I *can* believe it: Quantitative evidence for closed-class category knowledge in an English-speaking 20- to 24-month-old child

ALANDI BATES, LISA PEARL, & SUSAN R. BRAUNWALD
University of California, Irvine[*]

1 Introduction

Adults are believed to have abstract syntactic categories that they use to generate their observable utterances (e.g., closed-class categories like NEGation and open-class categories like VERB to generate *don't go*). However, there's been significant debate about when children develop syntactic categories and how to accurately assess what category knowledge they have when. We review prior approaches to assessing children's developing knowledge of syntactic categories, and then present our quantitative approach, which synthesizes insights from this prior work. This allows us to (i) define possible child representations for multi-word combinations, and (ii) calculate the observed vs. expected linguistic production properties for each possible representation. We use this approach to investigate the existence of both closed-class and open-class syntactic categories in a 20- to 24-month-old child's verb phrases. We evaluate whether the child's observed production matches the expected production when the child uses a specific category representation, and find that the child's productions are compatible only with representations that have adult-like closed-class categories (NEG, AUXiliary), but not adult-like open-class categories (NOUN, VERB). We conclude with implications for the development of syntactic categories.

2 Syntactic category knowledge in children

There hasn't been a clear consensus for when children develop syntactic categories, whether open-class or closed-class. Some studies suggest that knowledge of certain categories – either rudimentary or adult-like – may be in place as early as age 2 (Pinker 1984; Valian

[*] We have benefited from comments and suggestions from the Computation of Language Laboratory at UC Irvine, and the audiences at UROP 2017 at UC Irvine, CAMP 2017 at UCLA, and the Berkeley Linguistics Society 2018 meeting. All errors are of course our own and not at all their fault.

1986; Capdevila i Batet and Llinàs i Grau 1995; Booth and Waxman 2003; Rowland and Theakston 2009; Theakston and Rowland 2009; Yang 2010, 2011; Shin 2012; Meylan et al. 2017), while others argue that such knowledge only emerges much later (Pine and Lieven 1997; Tomasello 2004; Kemp et al. 2005; Tomasello and Brandt 2009; Theakston et al. 2015). Taken together, there seems to be some agreement that children may have rudimentary knowledge of open-class categories (NOUN, ADJective) fairly early, but don't refine these into adult-like open-class categories until later. However, for closed-class categories (DETerminer, NEG, AUX), there isn't yet consensus on when either rudimentary or adult-like versions of these categories develop. This may be due in part to the different quantitative analysis approaches that prior research has adopted. More generally, many prior studies demonstrate that there's utility in quantitatively analyzing children's productions to determine the nature of their underlying representations. However, there are several ways to go about this analysis.

Notably, many prior approaches harnessed the intuition that syntactic categories allow children to transfer knowledge about how a word from one category (a NEG like *don't*) combines with words from another category (VERBs like *go* and *believe*) in order for the child to generate novel productions. For example, if the two-word combination *don't go* hasn't been heard before, then it must have been generated based on units that are more abstract than individual lexical items. This combinatory *productivity* – that is, the generation of combinations that haven't been heard before – is a sign of abstract syntactic category knowledge. Yet, what about combinations that have in fact been heard before? How do we know if children are generating them in a productive way that relies on syntactic categories (the way adults would) or instead in some other way that relies on the individual lexical items? This is where prior approaches diverge from each other. Below, we describe our approach, which is inspired primarily by Yang (2010), Yang (2011), and Pine et al. (2013).

3 Possible child category representations for multi-word combinations

We consider three types of syntactic category representation that very young children could use to form multi-word combinations (like *don't go*). The representation types differ with respect to whether the child produces multi-word combinations according to (i) the distribution of multi-word combinations in her input (NOT productive), (ii) both her input distributions and an internal category representation (SEMI-productive), or (iii) internal category representations alone (fully PRODuctive).

