Developmental Changes in the Relationship Between Grammar and the Lexicon

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

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

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

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

Abstract

How does abstract structure emerge during language learning? On some accounts, children's early syntax consists of direct generalizations from particular lexical items, while on others, children's syntactic structure emerges independently and follows its own timetable. Progress on differentiating these views requires detailed developmental data. Using parent reports of vocabulary and grammar abilities, previous analyses have shown that early syntactic abstraction depends on the growth of the lexicon, providing support for lexicalist theories. Here we replicate and extend these findings, showing that they hold across four languages in a very large sample of children. We also show that there are measurable effects of age over and above effects of lexical development, and that these effects are greater for aspects of language ability that are more closely tied to syntax. These findings provide evidence for non-lexical contributions to the growth of syntactic abstraction, whether from domain-general or language-specific mechanisms.

Keywords: Language acquisition; word learning; morphology; syntax; development.

Introduction

A child as young as two or three years old (who happens to be acquiring English) can hear someone say *Alice glipped the blicket* and draw a wealth of inferences from the morphological and syntactic structure of that utterance: that *Alice* and *blicket* are entities in the world and *glipping* is an action; that Alice is the one doing the glipping and the blicket is the thing being glipped; that glipping occurred in the past (rather than ongoing in the present, as in *Alice is glipping the blicket*); that a singular blicket was involved (rather than multiple, as in *Alice glipped the blickets*). What mechanisms underlie the formation of generalizations that support such inferences? Does abstract structure in language emerge from the interactions of individual words, or is it acquired and represented separately?

According to lexicalist theories, morphosyntactic structure emerges from graded generalizations on the basis of lexical items, and at least early in development, there may be little or no representation of morphosyntactic rules or regularities *per se* (Tomasello, 2003). Even if syntactic structures are eventually represented, 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 (Chomsky, 1981; Baker, 2005), grammar emerges independently from lexical knowledge following its own, largely maturational, timetable. According to these theories, older children should have more grammatical competence than younger children, and this growth in competence should be driven by factors that are independent of the amount of language input they receive and so of the size of their vocabulary.

Developmental data that explore relations between lexicon, grammar and age are critical to resolving issues related to this fundamental theoretical debate. Early efforts used data from the MacArthur-Bates Communicative Development Inventory (CDI), a widely-used assessment tool in which parents report which words their child produces on a checklist organized by lexical-semantic categories. Children's vocabulary size can thus be estimated over the entire checklist, or for sub-categories. The CDI also provides indices of grammar learning by asking about children's use of inflected forms (e.g., *walked*) and the complexity of their word combinations (e.g., *kitty sleeping / kitty is sleeping*). Influential initial findings showed that early vocabularies tend to be composed primarily of nouns, while verbs and closed-class forms, that might support the transition into complex sentences, are typically acquired later (Bates et al., 1994). Further, across different populations and languages, global estimates of grammatical development were more strongly predicted by overall vocabulary size than by age, providing support for lexicalist theories (see Bates & Goodman, 1999 for a review).

While impressive in their time, the scope and power of these early studies were nevertheless limited, relying on relatively small norming samples (1000–2000 children) with few opportunities for direct comparisons of the nature or extent of these relations across languages (cf. English vs. Italian). The current study addresses these limitations using data from Wordbank (`wordbank.stanford.edu`), a new web-based tool that aggregates pre-existing samples of CDI data into a consistent format across forms and languages. While still in development, the resulting database is already considerably larger than those previously available, and thus allows analyses of lexical-grammar-age relations with enhanced statistical power and broader cross-linguistic representation. Here, we present data from 19,822 children between 16 and 32 months old, using parallel adaptations of the CDI Words & Sentences form in four languages: English, Spanish, Norwegian, and Danish.

In this study, we replicate classic findings of strong lexicon-grammar relations and patterns of vocabulary composition across four languages. We extend these findings through novel analyses afforded by the Wordbank database. We explore a hypothesis that was not explicitly tested in these earlier studies, namely, that there is age-related variance in grammatical development that is unexplained by vocabulary. The identification of age-related variance would suggest the presence of developmental processes that regulate grammar learning, above and beyond those captured by measures of vocabulary size.

