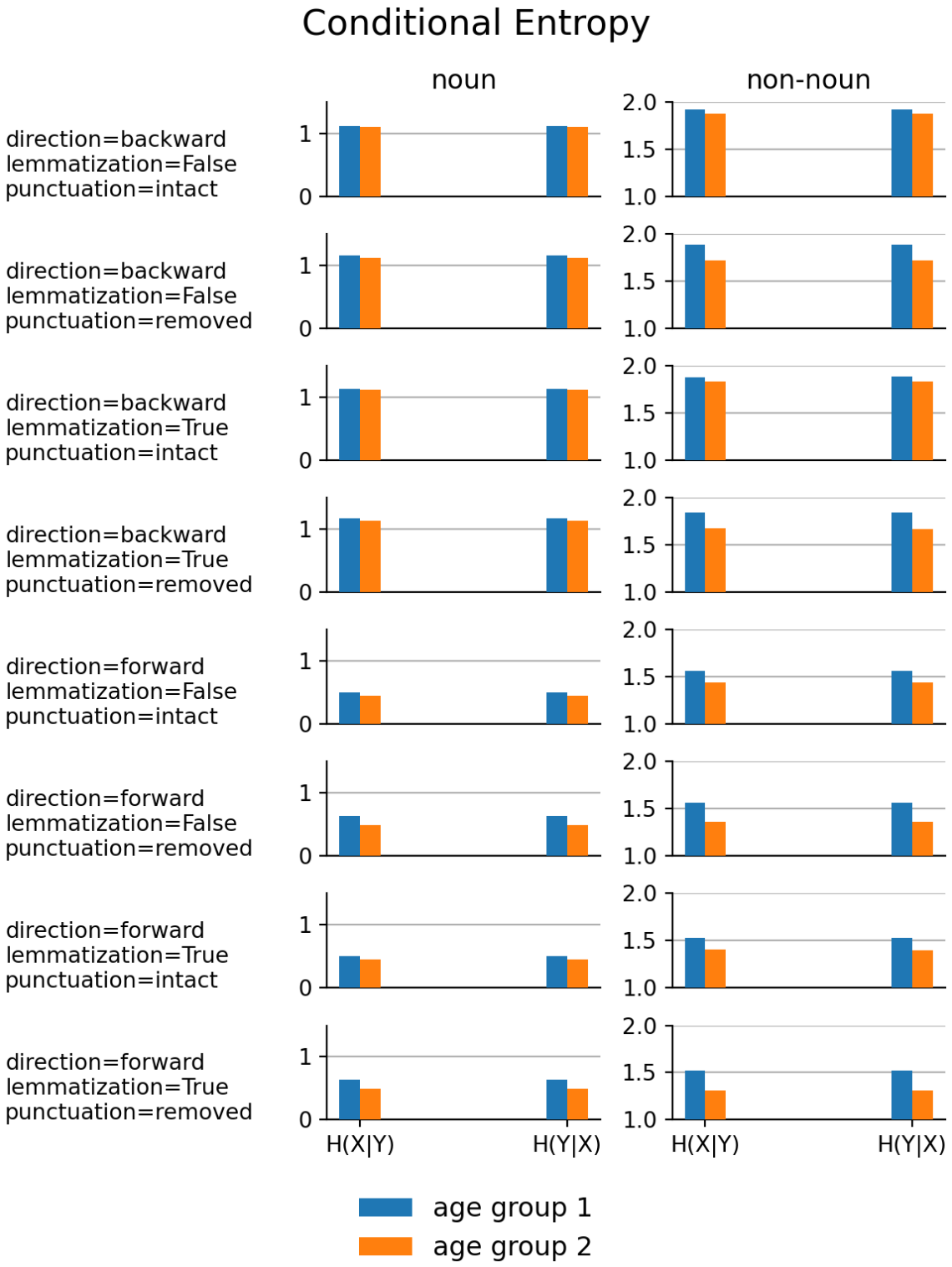

Figure 4: Comparison of conditional entropies between age group 1 and 2. Left panels are for nouns, and right panels for non-nouns. The three factors punctuation, lemmatization, and context direction are varied across rows. The y-axis units are bits.

under our hypothesis.

