Developmental Change in the Relationship Between Lexical and Grammatical Acquisition

Mika Braginsky
mikabr@stanford.edu
Department of Psychology
Stanford University

Daniel Yurovsky
yurovsky@stanford.edu
Department of Psychology
Stanford University

Virginia Marchman
marchman@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

TODO: Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Keywords: TODO

Introduction

Does abstract structure in language emerge from the interaction of individual words, or are syntactic structures represented separately? On lexicalist theories of grammatical development, syntactic structure emerges from graded generalizations on the basis of lexical items and there may be little or no representation of syntactic rules or regularities per se, at least early in development (Tomasello, 2001, 2003). Even if syntactic structures are eventually represented, these representations should be directly related to their support in more concrete lexical structure (Bannard, Lieven, & Tomasello, 2009; ?, ?).

In contrast, on more nativist theories like principles and parameters, grammar is predicted to emerge independently from lexical knowledge and on its own (largely maturational) timetable. On these theories, older children should have more syntactic competence, largely independent from the amount of language input they receive and hence from the size of their vocabulary.

Developmental data can help resolve this conflict by providing estimates of the relationship between grammar, lexicon, and age. Children's vocabulary growth is known to be strongly predictive of their grammatical competence (Bates & Goodman, 1997; C. Caselli, Casadio, & Bates, 1999). However, competing theoretical accounts make different predictions about the relationship between vocabulary and grammatical acquisition across development. On a more lexicalist account, the lexicon provides the basis for grammatical generalizations, so the relationship between the lexicon and grammar should not change with age. On the other hand, age-based changes in this relationship suggest the presence of developmental processes that regulate grammatical acquisition above and beyond lexical acquisition.

We probe the lexicon-grammar relationship further by delineating grammatical development into more targeted measures that 1) capture distinctions between morphological and syntactic knowledge; and 2) characterize the composition of the lexicon by grammatical function. For each of these, we examine the relationship between the measure and the size of the lexicon, and determine how age affects the relationship. A measure's relationship not changing with age indicates that vocabulary size predicts it thoroughly, while age-related change is indicative of the presence of some developmental shift not captured by vocabulary size.

Our measures of both lexical and grammatical development, like those of Bates et al., are derived from the MacArthur-Bates Communicative Development Inventories (CDIs; Fenson et al., 1994, 2007), a widely-used family of parent-report instruments for easy and inexpensive data-gathering about early language acquisition. We use data from adaptations of the CDI into four languages: English, Spanish, Norwegian, and Danish to establish cross-linguistic generalizability to our findings.

Each language's adaptation includes a vocabulary checklist, a word form section consisting of morphological inflections, and a complexity section consisting of pairs of syntactically simple/complex sentences. Vocabulary size as measured by the CDI correlates with both laboratory measurements and naturalistic observation (see (Fenson et al., 2007) for a review) and score on the complexity section correlates with MLU and other measurements of grammatical development [TODO: citation needed] [TODO: justification for word form][TODO: justifications for other languages].

[TODO: move some of the previous work summaries from the section sub-intros here? not sure how to structure this optimally or how to transition from into to paper]

General Methods

TODO: some text here

CDI Form Database

We implemented Wordbank (`wordbank.stanford.edu`), a structured database of CDI data, to aggregate and archive CDI data across languages and labs. By collecting language development data at an unprecedented scale, Wordbank enables the exploration of novel hypotheses about the course of lexical and grammatical development. [TODO: more about wordbank? what's useful?]

Wordbank now includes data in four languages: English, Spanish, Norwegian (?, ?, ?), and Danish (?, ?), including both cross-sectional and longitudinal data. [TODO: more about how/where data came from, not sure how to describe it]

CDI Measures

In general, CDI forms contain both vocabulary checklists and other questions relevant to the child's linguistic development. All of the data used here is from the Words and Sentences (Toddler) CDI instruments of each language, administered to children ages 16 months to 32 months. Each of these instruments includes a vocabulary section, which asks whether the child produces each of around 700 words form a variety of semantic and syntactic categories; a word form section, which asks whether the child produces each of around 30 morphologically inflected forms of nouns and verbs; and a complexity section, which asks whether the child's speech is most similar to the syntactically simpler or more complex versions of around 40 sentences. Each language's instrument is not merely a translation of the English words and sentences, but rather constructed and normed specifically to reflect the lexicon and grammar of that language. Table 1 shows, for each language, the number of items in each of these categories [TODO: also include examples of each?].