We also further probe the nature of lexical and grammatical development using sub-portions of the CDI forms. In our study of lexicon-grammar relations (Analysis 1), we delineate the grammar sections into items that reflect a broad distinction between inflectional morphology vs. sentence-level syntactic knowledge. We predict that age-related contributions to grammar should be evident to larger extent for syntax than morphology. In our study of vocabulary composition (Analysis 2), we leverage this technique to determine if age-related contributions vary across word classes. In particular, we predict that acquisition of predicates (verbs and adjectives) and function words should be relatively more dependent on syntactic factors than noun development, and thus should exhibit a greater influence of age.

We begin by describing the Wordbank database, the CDI measures, and our general analytic approach. We then describe two sets of analyses exploring the contribution of age to lexicon-grammar links (Analysis 1) and patterns of vocabulary composition (Analysis 2). These analyses reveal greater effects of age on aspects of grammar that are more aligned with syntax than with morphology, and greater effects of age on predicates and function words than on nouns. In the General Discussion, we consider potential domain-specific and domain-general explanations consistent with these findings.

Analyses

Methods

CDI Form Database We used Wordbank, a structured database of CDI data, to aggregate and archive CDI data across languages and labs. Wordbank is a repository that stores CDI data in a relational database for easy querying and analysis. By collecting language development data at an unprecedented scale, Wordbank enables the exploration of novel hypotheses about the course of lexical and grammatical development. At the time of writing, Wordbank includes data in four languages: English (Fenson et al., 2007), Spanish (Jackson-Maldonado, Thal, Marchman, Bates, & Gutiérrez-Clellen, 1993), Norwegian (Simonsen, Kristoffersen, Bleses, Wehberg, & Jørgensen, 2014), and Danish (Bleses et al., 2008), with both cross-sectional and longitudinal data. This dataset encompasses norming data from each language as well as a number of smaller-scale studies, some of which did not provide data from the grammar sections. Table 1 presents the total number of administrations and the number of administrations for which grammar (Word Form and Complexity) data were also available.

	Total admins	With grammar
Norwegian	10095	8505
English	5595	4137
Danish	3038	2074
Spanish	1094	1094
Total	19822	15810

Table 1: Number of administrations of the CDI (with and without grammar) in each language.

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

Table 2: Number of items in each section of each CDI instrument.

CDI Measures In all four languages, the CDI forms contain both vocabulary checklists and other questions relevant to the child's linguistic development. All of the data reported here come from the Words & Sentences form, administered to children ages 16–32 months. Each of these instruments includes a Vocabulary section, which asks whether the child produces each of around 700 words from a variety of semantic and syntactic categories (e.g., *foot*, *run*, *so*); a Word Form section, which asks whether the child produces each of around 30 morphologically inflected forms of nouns and verbs (e.g., *feet*, *ran*); 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 (e.g., *two foot / two feet*, *there a kitty / there's a kitty*). Each language's instrument is not just a translation of the English form, but rather was constructed and normed to reflect the lexicon and grammar of that language. Table 2 shows, for each language, the number of items in each of these sections.

To analyze lexical and grammatical development, we derive several measures. Each child's vocabulary size is computed as the proportion of 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 to make the scales comparable across languages.

Analysis 1: Syntax and Morphology

By two years, most children have a sizable working vocabulary, including verbs, prepositions, and closed class forms that perform grammatical work. They are also beginning to use multi-word combinations (e.g., *mommy sock*) and may demonstrate productive use of inflectional morphemes (e.g., past tense *-ed*). Previous studies have found a strong connection between the size of the lexicon and grammatical development as measured by the Complexity section, in many languages including English, Italian, Hebrew, and Spanish (see Bates & Goodman, 1999). However, no studies have had the power and cross-linguistic representation to go beyond this initial finding to explore relations to grammatical items that vary in morphosyntactic features. We extend it by examining grammatical development using two measures: the Word Form checklist as a window into morphology and the Complexity checklist as a window into syntax. For each measure, we investigate the effects of vocabulary size and age.