A child using a NOT productive representation can only generate a multi-word combination if she's heard it in her input (e.g., *don't go* → (*don't*+*go*)$_{Input}$). So, any multi-word combination she generates is effectively a memorized amalgam; how often she generates a particular amalgam depends on how frequently that amalgam was in her input. This contrasts with a child using a SEMI-productive representation, who relies on an internal category for generating one part of the multi-word combination and her input combinations with that category for generating the other part (e.g., *don't go* → (AUX+*go*)$_{Input}$). Here, if she's heard *go* used with an AUX – any AUX, not just *don't* – she can generate *don't go* this way. So, the child can generate some novel expressions, but still relies on input distributions when the expressions involve words that aren't part of a syntactic category. However, a child with a fully PRODuctive representation can generate novel combinations by relying on her internal syntactic categories alone, rather than input distributions of multi-word combinations (e.g., *don't go* → AUX+VERB). That is, the child draws on her internal category knowledge when generating utterances the way we believe adults typically do, and has the greatest capacity for novel multi-word combinations.

4 How can we quantitatively measure representational knowledge?

4.1 Lexical overlap as a measure of category knowledge

Lexical overlap is often used as a measure for productivity (Yang 2010, 2011; Pine et al. 2013), and is meant to capture the intuition that words in one category can be freely combined with words from another. That is, category members are effectively interchangeable in those combinations. For example, an AUX category would allow any of its member words (e.g., *don't*, *do*, *can*, etc.) to combine with verbs like *go*. So, we would expect to see multiple auxiliaries used with any given verb (e.g., *don't go, do go, can go*, etc.) – that is, there would be *overlap* in the use of auxiliary *lexical* items. So, to assess a category, we need to examine its lexical overlap with respect to words that the category can combine with. For example, when assessing AUX, we can look at how many verbs have lexical overlap when it comes to auxiliaries.

We assess both the *Observed* lexical overlap present in a speaker's productions and the *Expected* lexical overlap if the speaker used a particular representation to generate those productions. While there's only one Observed score per potential category (e.g., an AUX$_{Observed}$ for how auxiliaries combine with verbs), there's an Expected score for each potential representation the speaker could be using to generate her productions. If the expected overlap for a particular representation matches the observed overlap well enough, this indicates that representation is compatible with the speaker's output.

At the category level, the two representations are that the category is (i) present (i.e, $\{$*don't, do, can*, etc.$\}$ $\in$ AUX), or (ii) absent (i.e., *don't, do, can*, etc. are simply individual words that aren't interchangeable syntactically). At the multi-word combination level, we focus on combinations made up of two potential categories (e.g., *don't go*, which could involve AUX and VERB). For these combinations, there are three possible representations: NOT, SEMI, and fully PRODuctive. More specifically, a NOT productive representation has both categories absent; a SEMI-productive representation has one category present and the other absent; a fully PRODuctive representation has both categories present.

4.2 Calculating Observed and Expected overlap

We first describe how to calculate the lexical overlap for a potential category with respect to a set of words it combines with. This is the core calculation that will be used for calculating Observed and Expected overlap scores for multi-word combinations. We then describe how to calculate the Observed overlap for multi-word combinations and the Expected overlap for each of the three representation types (NOT, SEMI, and PROD).

For a potential category whose status is $Unknown$ (like AUX), we look at the lexical overlap in words which that category combines with (like verbs, which would be $w_{comb} \in Combine$ in (1)). Lexical overlap itself is defined very conservatively, following previous studies using it (Yang 2010, 2011; Pine et al. 2013): if more than one word $w_{unk} \in Unknown$ (e.g., both *don't* and *can*) appears in combination with a word $w_{comb} \in Combine$ (e.g., *go*), then lexical overlap for w_{comb} is 1. Otherwise, if w_{comb} only ever appears in combination with a single word $w_{unk} \in Unknown$ (e.g., *don't go* is the only combination of an auxiliary with *go*), lexical overlap is 0. This is $overlap_{w_{comb}}$ in (1). The total overlap $overlap_{Combine}$ is the lexical overlap average across all words that the potential category can combine with ($w_{comb} \in Combine$). For example, this would be the lexical overlap average across all verbs when assessing potential category AUX on how it combines with verbs. So, if there are 50 verbs that combine with auxiliaries in the data sample, then individual overlap scores $overlap_{w_{comb}}$ are calculated for each of these 50 verbs, and the average is taken of all 50 scores.