To analyze lexical and grammatical development, we derive several quantities based on these measures. Each child's vocabulary size is the proportion of the words on the corresponding CDI form that the child is reported to produce. Similarly, each child's word form score is the proportion of word forms they are reported to produce, and their complexity score the proportion of complexity items for which they are reported to use the more complex form. We compute all of these quantities as proportions rather than numbers of items to make the scales comparable across languages.

	Vocabulary	Word Form	Complexity
English	680	25	37
Spanish	680	24	37
Norwegian	731	33	42
Danish	725	29	33

Table 1: Overview of instruments in each language: number of items in each relevant section.

Syntax and Morphology

By the age of two years, most children have a sizable working vocabulary, including verbs, prepositions and closed class forms that perform grammatical work functions. They are also beginning to use two- or three-word combinations (e.g., "mommy sock") and may demonstrate productive use of inflectional morphemes (e.g., past tense "-ed"). As with vocabulary, there is sizable variation in when and how children move into grammar, and children who are more advanced in early vocabulary are more advanced in grammatical development as well (Bates & Goodman, 1999). This suggests that the mechanisms guiding vocabulary and grammar learning are highly interdependent (Tomasello, 2003; Bresnan, 2001), a view at odds with the nativist assumption that grammar emerges independent of the lexicon (Chomsky, 1981).

Previous studies have found a strong connection between the size of the lexicon and grammatical development as measured by the complexity section of the CDI. A consistent nonlinear relationship appears across a variety of languages, including English (Bates et al., 1994; Fenson et al., 1994), Italian (C. Caselli et al., 1999), Hebrew (Maital, Dromi, Sagi, & Bornstein, 2000), and Spanish (Jackson-Maldonado, 2003). However, no studies have had the power and cross-linguistic representation to go beyond this initial finding. We extend it by examining grammatical development using two different measures: word form as a window into morphology and complexity as a window into syntax. For each measure, we investigate the interaction of vocabulary size and age in a variety of languages.

Analysis

For each language in our sample, we wanted to estimate how much of a child's syntactic and morphological development was left to predicted after knowing that child's vocabulary size. Specifically, we asked whether knowing a child's age provided additional predictive power over and above vocabulary size. To estimate this predictive power of age, we fit regression models to each child's word-form and complexity scores. We modeled each child's score as a quadratic function of vocabulary size and interaction with the child's age in months. Figure 1 shows a scatter plot of these data and models. Each dot represents an individual child's score on each measure, while curves show the relationship between that measure and vocabulary size. As seen most clearly for English and Norwegian, the curves for complexity show a characteristic fan, while the curves for word form do not, suggesting that the relationship between vocabulary size and complexity score is modulated by age, while the relationship between vocabulary size and word form score is not (or is to a lesser extent). The Spanish and Danish data show less of a clear complexity curve fan, possibly because of the relatively small number of data points in the youngest age bin.

Because of the size of our samples, all main effects and interactions were highly significant in each language. To test our prediction—that vocabulary development predicts more