Results We wanted to estimate how much variance in children's syntactic and morphological development remains af-

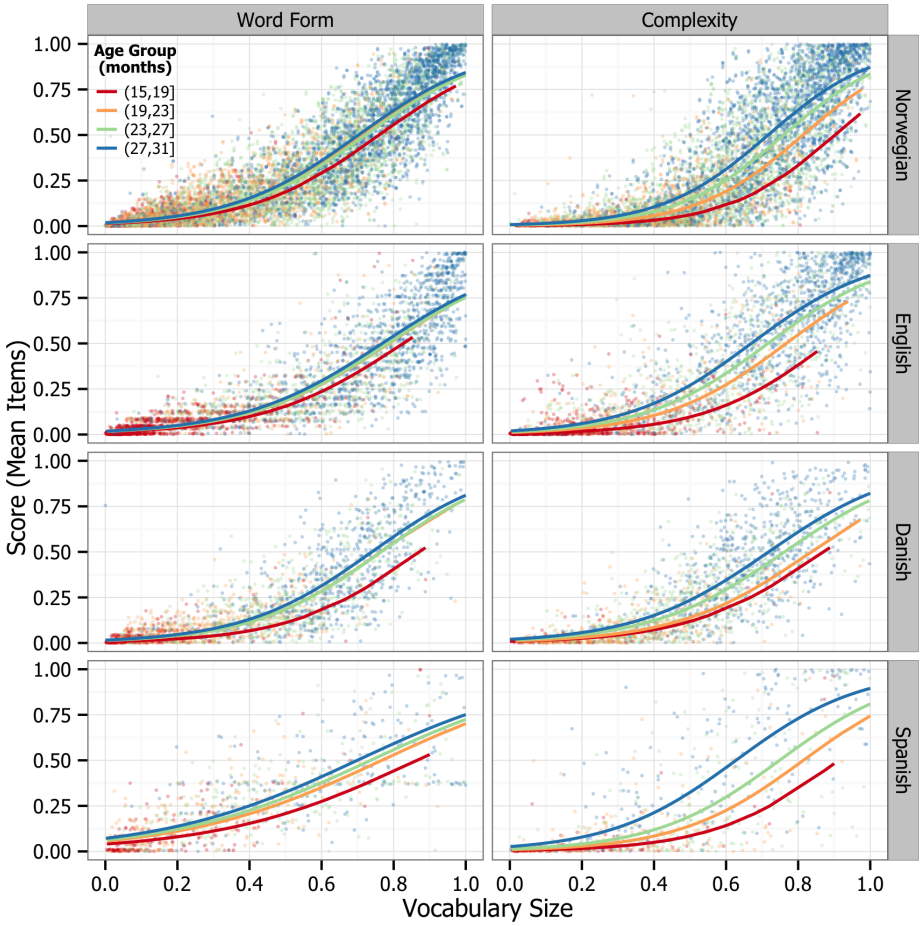

Figure 1: Each point shows an individual child, indicating their total vocabulary size and Word Form or Complexity score, with color showing their age bin (English $n = 4137$; Spanish $n = 1094$; Norwegian $n = 8505$; Danish $n = 2074$). Panels show different languages, and curves are logistic regression models fit separately for each language and measure. The models were specified as `score ~ vocab + age`.

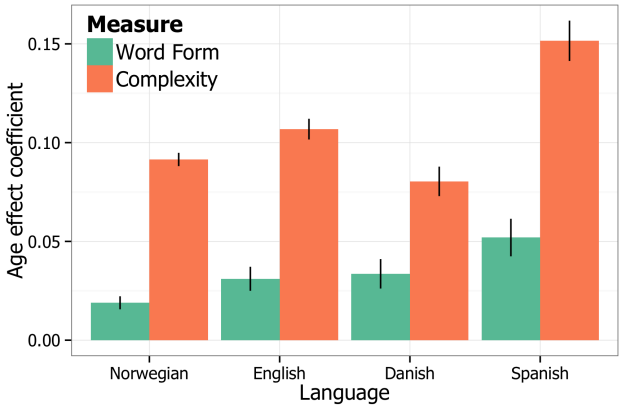

Figure 2: For each language and measure, the model's age effect coefficient. Ranges show the 95% confidence interval of the coefficient estimate. Across languages, Complexity has a substantially larger age effect than Word Form.

All data and code for these analyses are available at `https://github.com/dyurovsky/cdi-grammar`

ter accounting for that child's vocabulary size. Specifically, we asked whether age provides additional predictive power beyond vocabulary size. To estimate this effect, we fit logistic regression models to each child's Word Form and Complexity scores, predicting score as a function of vocabulary size and age in months. For all languages and measures, the evidence is overwhelmingly in favor of the model using both vocabulary and age as predictors, as compared to the model using only vocabulary (the smallest difference in AIC is 76).