II. Results

The results of our conditional entropy analyses of AO-CHILDES are shown in Figure 4. We excluded conditions in which co-occurrence matrices were normalized because conditional entropy requires raw frequency as input. Our 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, 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, we examined differences for $H(X|Y)$, for which we developed predictions. These values are represented by the first group of bars in the left panels. Across all eight conditions, we observed a larger $H(X|Y)$ for age group 1 compared to 2, as predicted. The average difference was 0.06 +/- 0.05 (std), ranging from differences as low as 0.01 bits to 0.15 bits[11]. Interestingly, this effect appears independent of punctuation, which suggests that punctuation symbols are not the only entropy-maximizing contexts that can potentially combat fragmentation of the noun category. That said, it is also possible that our analysis is simply not useful for pining down the role of punctuation symbols in fragmentation, because their removal most likely resulted in counting of co-occurrences with words that are *equally unhelpful* for predicting the identify of nouns; the first word of a new sentence often has little to do semantically with the final noun of the previous sentence. This means that even if punctuation played a prominent role in combating fragmentation in sub-corpus 1 of AO-CHILDES, our method of counting co-occurrences across sentence-boundaries may have interfered with our ability to detect such an effect. Nonetheless, some evidence for the idea that punctuation symbols are important entropy maximizing contexts is the 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, we compared $H(Y|X)$ across age groups, 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. We found that $H(Y|X)$ behaves qualitatively similar to $H(X|Y)$ in that it was larger in age group 1 across all eight conditions. On average, the values for age group 1 were 0.06 +/- 0.05 (std) larger, with differences ranging as low as 0.01 bits to 0.15 bits.

Interestingly, the patterns we have so far discussed also hold when tracking co-occurrence statistics of non-nouns (shown in the right panels of Figure 4). Whereas the entropy estimates are overall larger (ranging from 1 to 2 bits as opposed to 0 to 1 bit), their qualitative pattern across conditions is virtually identical 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.

Overall, the results of comparing $H(X|Y)$ and $H(Y|X)$ across age groups show that, over the course of developmental time, it becomes increasingly less difficult to predict nouns given their lexical contexts and vice versa in speech to children. More importantly, we 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 speech to children. The increase in syntactic complexity associated with age, such as lexical and constructional diversity are all likely to be involved in this shift, and cumulatively, peel away the protective layers

[11] 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.

of entropy-maximizing contexts that present nouns as a more coherent category in children's early language input.

Had one been forced to guess the pattern of conditional entropies naively, one should have based such a guess on the difference in the shape of the two co-occurrence matrices. This would have led one to predict that $H(X|Y)$ would be smaller, not larger, for age group 1. This can be verified by a brief look at Figure 3: The further one moves away from the regime in which entropy-maximization takes place (from left to right along the x-axis), the more likely it is that the curves representing $H(X|Y)$ of age group 1 fall below the simulated conditional entropy of age group 2. We interpret the fact that we observed precisely the opposite pattern as evidence that something more involved than simple differences in the number of nouns or their contexts must be invoked to explain the difference in fragmentation, such as entropy-maximization.

In order to have any effect, entropy-maximizing contexts must be relatively rare - if every context were 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. Indeed, our simulations show that entropy-maximization in speech to younger children only predicts our 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 30 contexts, see Figure 3). This relates to our previous finding that the age-related difference in fragmentation disappeared when normalizing the co-occurrence matrices prior to analysis. Normalization removed differences in the amount of total variance that each column could account for, which could have eliminated the potentially disproportionate contribution of a potentially small set of contexts - those that are entropy-maximizing.

VII. General Discussion

Language acquisition unfolds incrementally over a time span covering many years from infancy to early adulthood, and is accompanied not only by changes in neural, cognitive, and social development, but changes in language input. This chapter addresses questions of broad interest in language acquisition regarding the ways in which caregiver input may be structured to scaffold the learning process. In particular, our work emphasizes the importance of viewing language acquisition as an incremental learning problem. The incremental nature of learning introduces novel questions and problems that few researchers have addressed in the past. For example, incremental learning systems are vulnerable to distribution shifts, that occur when language samples are generated from a non-stationary distribution. These distribution shifts are well documented in the child language acquisition Literature (Gallaway & Richards, 1994; Lieven, 1994; Montag et al., 2015; Pine, 1994; Richards, 1994; Snow & Ferguson, 1977), but are largely ignored in most models of language learning. In order to build more ecologically and developmentally plausible models of language acquisition, we first require a better understanding of the incremental organization of language input. Toward that end, we presented a novel technique for measuring qualitative shifts in co-occurrence data across time, and applied it to co-occurrence patterns of nouns in child-directed speech.