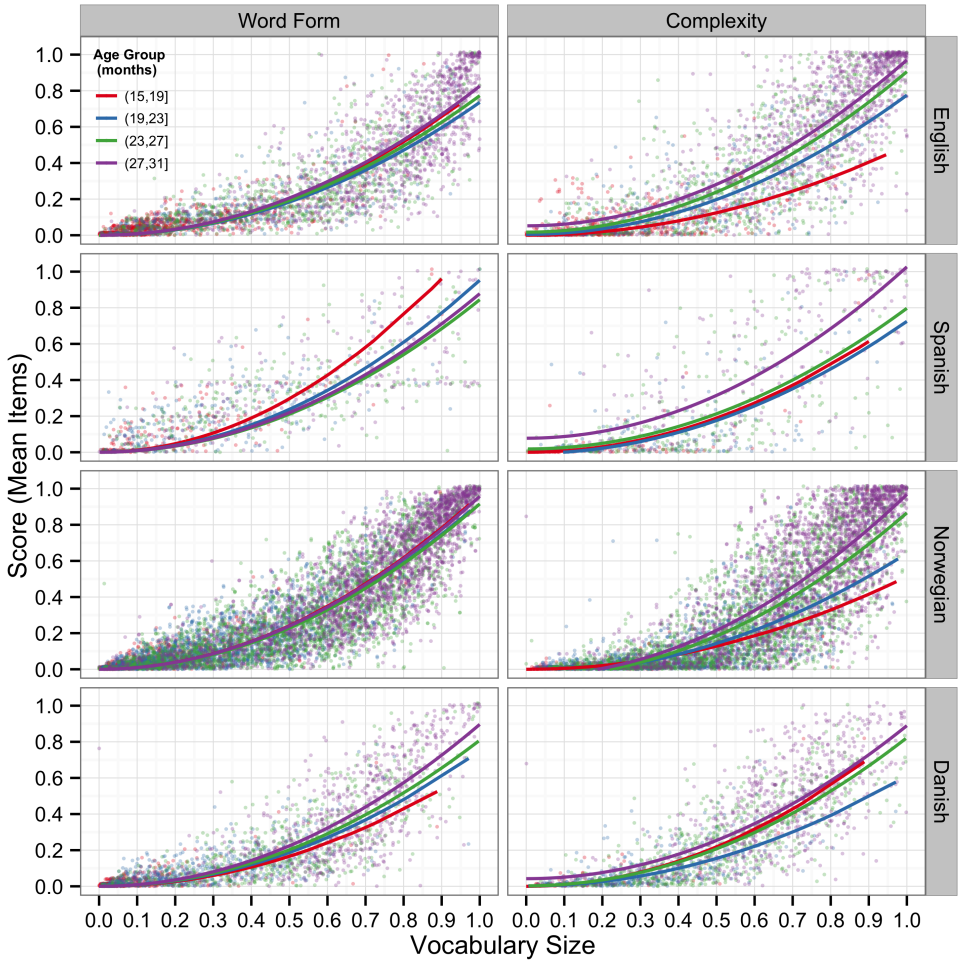


Figure 1: Each point shows an individual child, indicating their total vocabulary size and word form score or complexity score, with color showing their age bin (English $n = 4137$; Spanish $n = 1094$; Norwegian $n = 8505$; Danish $n = 2074$). Panels show different languages, and curves are regression models fit separately for each language and measure. The models were specified as `score` $\sim$ `vocab`^{2} `* age`.

of the variability in children's morphological than syntactic development—we compared the coefficients in our models across languages and measures. Figure 2 shows the interaction between age and vocabulary size for each of these models across languages. In each model, the age-related interaction coefficient is substantially larger for complexity than for word form, indicating that for complexity more than word form. TODO: report some metrics on the models?

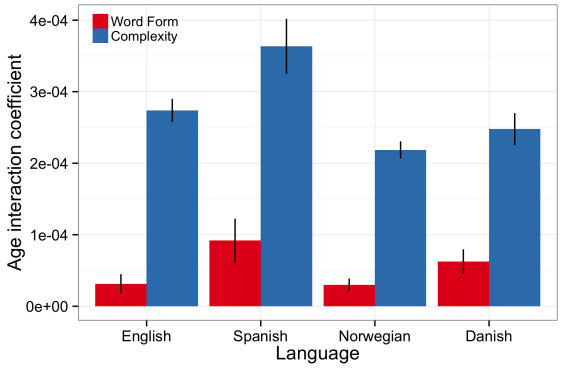
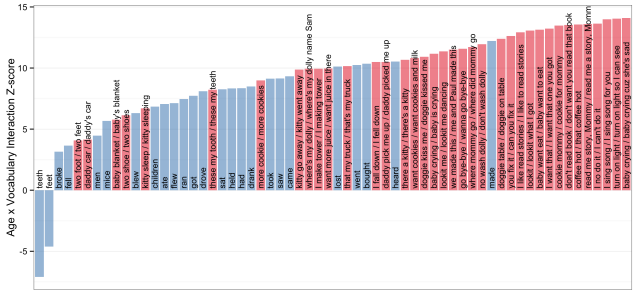


Figure 2: For each language and measure (word form and complexity), the coefficient of the interaction between vocabulary size and age. Error bars indicate the $\pm$1SE (TODO: is that true?). Across languages, complexity has a substantially larger interaction with age effect than word form, suggesting that less of children's syntactic development is predicted by their vocabulary growth.