Figure 1 shows data and models: each dot represents a child's score on a measure, while curves show the relationship between score and vocabulary size. For all languages, the curves for Word Form are near overlapping, showing little differentiation across age groups. This indicates only small contributions of age above and beyond vocabulary. In contrast, the curves for Complexity show a characteristic fan across age groups, indicating that the relationship between vocabulary size and complexity score is modulated by age.

Because of the size of our samples, all main effects and interactions are highly significant. To assess the extent of the age contribution to children's morphological and syntactic

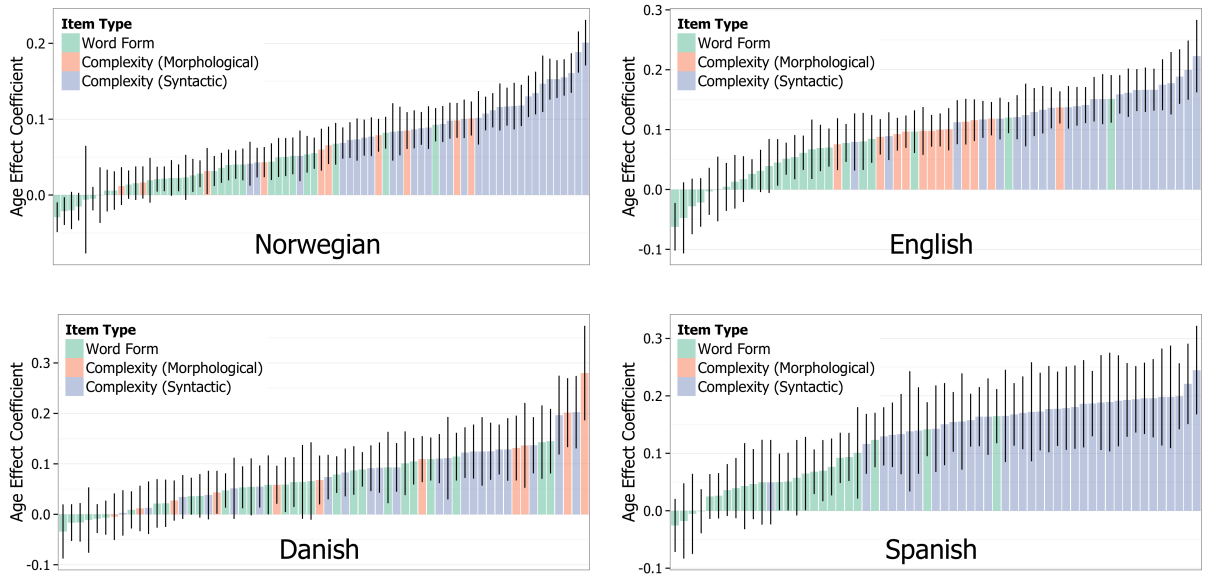

Figure 3: For each language and item, the model's age effect coefficient. Ranges show the 95% confidence interval of the coefficient estimate. Across languages, Word Form items tend to have smaller age effects, Morphological Complexity items tend to have middling age effects, and Syntactic Complexity items tend to have larger age effects. (Note: No Spanish complexity items had exclusively morphological content.)

development, we compared the coefficients of Word form and Complexity models. Figure 2 shows the coefficient of the age effect for each measure across languages. In each language, the age effect coefficient is substantially larger for Complexity than Word Form, indicating a greater age effect on those items that generally align with syntax than morphology.

Given the heterogeneous nature of the CDI instruments, particularly in the Complexity sections, we further broke down these items by classifying them as capturing more morphological or more syntactic phenomena. Items for which the difference between the simple and complex sentences is in the inflection of a noun or verb (such as *doggie kiss me / doggie kissed me*) were coded as Morphological. The remainder of the items were coded as Syntactic, since they involve the use of some sentence-level syntactic construction (such as *doggie table / doggie on table*).

We then fit predictive models as above separately for every item. Figure 3 shows the age effect coefficient of each item. In general, there is a three-way split: age effects are smallest for Word Form items, then Morphological Complexity items, and largest for Syntactic Complexity items, suggesting more syntactic phenomena have greater age contributions.