Our method is based on the notion of "fragmentation", which occurs when the context word distributions of a lexical category, such as nouns, are made less similar in terms of their distributional co-occurrence statistics. It can be helpful to think about different factors that make the statistical regularities of nouns less similar. These can be idiosyncratic phonological and morphological factors (e.g. consonant-initial nouns frequently co-occur with *a* and vowel-initial nouns frequently co-occur with *an* instead), but more often these are due to syntactic and semantic constraints, resulting

in co-occurrence patterns that divide the noun category into subcategories. These distributional patterns invariably reduce the similarity between these subcategories, and thereby obscure the fact that members of different subcategories are still members of the larger noun category. In such a situation, a distributional analysis may be able to identify only separate sub-categories (fragments) of the larger noun category rather than the noun category itself. We have suggested this can be detrimental in the early stages of language acquisition, during which initial representations have the potential for influencing (negatively or positively) later learning.

Our novel analysis method revealed that the distributional patterns within the noun category is less fragmented in speech input available to younger compared to speech input available to older children. As we have described, this fragmentation is disadvantageous for the discovery of larger lexical classes from co-occurrence data. To explain how speech to younger children is less fragmented, we developed a hypothesis building on the idea of anchoring (Cameron-Faulkner et al., 2003) (and the closely related "slot entropy" discussed at length by Matthews and Bannard (2010)). To reduce fragmentation of the noun category - inevitable in large corpora of natural language data - nouns must repeatedly and systematically occur with a small subset of lexical contexts, which we termed entropy-maximizing contexts. This is precisely what anchor points do in speech to children. Working with a model in which entropy-maximizing contexts are allowed to "anchor" nouns in a small subset of frequent contexts, we generated quantitative predictions for the difficulty of predicting which noun occurred given information about its lexical context. The predictions of this model aligned with the actual results observed in child-directed speech. Overall, our work supports the idea that 1) the noun category is less fragmented in speech to younger children, and that 2) entropy-maximizing contexts provide a strong theoretical framework for understanding this difference in age-related fragmentation.

I. Implications for theories and models of language acquisition

Our hope is that a better understanding of how distributional regularities in speech to children change across development sets the stage for more developmentally plausible models of lexical category acquisition. In particular, there are two age-related phenomena in the acquisition literature which clearly illustrate the need for models and explanations based on gradual and scaffolded formation of linguistic representations. These two phenomena are progressive differentiation, and the noun bias. We will discuss each in turn.

I.1 Hierarchically-structured Knowledge and Progressive Differentiation

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 (Clark et al., 1985; 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 (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 simple 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.

Our work 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.

Our work also indicates that as children become more familiar with broad linguistic distinctions such as part-of-speech, their language environment gradually reveals finer-grained distinctions that might otherwise slow learning at an early age. Our notion of entropy-maximizing lexical contexts suggests that the highly repetitive and simple constructions in speech to younger children may play an important role in concealing these finer-grained distinctions during the earliest stages of learning. In this light, the benefit of anchor points and other entropy-maximizing lexical contexts goes beyond supporting specific linguistic abstractions (Cameron-Faulkner et al., 2003): Such distributional cues could help learners avoid complex hypotheses that are unproductive at an early stage of acquisition or are costly to revise. We think that novel insights can be gained from conceptualizing language acquisition not as a straight march towards abstract knowledge, but as a process that uses abstraction as a tool to shape and revise existing knowledge and to make predictions about and organize new information. Abstract linguistic knowledge, such as lexical categories, need not be considered as mere output of acquisition, but also as stepping stones that support the formation of novel hypotheses, and, critically, help a learner *avoid* certain inferences that might otherwise lead them astray.