$$\begin{aligned} overlap_{w_{comb}} &= \begin{cases} 1\colon w_{comb}\ occurs\ with\ > 1\ word\ w_{unk} \in Unknown \\ 0\colon w_{comb}\ occurs\ with\ only\ 1\ word\ w_{unk} \in Unknown \end{cases} \\ overlap_{Combine} &= \frac{\sum_{w_{comb} \in Combine} overlap_{w_{comb}}}{|Combine|} \end{aligned} \tag{1}$$

For a multi-word combination involving two potential categories (e.g., AUX+VERB), observed overlap can be calculated with respect to each category (e.g., with respect to verbs when assessing AUX and with respect to auxiliaries when assessing VERB). The observed overlap calculation is just as in (1), shown in (2) over the set of speaker productions that involve those kind of multi-word combinations S_{Obs} (e.g., all combinations of auxiliaries+verbs for AUX+VERB).

$$Observed = overlap_{Combine}(S_{Obs}) \tag{2}$$

Expected overlap, as mentioned, depends on the representation the speaker uses to generate her multi-word combinations. A more detailed walk-through of the Expected overlap calculation for all three representation types is in Appendix A in the supplementary materials[1], but we sketch the core intuitions here.

A child using a NOT productive representation (e.g., *don't go* → *don't+go*) generates multi-word combinations as memorized amalgams from her input, based on the frequency of those input amalgams. To simulate this, we generate multi-word combination data samples $S_{Exp_{Not}}$ that are the same size as the observed speaker multi-word combination sample

[1] Available at http://sites.uci.edu/alandibates/files/2018/07/Bates_Pearl_Braunwald_2018_BLS.pdf and www.socsci.uci.edu/~lpearl/papers/BatesPearlBraunwald2018_BLS.pdf.

S_{obs}; these samples are drawn from the speaker's input. So, if there are 100 auxiliary+verb combinations in the speaker's output, we generate 100 auxiliary+verb combinations, based on the auxiliary+verb distribution in the speaker's input.

Suppose we have a word w_{unk} (like *don't*) from a category with $Unknown$ status (like AUX). The combinations that w_{unk} is generated with depend on the combinations from the speaker's input that w_{unk} appeared with. So, the probability of sample s_i being w_{unk}+w_{comb} (e.g., *don't*+*go*) depends on how often w_{unk}+w_{comb} appeared in the speaker's input ($p_{w_{unk}w_{comb_{Input}}}$). We then calculate the lexical overlap of the NOT sample and use that as the Expected overlap for a child using the NOT productive representation (3).

$$\begin{gathered} s_i \in S_{Exp_{Not}}, s_i = w_{unk}w_{comb} \ \propto p_{w_{unk}w_{comb_{Input}}} \\ Expected_{Not} = overlap_{Combine}(S_{Exp_{Not}}) \end{gathered} \tag{3}$$

We can use a similar approach to calculate the Expected overlap for the SEMI-productive representation (e.g., *don't go* → AUX+*go* or *don't*+VERB). For simplicity, we abbreviate the word from the category as w_{+cat} and the word not from a category as w_{-cat}. Then, to generate combination $w_{unk}w_{comb}$, the child relies on her internal category representation to generate word w_{+cat} and looks to her input to see how often words from this category combine with word w_{-cat}. So, she would generate combination $w_{unk}w_{comb}$ with about the same frequency she heard examples of either $Unknown$+w_{comb} (if $Unknown$ is the category) or w_{unk}+$Combine$ (if $Combine$ is the category). To simulate this process, we generate multi-word combination data samples $S_{Exp_{Semi}}$ that are the same size as the observed speaker multi-word combination sample S_{obs}. The probability of multi-word sample $s_i \in S_{Exp_{Semi}}$ involving a specific word $w_{+cat} \in Category$ combined with w_{-cat} depends on how often any word in $Category$ combines with w_{-cat} in the speaker's input ($p_{Category}\ p_{w_{-cat_{Input}}}$). We then calculate the lexical overlap for the SEMI sample and use that as the Expected overlap for a child using a SEMI-productive representation (4).