Discussion Building on previous analyses that showed a strong relationship between lexical and grammatical development, we added age into this relationship. Across languages, our measures of syntactic development consistently showed greater age modulation than measures of morphological development. Further distinguishing between items that were more reflective of morphology than syntax, we again found greater age effects for more syntactic items. Thus, this analysis provides evidence for a relationship between syntactic development and age, not captured by lexical development.

Analysis 2: Vocabulary Composition

Early vocabulary development is typically characterized by learning of names for caregivers and common objects, while later in development, children tend to diversify their vocabulary by increasing the proportion of predicates (verbs and adjectives) and closed class words. This over-representation of nouns has been found across a number of analyses and in a variety of languages (Bates et al., 1994; M. Caselli et al., 1995; Bornstein et al., 2004), though not all (Tardif, 1996; Choi & Gopnik, 1995). 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 predicates and function words 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 account should syntactic information be a primary information source (though of course syntax might be more informative for abstract nouns). In contrast, for other types of words, syntax should be more important for learning their meaning.

On the syntactic bootstrapping hypothesis (Gleitman, 1990; Fisher, Gertner, Scott, & Yuan, 2010), verbs especially 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)). A similar argument can be made for adjectives, since identification of the modified noun is simi-

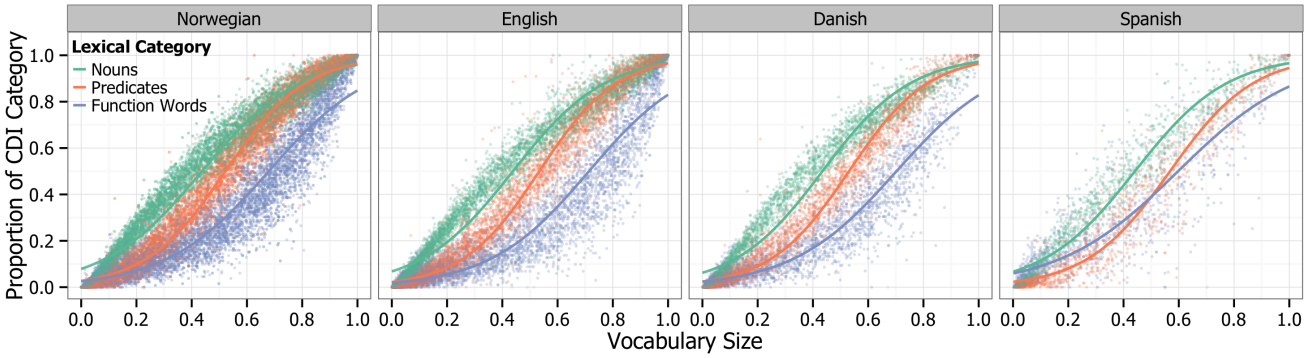

Figure 4: Proportion of a particular CDI category, plotted by total vocabulary size. Each point shows an individual child, with color showing their noun, predicate, and function word vocabulary. Panels show different languages, and curves are regression models fit separately for each language, specified as `proportion ~ vocab + age` (English $n = 5595$; Spanish $n = 1094$; Norwegian $n = 10095$; Danish $n = 3038$).

larly critical for inferring the meaning of the modifier. And by the same logic, function words should be even harder to learn without some understanding of their syntactic role. Thus, if syntactic development is related in some way to age, we should see larger age effects on predicate and function word vocabulary than on noun vocabulary.

Results Each CDI form contains a mixture of words in different classes. We adopt the categorization of Bates et al. (1994), splitting words into nouns, predicates (verbs and adjectives), function words, and other words. For each child's vocabulary, we compute the proportion of the total words in each of these categories that they are reported to produce.

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 a child's knowledge of a particular class, while curves show the relationship between a class and the whole vocabulary. If categories grow independently of one another, these curves should approximate the diagonal. This pattern is not what we 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.

Next, we measure the contribution of age to vocabulary composition. We fit a logistic model to all children's data for each word class, predicting word-class proportion as a function of total vocabulary and age (as in Analysis 1). Figure 5 shows age coefficients for each of these models across languages. In both English and Norwegian (and to some extent Danish and Spanish), the age coefficient is substantially larger for predicates and for function words than for nouns. This asymmetry can be interpreted as evidence that, for two vocabulary-matched children, the older child would tend to have relatively more predicates and function words than the younger. For Danish and Spanish, this effect is clear for function words but not predicates.