I.2 The Noun Bias

An important debate in the language acquisition literature which our work relates to is the question of why nouns are among the most early acquired category of words. For instance, nouns pre-dominate in children's early lexicon, and are much more frequently produced than other word classes (Dromi, 1999; Gentner, 1978; Jackson-Maldonado et al., 1993; Nelson et al., 1993). Further, children make fewer errors related to nominal compared to adjectival meanings, both in production and comprehension (Carey, 1982; Huttenlocher & Smiley, 1987; Maratsos, 1988; Naigles & Gelman, 1995). Similarly, children as old as 4 years are much less likely to correctly extend a novel word when presented in an adjectival frame (e.g., *this is a daxy one*) than when it is presented in a nominal frame (e.g., *this is a dax*) (Au & Laframboise, 1990; Imai & Gentner, 1997; L. B. Smith et al., 1992). Overall, experiments in which novel adjectives are taught to children often produce less successful results and are often more dependent on context than experiments in which novel nouns are taught. To account for these and many other related findings, collectively pointing to the so-called noun bias,

many researchers have argued that children have built-in expectations about how to map words onto their referents. Most prominently, Gentner (1982) proposed that children's perceptual systems are biased to pick out whole objects as opposed to parts or properties which are less readily individuated. If children are endowed with systems that bias their perception towards whole objects, this would explain their greater success in learning novel nouns relative to other words that pick out properties or actions.

A different line of work, more concerned with age-related biases present in children's learning environments than those already present at birth, has demonstrated that nouns themselves are privileged. In speech to children, nouns are more likely to occur in utterance-final positions (Freudenthal et al., 2013), in contexts with more high-frequency words (Goldfield, 1993) relative to other classes, and in contexts that contain more familiar words (Cameron-Faulkner et al., 2003). More strikingly, Aslin et al. (1996) have shown that when speaking to 12-month-olds, caregivers are twice as likely to violate grammaticality when placing a noun in utterance-final position compared to talking to adults. This is strong evidence that caregivers go to great lengths to make nouns more salient when speaking to their children. Some researchers interpret such findings as a potential explanation why children find it easier, initially, to learn nouns. For example, Cameron-Faulkner et al. (2003) pointed out that the fact that English-speaking children hear nouns within syntactic contexts that are much more consistent than those in which they hear verbs, challenges competing accounts of the noun bias based on cognitive factors. Another example is the connectionist model of Gasser and Smith (1998) which - without built-in perceptual biases - learned categories resembling nouns faster than it learned categories resembling other word classes. Our findings further support such claims by highlighting the privileged status of nouns in the statistical structure of speech to children. Because the noun category is less fragmented in speech to younger children, and nouns are more likely to occur in entropy-maximizing contexts that make individual nouns difficult to predict, nouns may be easier, at first, to discover and abstract from their distributional contexts. Preliminary support for this claim can be found in the work by Freudenthal et al. (2016) who showed that a distributional learning system sensitive to age-related changes in the lexical contexts of nouns was able to simulate the noun bias in children's word learning in the absence of built-in perceptual or conceptual biases.

II. Limitations and Future Directions

Our work has limitations that could be followed up with further research. For instance, we have examined a relatively restricted set of lexical contexts. While we have examined both forward and backward co-occurrences, our analysis of context 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). We have not included non-adjacent context words because this opens up many questions about the best way to represent those contexts. For instance, should word order be preserved? Should the representation context words be adjusted in proportion to their distance to the target word? Variations along these dimension can make a large difference in the ability of distributional semantic models to capture performance on a variety of semantic tasks (Bullinaria & Levy, 2007). Analyzing co-occurrences spanning greater distances would likely make analyses of fragmentation more sensitive. 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 like to occur at greater distances. This idea, combined with the fact that sentence length and syntactic complexity, factors that tend to contribute to fragmentation, are more frequent in speech to older children (Foushee et al., 2016; Hayes & Ahrens, 1988;

Kirchhoff & Schimmel, 2005), suggests that our analysis is underestimating the extent to which speech to younger children helps protect against noun category fragmentation. Nonetheless, our results - using only limited context information - provide an important existence proof which we hope will motivate future work on this topic.

However, our exclusion of context words at greater distances has an additional, more subtle, consequence: 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 (1.) and (2.) below, which both provide the diagnostic signal *some X* for the mass nouns *juice* and *coffee*.