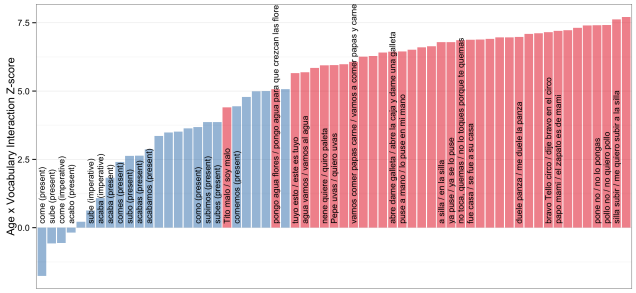
Given the heterogeneous nature of the CDI instruments, particularly of the complexity sections, we further break down the items of the complexity sections by classifying them as capturing more morphological or more syntactic phenomena. [TODO: stuff about our coding methods here]. We then fit predictive models as above, but separately for each word form and complexity item. Figure 3 shows the z-score ([TODO: why coefficients some places and z-scores other places?]) of the age-interaction terms of the models for each item. [TODO: code the items and analyze these results post-coding].

Discussion

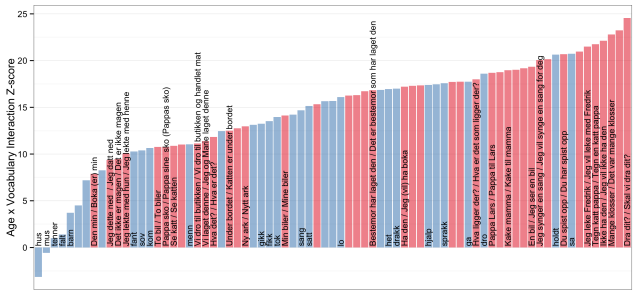
We build on previous analyses that showed a strong relationship between lexical and grammatical development, by factoring in the interplay of age with this relationship. We further distinguished in this analysis between measures more reflective of morphology and measures more reflective of syntax, and found that syntactic development broadly shows greater age modulation than morphological development. Thus, this analysis provides circumstantial evidence for the relationship between syntactic development and age, independent of the growth of the lexicon.



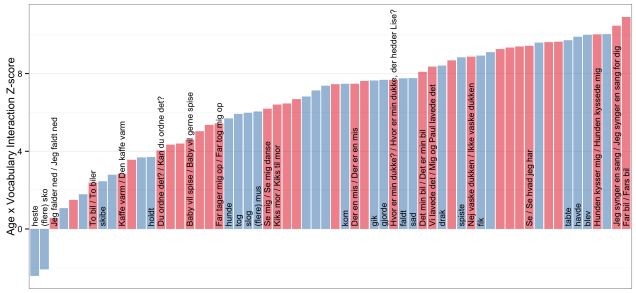
(a) English



(b) Spanish



(c) Norwegian



(d) Danish

Figure 3: For each language and item, the z-score of the model's age and vocabulary size interaction term. Across languages, complexity items tend to have a substantially larger age effect than word form items. TODO: get non-ascii characters to display; make things big enough to be legible; use coefficients instead of z-scores for consistency?

Vocabulary Composition

Early vocabulary development is typically characterized by learning of names for caregivers and common objects, while later in development, children tend to increase their vocabulary by increasing the proportion of predicates (verbs and adjectives). This over-representation of nouns has been found across a number of analyses and in a variety of languages (Bates et al., 1994)(TODO: others?), though not all (M. Caselli et al., 1995). [1] For our purposes we are interested in using these analyses of vocabulary composition to test for the same kind of age-related differences that we found in the complexity and word-form analyses.

We predict that the proportion of verbs in children's vocabulary should be relatively more affected by age than nouns. Concrete nouns are hypothesized to be learned initially from both co-occurrences between words (Yu & Smith, 2007) and by social cues to reference to particular objects (Bloom, 2002). On neither of these accounts should syntactic information be a primary information source (though of course syntax might be more informative for abstract nouns). In contrast, for verbs, syntax has been argued to be crucial for learning. On the syntactic bootstrapping hypothesis (Gleitman, 1990; ?, ?), verbs are learned by mapping the syntactic structure of utterances to the thematic structure of observed events, for example by noticing that the subject of a sentence matches the agent in one particular ongoing event but not another ("the cat is fleeing the dog" matches FLEES(CAT, DOG) but not CHASES(DOG,CAT)). Thus, if syntactic development is related in some way to age, we should see larger age effects on verb representation than noun representation.