$$\begin{gathered} s_i \in S_{Exp_{Semi}}, s_i = w_{+cat}w_{-cat} \ \propto p_{Category}\ p_{w_{-cat_{Input}}} \\ Expected_{Semi} = overlap_{Combine}(S_{Exp_{Semi}}) \end{gathered} \tag{4}$$

A child with a fully PRODuctive representation (e.g., *don't go* → AUX+VERB) generates her multi-word combinations by relying on internal category representations for both words. Yang (2010; 2011) describes an analytical solution for the Expected lexical overlap when both categories exist (5). We can use this here, rather than generating expected samples and calculating lexical overlap for those samples. The key intuition involves the definition of lexical overlap, where a word w_{comb} shows lexical overlap if more than one word $w_{unk} \in Unknown$ combines with w_{comb}. So, we can calculate this analytically as 1 minus the probability that w_{comb} will (i) never appear with any word in $Unknown$, or (ii) only appear with a single word in $Unknown$. This is equivalent to the formula in (5) for the Expected overlap for word w_{comb}, whose derivation is discussed more fully in Appendix A of the supplementary materials. All word probabilities are estimated based on the speaker's productions of w_{comb} and w_{unk} (i.e., $p_{w_{comb}} = p_{w_{comb_{Obs}}}$, $p_{w_{unk}} = p_{w_{unk_{Obs}}}$). This is because all words in these combinations are generated from an underlying internal category, and so don't rely on the speaker's input. As with the original calculation of lexical overlap, these individual word overlaps are averaged to get the Expected overlap.

$$\begin{aligned} overlapprod_{w_{comb}} &= 1 - P(no\, w_{comb}) - P(only\ 1\ w_{comb}) \\ &= 1 + (|Unknown| - 1)(1 - p_{w_{comb}})^{S_{obs}} \\ &- \sum_{w_{unk} \in Unknown} (p_{w_{comb}} * p_{w_{unk}} + 1 - p_{w_{comb}})^{S_{obs}} \\ Expected_{Prod} &= \frac{\sum_{w_{comb} \in Combine} overlapprod_{w_{comb}}}{|Combine|} \end{aligned} \tag{5}$$

5 Data

Given that syntactic category knowledge may be present as early as two years old, we investigated data from a child (hereafter **L**) just before the age of two. L's productions between 20 and 24 months were hand-recorded in daily diary data in the Susan R. Braunwald Language Acquisition Diaries (Braunwald 2015), and represent a rich cross-contextual sample of L's whole language acquisition experience. We also included child-directed mealtime input from L's caretaker when L is between 20 and 24 months, which are in the Braunwald corpus (Braunwald 1995) in CHILDES (MacWhinney 2000). Between these two datasets, we had a dense longitudinal sample from the same child of both her output (from the daily diary data) and her input (from the child-directed mealtime speech).

Because verbs are often considered the backbone of language, given the wealth of information they encode about events and their participants (Gleitman 1990; Tomasello and Merriman 1995), we focused our investigation on syntactic categories in verb phrases (VPs): VERB itself, along with NEG, AUX, ADJective, PREPosition, and NOUN. The VPs from L and her caretaker were manually extracted and syntactically annotated, yielding 2,154 child-produced VPs from L and 2,184 adult-produced VPs from L's caretaker.[2] We additionally restricted our analyses to lexical items shared by L and her caretakers to facilitate comparisons between their lexical overlaps, as Pine et al. (2013) note that differing vocabulary sizes can disrupt comparisons between subjects. This yielded 105 verbs total. Because the quantitative analysis we use requires sample sizes that are sufficiently large, we decided to only include potential categories where there were at least 100 tokens in L's productions (Goldin-Meadow and Yang (2016), Charles Yang, p.c.). For example, at least 100 instances of nouns combining with verbs were needed to include the potential category NOUN. This led to us including two open-class categories (VERB and NOUN) and two closed-class categories (AUX and NEG). Table 1 shows the types and tokens from L and L's caretaker for each potential category and multi-word combination involving those potential categories.[3]