1. *Can I have some juice please?*

2. *I want some more of that coffee!*

The difference is that the diagnostic signal in (2.) is not adjacent to the target word as in (1.). 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.

A second question for further research concerns the relationship between fragmentation and learnability. While the output of our fragmentation analysis is highly interpretable (partially due to being bounded between 0 and 1), it is not necessarily the case that higher fragmentation conclusively implies that a particular dataset is more difficult to learn from. In the contrary, higher fragmentation could indicate a rich set of hierarchical relationships that do not necessarily obscure or eliminate the discover-ability of the super-ordinate category. In practice it is possible that the superordinate category is perfectly signaled by the presence of one or a small set of context words, and all other context words simply enrich the superordinate category by signaling finer grained distinctions. Put differently, fragmentation does not necessarily imply that the statistical cues that underlie sub-ordinate categories occur *at the expense of* superordinate category cues - different sets of cues can coexist without necessarily coming into conflict with each another. The precise relationship between learnability and fragmentation would require specifying formal learning models and how they would handle these intermixed superordinate and subordinate category cues.

This leads to a third potential avenue for future research. While many of our corpus analyses predict that distributional learning systems would benefit by being trained in the order in which children actually experience language, in this chapter we did not actually perform any simulations that demonstrate a benefit for learning from less fragmented data first. The idea that learning from input ordered by increasing complexity may facilitate learning of lexical categories was first explored by J. L. Elman (1993), who found that syntax acquisition in an artificial grammar was only possible when training a recurrent network on input that "starts small" in terms of the syntactic complexity of sentences in the input. However, it is not yet known whether this finding can be replicated using larger, and more naturalistic language corpora. We are currently testing this prediction using computational modeling. Additionally, not every learning algorithm is likely to benefit from input that is scaffolded in the way we have demonstrated. For instance, batch learning (non-incremental) algorithms can be excluded on first principles, but what about more subtle differences between incremental learning systems, such as whether distributed versus localist representations, or supervised versus. unsupervised systems are used?

Another potential research direction is to better characterize the source of fragmentation. Does fragmentation in speech to older children

originate in linguistically relevant distinctions, or is fragmentation simply the inevitable result of the messiness of natural language - a combination of the large vocabulary and the virtually infinite potential for combining words? In other words, does the increase in fragmentation over time actually reveal important syntactic or semantic distinctions, or is fragmentation the result of lexically-specific idiosyncrasies irrelevant to theories of semantic or syntactic development? Similarly, what does the increase in syntactic complexity in speech to children have to do with fragmentation, if anything? Does the age-related increase in syntactic complexity drive fragmentation, or are the two independent?

Our corpus analysis were purposefully restricted to nouns in child directed speech because they are among the earliest learned words and therefore the most likely candidate for age-related fragmentation. However, our analyses of the change in conditional entropy of non-nouns in child directed speech suggest that other word classes might be similarly influenced by fragmentation. Toward that end, another future direction is a more systematic survey of fragmentation that includes a broader set of word classes. Such work has the potential to reveal whether age-related increase in fragmentation is a noun-specific phenomenon, and whether there are principled differences in fragmentation between, for instance, function words, and content words.

Because our methodology for evaluating fragmentation is compatible with any co-occurrence data, not just lexical co-occurrences, it might be useful to apply our method for quantifying fragmentation in other domains in psychology (e.g. learning visual concepts). Further, our method could be extended or employed in novel ways, such as for identifying subsets of data that are particularly fragmented (e.g. in a data pre-processing step). To make our method more appropriate for analyzing co-occurrence data which is discrete rather than continuous, it would be useful to improve our method with a matrix decomposition methodology other than standard singular value decomposition, which respects the discrete nature of co-occurrence data (Buntine & Jakulin, 2004).

Further, while we have developed and empirically tested a hypothesis for explaining the qualitative shift in the co-occurrence data in speech to children, our explanation makes little contact with questions about the psychology and ecology of caregivers who are actually producing the speech we have been examining. Questions about the constraints on producers, and the child-caregiver dyad are interesting directions to pursue in the future. For example, are caregivers purposefully avoiding certain constructions when speaking to children, and do they adjust the complexity of their speech in accordance with the linguistic competence of their child?