Analysis

Each CDI form contains a mixture of words in different classes. We adopt the categorization of Bates et al. (1994), who split words in nouns, predicates (adjectives, adverbs, and verbs), function words, and other words. Then for each child's vocabulary, we compute the proportion of the total words in each of these categories that they are reported to produce. This yields a set of proportions for each child.

For each of the four languages in our sample, we plot these proportions against total vocabulary. These functions are shown in Figure 4: each dot represents an individual child's knowledge of a particular class, while curves show the relationship between that class and the entirety of the vocabulary. If categories grow independently of one another, these curves should approximate the diagonal. This pattern is not what either we or Bates et al. observe however: Across the languages in our sample, nouns are systematically over-represented in smaller vocabularies (shown by a curve that is above the diagonal), while function words—and to some extent, predicates—are under-represented.

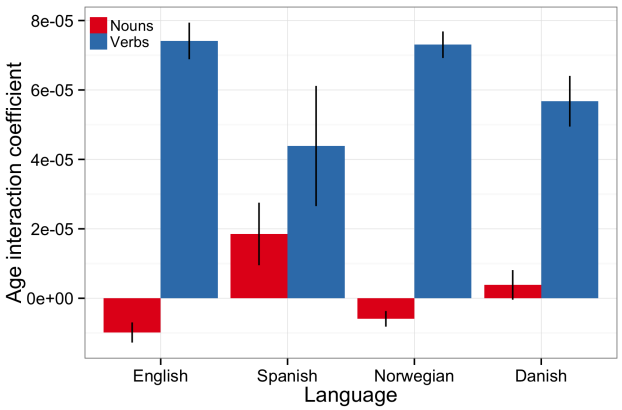


Figure 5: For each language and part of speech (nouns and verbs), the coefficient of the model's age and vocabulary size interaction term. Across languages, verbs have a substantially larger age effect than nouns. TODO: is this plot broken/wrong?

Next, we measure the contribution of age to vocabulary composition. We fit a linear model to all children's data for each word class, predicting word-class proportion as a linear and quadratic (TODO) function of total vocabulary. We then investigated the interaction between total vocabulary and age. Because of our theoretical interest in the relationship between age and syntactic development, we focus here specifically on nouns and verbs. Figure 5 shows coefficients for each of these models across languages. In each model, the age-related interaction coefficient is substantially larger for verbs than for nouns. This asymmetry can be interpreted as evidence that for two vocabulary-matched children, the older would tend to have relatively more verbs than the younger, and this effect was larger for children with overall larger vocabularies. TODO: report some metrics on the models?

Discussion

We replicated previous analyses showing an over-representation of nouns in the developing lexicon and a relative under-representation of verbs. We also predicted that—if syntactic generalization was in some way tied to age—verbs would show relatively more influence than nouns. This prediction was confirmed across all four languages we examined. Thus, this analysis provides additional circumstantial evidence for a relationship between syntactic development and age, independent of the growth of the lexicon.

[1] Differences in early vocabulary composition have been argued to emerge from typological differences (e.g., word order, subject drop), and from cultural practices (e.g., focus on picture book reading) (Tardif, Gelman, & Xu, 1999; Gopnik, Choi, & Baumberger, 1996; Choi & Gopnik, 1995)—we are agnostic as to the source of this variability.