6 Evaluating the possible representations

Because we consider four potential categories in VPs, a child's complete category representation involves something about each potential category. In particular, for each of the four categories, the child either has a category representation for it (e.g., all verbs categorized as VERB) or doesn't (e.g., all verbs treated as individual words). This yields 16

[2] We note that L's verb usage seems typical of her age group, as assessed by a corpus analysis of verb production frequency from 93 North American English children between the ages of 20 and 24 months from the CHILDES database (MacWhinney 2000). In particular, based on L's verbs and the verbs used by these 93 children (10432 verb tokens and 322 verb types), L used 16 of their 20 most frequent verbs, and they collectively used 15 of her 20 most frequent verbs.

[3] We note that the multi-word combination counts were analyzed irrespective of order. So, for example, VERB+NOUN combinations include instances such as *I go* (NOUN VERB) and *have coffee* (VERB NOUN). We also note that contractions (e.g., *don't*) were analyzed as belonging to both potential categories their components were from. For example, *don't* was counted as an instance of both an AUX and a NEG, because *do* is an AUX for adults and *n't* is a NEG for adults.

Table 1: Types and tokens of potential categories and multi-word combinations involving those categories in the VPs of L and L's caretaker.

	L		**L's caretaker**	
Potential category	Types	Tokens	Types	Tokens
VERB	105	2642	105	3164
NOUN	504	2330	617	2606
AUX	21	198	38	454
NEG	6	114	11	104
Multi-word combination	Types	Tokens	Types	Tokens
VERB+NOUN	1111	2330	1426	2606
VERB+AUX	95	198	239	454
VERB+NEG	42	114	61	104

possible category representations (2^4) the child might have. The completely NOT productive representation (Rep$_{\text{NOT}}$) has no categories – that is, all four potential categories are absent, and the words that would be in them are represented only as individual words. In contrast, the fully PRODuctive representation (Rep$_{\text{PROD}}$) is the one where all four potential categories are present for the child, just as they are for adults like L's caregiver. The remaining possible category representations are SEMI-productive (Rep$_{\text{SEMI}}$), because they involve at least one category present and at least one absent. For example, one possible Rep$_{\text{SEMI}}$ would have VERB and NEG as categories while nouns and auxiliaries would be represented as individual words only.

We can evaluate each category representation based on how well its Expected lexical overlap matches L's Observed lexical overlap. Because we have four potential categories (VERB, NOUN, AUX, NEG), we can calculate lexical overlap scores for all multi-word combinations involving these potential categories. More specifically, for any multi-word combination, we calculate lexical overlap scores with respect to the *Unknown* category, and either word in the multi-word combination can be the *Unknown* category being assessed. For example, in VERB+NOUN multi-word combinations, either the VERB or the NOUN could be assessed as the *Unknown* category while the other category serves as the collection of words in *Combine* (i.e., *Unknown*=NOUN and *Combine*=VERB, or vice versa). The Observed and Expected lexical overlap scores are calculated based on which set of words is *Unknown* and which set is *Combine*.

Importantly, the Expected calculation depends on the status of *Unknown* and *Combine* in the category representation itself. For example, consider *Unknown*=NOUN while *Combine*=VERB. A representation where both categories were absent would calculate the Expected overlap score using $Expected_{Not}$, a representation where NOUN was present while VERB was absent would use $Expected_{Semi}$, and a representation where both categories were present would use $Expected_{Prod}$. A more detailed walk-through of this calculation is in Appendix B of the supplementary materials.

Once we have the Observed and Expected lexical overlap scores for a category representation, how do we tell that they match sufficiently? Following Goldin-Meadow and Yang (2016), we use Lin's Concordance Correlation Coefficient (**LCCC**, represented with ρ_c: Lawrence and Lin (1989)) to assess agreement between the Observed and Expected overlap. LCCC measures the agreement between two sets of observations on a scale from -1 to 1, with a ρ_c of -1 indicating perfect disagreement, 1 indicating perfect agreement, and 0 indicating no agreement. So, given that there are multiple lexical overlap scores for each category representation (one for each legitimate multi-word combination within a particular category representation), we assess ρ_c for the Observed vs. Expected overlap scores within that category representation (Table 2).