VIII. Conclusion

While modeling efforts have greatly advanced our understanding of the possible mechanisms underlying learning from lexical context, most work has disregarded the role that the slowly changing language environment of children plays in organizing and scaffolding language acquisition. With rare exceptions (J. L. Elman, 1993; Freudenthal et al., 2016; Huebner & Willits, 2018), most distributional models of lexical category acquisition are trained without regard to the dramatic changes in language complexity that accompany the period between early infancy and school-age. In this chapter, we demonstrated the non-stationary aspect of child-directed speech in terms of age-related increase in fragmentation of nouns, and developed a theoretical framework for understanding the qualitative shift in lexical distributions that underlie age-related fragmentation. Our results have broad implications for developmentally-plausible training of models of language learning. For example, providing language input to incremental learning systems in the order in which children actually experience it, should endow learners with strong inductive biases for navigating the complex co-occurrence signals in natural language data. As Aristotle said, "Well begun is half done."

References

Alishahi, A., & Chrupala, G. (2012). Concurrent acquisition of word meaning and lexical categories. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 643–654.

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Lawrence Erlbaum Associates.

Au, T. K.-f., & Laframboise, D. E. (1990). Acquiring color names via linguistic contrast: The influence of contrasting terms. *Child Development*, *61*(6), 1808–1823.

Bertenthal, B. I., Rose, J. L., & Bai, D. L. (1997). Perception–action coupling in the development of visual control of posture. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(6), 1631.

Borovsky, A., & Elman, J. (2006). Language input and semantic categories: A relation between cognition and early word learning. *Journal of child language*, *33 4*, 759–90.

Bower, G. H. (1970). Organizational factors in memory. *Cognitive psychology*, *1*(1), 18–46.

Braine, M. D. (1963). On learning the grammatical order of words. *Psychological review*, *70*(4), 323.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–B44.

Broen, P. (1972). The verbal environment of the english-learning child. *ASHA Monographs*, *17*.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, *39*(3), 510–526.

Buntine, W. L., & Jakulin, A. (2004). Applying discrete pca in data analysis. *20th conference on uncertainty in artificial intelligence (uai)*.

Byrge, L., Sporns, O., & Smith, L. B. (2014). Developmental process emerges from extended brain–body–behavior networks. *Trends in cognitive sciences*, *18*(8), 395–403.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cogn. Sci.*, *27*(6), 843–873.

Carey, S. (1982). Semantic development: The state of art. In E. Wanner & G. L. R. (Eds.), *Language acquisition: The state of art*. Cambridge University Press.

Casey, B., Giedd, J. N., & Thomas, K. M. (2000). Structural and functional brain development and its relation to cognitive development. *Biological Psychology*, *54*(1), 241–257. https://doi.org/https://doi.org/10.1016/S0301-0511(00)00058-2

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.

Chomsky, N. (2002). *Syntactic structures*. Mouton de Gruyter.

Clark, E. V. (1973). What's in a word? on the child's acquisition of semantics in his first language. In T. E. Moore (Ed.), *Cognitive development and acquisition of language* (pp. 65–110). Academic Press.

Clark, E. V., Gelman, S. A., & Lane, N. M. (1985). Compound nouns and category structure in young children. *Child Development*, *56*(1), 84–94. http://www.jstor.org/stable/1130176

Cochran, B. P., McDonald, J. L., & Parault, S. J. (1999). Too smart for their own good: The disadvantage of a superior processing capacity for adult language learners. *J. Mem. Lang.*, *41*(1), 30–58.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, *8*(2), 240–247.

Cunningham, A. E., & Stanovich, K. E. (1998). What reading does for the mind. *American educator*, *22*, 8–17.

Dautriche, I., Swingley, D., & Christophe, A. (2015). Learning novel phonological neighbors: Syntactic category matters. *Cognition*, *143*, 77–86.

Dromi, E. (1999). Early lexical development. In M. D. Barrett (Ed.), *The development of language*. Psychology Press.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, *7*(2), 195–225.

Estes, K. G., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? statistical segmentation and word learning. *Psychological science*, *18*(3), 254–260.

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant behavior and development*, *10*(3), 279–293.

Fernald, A., & Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child development*, *83*(1), 203–222.