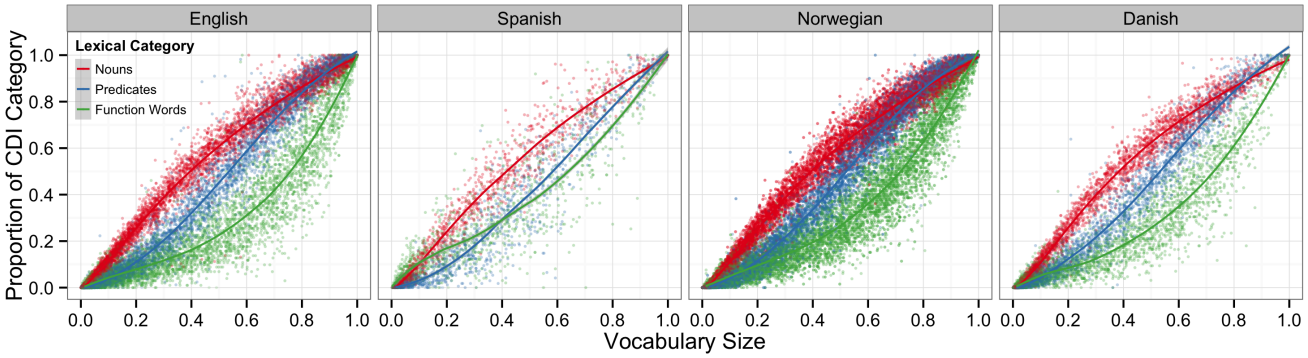


Figure 4: Proportion of a particular CDI category, plotted by total vocabulary size. Each point shows an individual child, with color showing their noun, predicate, and function word vocabulary. Panels show different languages, and curves are smoothing functions using loess. (English $n = 5595$; Spanish $n = 1094$; Norwegian $n = 10095$; Danish $n = 3038$) TODO: have fit model curves instead of smoothers?

General Discussion

We measured children's grammatical competence using different approaches: their reported usage of various morphological forms and syntactic constructions, and the substructure of their vocabularies by grammatical category. For each of these metrics, we used vocabulary size as a predictor and examined the interaction of age with this predictive relationship.

Across four languages, we find that syntax is modulated by age to a greater extent than morphology, and that verb proportion is modulated by age to a greater extent than noun proportion. Both of these findings suggest a place for developmental processes that facilitate grammatical acquisition, above vocabulary acquisition. This developmental change could range from something more domain-general like working memory to something more domain-specific like [TODO: how does P&P formulate this?]. In either case, it goes beyond a purely lexicalist account of grammatical acquisition.

Acknowledgments

Thanks to the MacArthur CDI Advisory Board, Dorthe Bleses, Kristian Kristoffersen, Rune Nørgaard Jørgensen, and the members of the Language and Cognition Lab.

References

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*(41), 17284.

Bates, E., & Goodman, J. (1999). On the emergence of grammar from the lexicon. In B. Macwhinney (Ed.), *The emergence of language.* Mahwah, NJ: Lawrence Erlbaum Associates.

Bates, E., & Goodman, J. C. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real-time processing. *Language and cognitive Processes*, *12*(5-6), 507–584.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., ... Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of child language*, *21*(01), 85–123.

Bloom, P. (2002). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Bresnan, J. (2001). *Lexical-functional syntax*. Wiley-Blackwell.

Caselli, C., Casadio, P., & Bates, E. (1999). A comparison of the transition from first words to grammar in english and italian. *Journal of child language*, *26*(01), 69–111.

Caselli, M., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J. (1995). A cross-linguistic study of early lexical development. *Cognitive Development*, *10*(2), 159–199.

Choi, S., & Gopnik, A. (1995). Early acquisition of verbs in korean: A cross-linguistic study. *Journal of child language*, *22*(03), 497–529.

Chomsky, N. (1981). Principles and parameters in syntactic theory. *Explanation in linguistics: The logical problem of language acquisition*, 32–75.

Fenson, L., Bates, E., Dale, P. S., Marchman, V. A., Reznick, J. S., & Thal, D. J. (2007). *Macarthur-bates communicative development inventories*.

Fenson, L., Dale, P., Reznick, J., Bates, E., Thal, D., Pethick, S., ... Stiles, J. (1994). Variability in early communicative development. *Monographs of the society for research in child development*, *59*(5).

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 3–55.

Gopnik, A., Choi, S., & Baumberger, T. (1996). Cross-linguistic differences in early semantic and cognitive development. *Cognitive Development*, *11*(2), 197–225.

Jackson-Maldonado, D. (2003). *Macarthur inventarios del desarrollo de habilidades comunicativas: User's guide and technical manual*. Paul H Brookes Pub Co.

Maital, S. L., Dromi, E., Sagi, A., & Bornstein, M. H. (2000). The hebrew communicative development inventory: Language specific properties and cross-linguistic generalizations. *Journal of Child Language*, *27*(01), 43–67.

Tardif, T., Gelman, S. A., & Xu, F. (1999). Putting the noun

bias in context: A comparison of english and mandarin. *Child Development*, *70*(3), 620–635.

Tomasello, M. (2001). Perceiving intentions and learning words in the second year of life. In *Language acquisition and conceptual development* (pp. 132–159). New York: Cambridge University Press.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

TODO: fix broken citations