Discussion We replicated previous analyses (Bates et al., 1994) showing an over-representation of nouns in the developing lexicon and an under-representation of predicates and function words. We also predicted that—if syntactic generalization was in some way tied to age—predicates and function words would show relatively more age influence than nouns. Although there was some cross-language variation in predicate terms, overall this prediction was confirmed across the four languages we examined. Thus, this analysis provides additional evidence for a relationship between syntactic development and age, independent of the growth of the lexicon.

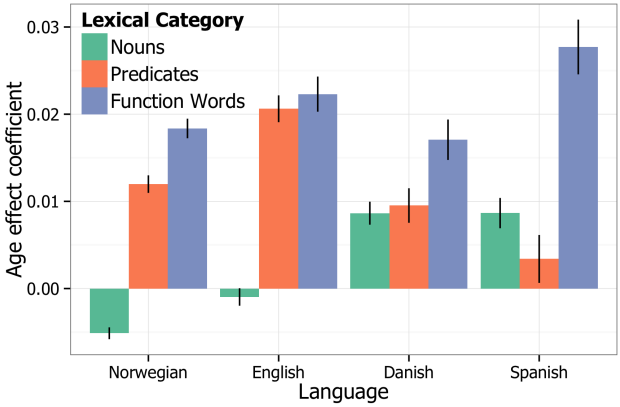

Figure 5: For each language and lexical category, the model's age effect coefficient. Ranges show the 95% confidence interval of the coefficient estimate. Across languages, predicates and function words tend to have a substantially larger age effect than nouns.

General Discussion

The current study revisits classic findings but also explores novel questions regarding lexicon-grammar relations and vocabulary composition through Wordbank, a newly-developed web-based tool for cross-linguistic analyses of large CDI datasets. Our results provided general support for a lexicalist view, in that, in four languages, variance in vocabulary production strongly aligned with variance in grammar. However, we also estimated additional age-related contributions, specifically contrasting the links to morphological forms vs. syntactic constructions, and for different lexical categories.

In general, we find that measures of grammar that are more closely aligned with syntax are modulated by age to a greater extent than those reflecting inflectional morphology. Also, we find that the trajectories of predicate and function word representation in the vocabulary are modulated by age to a greater extent than noun representation (albeit with some variability across languages). Both findings suggest a place for developmental processes that facilitate grammatical acquisition beyond pure lexical growth.

Our analyses suggest interesting new areas of research regarding possible mechanisms driving children's early lexical development and how those mechanisms might support children's transition from single words to more morphosyntactically complex utterances. One possibility is that these developments are dependent on maturational factors that operate on grammatical development in a domain-specific way, independent of lexical-semantic processes. Another possibility is that age-related effects represent more domain-general learning mechanisms, such as attention or working memory, that provide differential support for sentence-level processes than word-internal ones (Gathercole & Baddeley, 2014). Future studies should also explore the extent to which lexical and age-related processes are shaped, either independently or in tandem, by features of the learning environments that children experience (e.g., Weisleder & Fernald, 2013).

Questions about the nature of morphosyntactic representations in early language have often seemed deadlocked. But by mapping out developmental change across large samples and multiple languages, our findings here challenge theories across the full range of perspectives to more fully describe the mechanistic factors underlying the interaction of vocabulary, grammar, and development.

Acknowledgments

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

References

Baker, M. C. (2005). Mapping the terrain of language learning. *Language Learning and Development*, *1*(1), 93–129.

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., 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.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. O., & Basbøll, H. (2008). The Danish Communicative Developmental Inventories: Validity and main developmental trends. *Journal of Child Language*, *35*(03), 651–669.

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

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., ... Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development*, *75*(4), 1115–1139.

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*.

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

Gathercole, S. E., & Baddeley, A. D. (2014). *Working memory and language processing*. Psychology Press.

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

Jackson-Maldonado, D., Thal, D., Marchman, V., Bates, E., & Gutiérrez-Clellen, V. (1993). Early lexical development in Spanish-speaking infants and toddlers. *Journal of Child Language*, *20*(03), 523–549.

Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. N. (2014). The Norwegian Communicative Development Inventories: Reliability, main developmental trends and gender differences. *First Language*, *34*(1), 3–23.

Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from Mandarin speakers' early vocabularies. *Developmental Psychology*, *32*(3), 492.

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

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152.

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