Firth, J. (1961). *Papers in linguistics, 1934-1951*. Oxford University Press.

Fisher, C., Gertner, Y., Scott, R. M., & Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(2), 143–149.

Foushee, R., Griffiths, T., & Srinivasan, M. (2016). Lexical complexity of child-directed and overheard speech: Implications for learning. *CogSci*.

Freudenthal, D., Pine, J., Jones, G., & Gobet, F. (2013). Frequent frames, flexible frames and the noun-verb asymmetry. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*(35).

Freudenthal, D., Pine, J. M., Jones, G., & Gobet, F. (2016). Developmentally plausible learning of word categories from distributional statistics. *CogSci*.

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.

Gallaway, C., & Richards, B. J. (1994). *Input and interaction in language acquisition*. Cambridge University Press.

Gasser, M., & Smith, L. B. (1998). Learning nouns and adjectives: A connectionist account. *Language and cognitive processes*, *13*(2-3), 269–306.

Gelman, S. A., & Taylor, M. (1984). How two-year-old children interpret proper and common names for unfamiliar objects. *Child development*, *55*(4), 1535–1540.

Gentner, D. (1978). On relational meaning: The acquisition of verb meaning. *Child development*, *49*(4), 988–998.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Center for the Study of Reading Technical Report; no. 257*.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language learning and development: the official journal of the Society for Language Development*, *1*(1), 23–64.

Goldfield, B. A. (1993). Noun bias in maternal speech to one-year-olds. *Journal of child language*, *20*(1), 85–99.

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.

Gomez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436.

Graven, S. N., & Browne, J. V. (2008). Visual development in the human fetus, infant, and young child. *Newborn and Infant Nursing Reviews*, *8*(4), 194–201.

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

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap by age 3. *American educator*, *27*(1), 4–9.

Hayes, D. P., & Ahrens, M. G. (1988). Vocabulary simplification for children: A special case of 'motherese'? *Journal of child language*, *15*(2), 395–410.

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.

Höhle, B., Weissenborn, J., Kiefer, D., Schulz, A., & Schmitz, M. (2004). Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy*, *5*(3), 341–353.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Lang. Learn. Dev.*, *1*(2), 151–195.

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*, 133.

Huttenlocher, J., & Smiley, P. (1987). Early word meanings: The case of object names. *Cognitive Psychology*, *19*(1), 63–89.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cogn. Psychol.*, *61*(4), 343–365.

Imai, M., & Gentner, D. (1997). A cross-linguistic study of early word meaning: Universal ontology and linguistic influence. *Cognition*, *62*(2), 169–200.

Jackson-Maldonado, D., Thal, D., Marchman, V., Bates, E., & Gutierrez-Clellen, V. (1993). Early lexical development in spanish-speaking infants and toddlers. *Journal of child language*, *20*(3), 523–549.

Jakulin, A., & Bratko, I. (2003). *Quantifying and visualizing attribute interactions* [Preprint on webpage at https://arxiv.org/abs/cs/0308002].

Jayaraman, S., Fausey, C. M., & Smith, L. B. (2017). Why are faces denser in the visual experiences of younger than older infants? *Developmental psychology*, *53*(1), 38.

Keil, F. C. (1981). Constraints on knowledge and cognitive development. *Psychological review*, *88*(3), 197.

Kemp, N., Lieven, E., & Tomasello, M. (2005). Young children's knowledge of the "determiner" and "adjective" categories. *J. Speech Lang. Hear. Res.*, *48*(3), 592–609.

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.

Kodner, J. (2018). Syntactic category learning as iterative prototype-driven clustering. *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, 44–54. https://doi.org/10.7275/R5TQ5ZQ4

Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see the world differently. *Child development*, *85*(4), 1503–1518.

Lany, J., & Saffran, J. R. (2010). From statistics to meaning: Infants' acquisition of lexical categories. *Psychological science*, *21*(2), 284–291.

Lee, L. S.-Y. (2015). *On the linear algebraic structure of distributed word representations* [Preprint on webpage at https://arxiv.org/abs/1511.06961].

Lieven, E. V. (1994). Crosslinguistic and crosscultural aspects of language addressed to children.

MacWhinney, B. (2014). *The childes project: Tools for analyzing talk, volume ii: The database*. Psychology Press.