With ρ_c scores for each of the 16 possible category representations, we then need to decide which representations have a "good enough" match between Observed and Expected overlap. Unfortunately, there isn't a current consensus about what the threshold should be for good agreement with the LCCC (Altman 1990; McBride 2005). Given this, we decided to leverage L's input data, with the idea that L's caregiver had a fully productive category representation (Rep_{PROD}) involving VERB, NOUN, AUX, and NEG. Because of this, the agreement between the Observed overlap in L's caregiver's productions and the Expected overlap from the Rep_{PROD} category representation could serve as a "good enough" threshold of agreement. More specifically, because we believe the Rep_{PROD} category representation generated L's caregiver's productions, the ρ_c obtained for that representation is a reasonable cutoff for when a category representation in general matches sufficiently well with the observed data. We found ρ_c = 0.901 when comparing the Expected overlap from a Rep_{PROD} category representation against the Observed overlap in L's caretaker's productions. We take this value as our threshold for when L's possible category representations are sufficiently compatible with her output (Table 2).

Table 2: LCCC scores for the 16 possible category representations L could have, comparing her Observed lexical overlap against the lexical overlap Expected by each possible category representation. Representations with sufficient agreement (>0.901) are indicated.

Represention	VERB	NOUN	AUX	NEG	***LCCC*** ρ_c	**Sufficient agreement?**
Rep$_{\text{NOT}}$	✗	✗	✗	✗	0.873	
Rep$_{\text{PRODUCTIVE}}$	✓	✓	✓	✓	0.838	
Rep$_{\text{SEMI}_1}$	✓	✓	✓	✗	0.851	
Rep$_{\text{SEMI}_2}$	✓	✓	✗	✓	0.802	
Rep$_{\text{SEMI}_3}$	✓	✗	✓	✓	0.753	
Rep$_{\text{SEMI}_4}$	✗	✓	✓	✓	0.867	
Rep$_{\text{SEMI}_5}$	✓	✓	✗	✗	0.809	
Rep$_{\text{SEMI}_6}$	✓	✗	✓	✗	0.753	
Rep$_{\text{SEMI}_7}$	✓	✗	✗	✓	0.719	
Rep$_{\text{SEMI}_8}$	✗	✗	✓	✓	0.915	Yes
Rep$_{\text{SEMI}_9}$	✗	✓	✗	✓	0.891	
Rep$_{\text{SEMI}_{10}}$	✗	✓	✓	✗	0.765	
Rep$_{\text{SEMI}_{11}}$	✓	✗	✗	✗	0.715	
Rep$_{\text{SEMI}_{12}}$	✗	✓	✗	✗	0.794	
Rep$_{\text{SEMI}_{13}}$	✗	✗	✓	✗	0.850	
Rep$_{\text{SEMI}_{14}}$	✗	✗	✗	✓	0.935	Yes

Though agreement values range from ρ_c = 0.715 to 0.935, only two of the sixteen possible category representations are above the "good enough" threshold of 0.901. Both category representations are SEMI-productive (Rep$_{\text{SEMI}_8}$=0.915, Rep$_{\text{SEMI}_{14}}$=0.935). As Table 2 shows, neither category representation involves knowledge of the open-class categories VERB or NOUN. Instead, both involve knowledge of the closed-class category NEG (covering all 6 negations in L's productions) and one also includes knowledge of AUX (covering all 21 auxiliaries in L's productions).

7 General discussion

We find quantitative evidence for adult-like closed-category knowledge in a child before age two, but not for adult-like open-class category knowledge. This suggests open-class categories may take longer to develop into adult-like categories, compared with closed-class categories. Notably, this supports advocates of early closed-class category knowledge (Capdevila i Batet and Llinàs i Grau 1995; Rowland and Theakston 2009; Theakston and Rowland 2009; Yang 2010; Shin 2012; Meylan et al. 2017). This also aligns with the idea that closed-class categories may provide a way of breaking into the categorization of open-class lexical items (Höhle et al. 2004).