Maratsos, M. (1988). Crosslinguistic analysis, universals, and language acquisition. *The development of language and language researchers: Essays in honor of Roger Brown*, 121–152.

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.

Mervis, C. B. (1983). Acquisition of a lexicon. *Contemporary educational psychology*, *8*(3), 210–236.

Mintz, T. H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition*, *30*(5), 678–686.

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.

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science*, *26*(9), 1489–1496.

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8-and 12-year-old children and adults. *Journal of Experimental Psychology: General*, *144*(2), 447.

Naigles, L. G., & Gelman, S. A. (1995). Overextensions in comprehension and production revisited: Preferential-looking in a study of dog, cat, and cow. *Journal of Child Language*, *22*(1), 19–46.

Nelson, K., Hampson, J., & Shaw, L. K. (1993). Nouns in early lexicons: Evidence, explanations and implications. *Journal of Child Language*, *20*(1), 61–84.

Newport, E., Gleitman, H., & Gleitman, L. (1977). Mother, id rather do it myself: Some effects and non-effects of maternal speech style. In E. C. 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.

Phillips, J. R. (1973). Syntax and vocabulary of mothers' speech to young children: Age and sex comparisons. *Child development*, *44*(1), 182–185.

Pine, J. M. (1994). The language of primary caregivers. In C. Gallaway & B. J. Richards (Eds.), *Input and interaction in language acquisition*. Cambridge university press.

Pinker, S. (2009). *Language learnability and language development, with new commentary by the author: With new commentary by the author* (Vol. 7). Harvard University Press.

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

Redington, M., & Chater, N. (1998). Connectionist and statistical approaches to language acquisition: A distributional perspective. *Language and cognitive processes*, *13*(2-3), 129–191.

Richards, B. J. (1994). Child-directed speech and influences on language acquisition: Methodology and interpretation. In C. Gallaway & B. J. Richards (Eds.), *Input and interaction in language acquisition*. Cambridge University Press.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of verbal learning and verbal behavior*, *14*(6), 665–681.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Rogers, T. T., & McClelland, J. L. (2008). Précis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences*, *31*(6), 689.

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.

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.

Schieffelin, B. B., & Ochs, E. (1986). *Language socialization across cultures*. Cambridge University Press.

Schwering, S. C., Ghaffari-Nikou, N. M., Zhao, F., Niedenthal, P. M., & MacDonald, M. C. (2021). Exploring the relationship between fiction reading and emotion recognition. *Affective Science*, *2*(2), 1–9.

Seebach, B. S., Intrator, N., Lieberman, P., & Cooper, L. N. (1995). A model of prenatal acquisition of speech parameters. *How we learn; how we remember: Toward an understanding of brain and neural systems: Selected papers of leon n cooper* (pp. 364–367). World Scientific.

Seidenberg, M. S., & MacDonald, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cogn. Sci.*, *23*(4), 569–588.

Shi, R., & Melançon, A. (2010). Syntactic categorization in french-learning infants. *Infancy*, *15*(5), 517–533.

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711). https://doi.org/10.1098/rstb.2016.0051

Smith, L. B., Jones, S. S., & Landau, B. (1992). Count nouns, adjectives, and perceptual properties in children's novel word interpretations. *Developmental Psychology*, *28*(2), 273.

Snow, C., & Ferguson, C. (1977). Talking to children: Language input and acquisition: Language input and acquisition.

Swingley, D., & Humphrey, C. (2018). Quantitative linguistic predictors of infants' learning of specific english words. *Child development*, *89*(4), 1247–1267.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge University Press.

Tomasello, M. (2001). First steps toward a usage-based theory of language acquisition. *Cogn. Linguist.*, *11*(1-2).

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res.*, *37*, 141–188.

Valian, V. (1986). Syntactic categories in the speech of young children. *Dev. Psychol.*, *22*(4), 562–579.

Warrington, E. K. (1975). The selective impairment of semantic memory. *The Quarterly journal of experimental psychology*, *27*(4), 635–657.

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.

Yuan, S., Fisher, C., & Snedeker, J. (2012). Counting the nouns: Simple structural cues to verb meaning. *Child development*, *83*(4), 1382–1399.