One reason why adult-like closed-class categories might emerge sooner is that they have fewer members than their open-class counterparts. For instance, in our data sample, there were 6 negations and 21 auxiliaries, in contrast with 105 verbs and 504 nouns. So, with fewer members, it may be easier to cluster all relevant lexical items into their respective closed-class categories.

Another difference is the existence of salient semantic sub-classes within the open-class categories – these sub-classes might form a natural category for the child, rather than the child clustering all relevant lexical items into an adult-like NOUN or VERB. For example, 6-month-olds recognize concrete nouns specifically (Bergelson and Swingley 2012), and so this kind of noun might persist as a natural class even after children recognize and use other nouns. Similarly, 3- and 4-year-olds have distinct comprehension behavior for passives involving actional verbs like *hug* vs. non-actional verbs like *surprise, find, forget*, or *love* (Nguyen et al. 2016; Nguyen and Pearl 2018). So, the actional lexical feature (and others) may cause younger children to form categories for subsets of verbs, rather than having a single VERB category.

The existence of these potential child categories that are subsets of the adult category highlights an interesting area for future research: evaluate potential child-like categories against young children's productions, such as L's data. Recall that we only evaluated adult-like categories here, which encompass all relevant lexical items (e.g. all nouns for NOUN). However, the very same quantitative approach can be used to assess whether potential child-like categories, which may be subcategories of the adult versions (e.g., CONCRETE-

NOUN VS. NON-CONCRETE-NOUN), best match children's productions. All that's needed is to define which words belong to which child-like category, and the possible multi-word combination types involving these child-like categories. If a good enough match were found to the child's productions, this would provide quantitative support for a specific child-like category, distinct from the absence of any category or from the presence of an adult-like category. That is, we would have quantitative support for a particular developing category. We leave this exciting possibility for the future.

8 References

Douglas G Altman. *Practical statistics for medical research.* CRC press, 1990.

Elika Bergelson and Daniel Swingley. At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9): 3253–3258, 2012.

Amy Booth and Sandra Waxman. Mapping words to the world in infancy: On the evolution of expectations for nouns and adjectives. *Journal of Cognition and Development*, 4(3): 357–381, 2003.

Susan Braunwald. *Differences in the acquisition of early verbs: Evidence from diary data from sisters*, pages 81–111. Psychology Press, 1995.

Susan Braunwald. Susan R. Braunwald Language Acquisition Diaries. In *UC Irvine Special Collections and Archives*. UC Irvine, 2015.

Montserrat Capdevila i Batet and Mireia Llinàs i Grau. The acquisition of negation in English. *Atlantis*, pages 27–44, 1995.

Lila Gleitman. The structural sources of verb meanings. *Language Acquisition*, 1(1):3–55, 1990.

Susan Goldin-Meadow and Charles Yang. Statistical evidence that a child can create a combinatorial linguistic system without external linguistic input: Implications for language evolution. *Neuroscience & Biobehavioral Reviews*, 2016.

Barbara Höhle, Jürgen Weissenborn, Dorothea Kiefer, Antje Schulz, and Michaela Schmitz. Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy*, 5(3):341–353, 2004.

Nenagh Kemp, Elena Lieven, and Michael Tomasello. Young children's knowledge of the determiner and adjective categories. *Journal of Speech, Language, and Hearing Research*, 48(3):592–609, 2005.

I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

B MacWhinney. The CHILDES project: The database. Psychology Press, 2000.

GB McBride. A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. *NIWA Client Report: HAM2005-062*, 2005.

Stephan C Meylan, Michael C Frank, Brandon C Roy, and Roger Levy. The emergence of an abstract grammatical category in children's early speech. *Psychological Science*, 28 (2):181–192, 2017.

Emma Nguyen and Lisa Pearl. Do You Really Mean It? Linking Lexical Semantic Profiles and the Age of Acquisition for the English Passive. In *35th West Coast Conference on Formal Linguistics*, pages 288–295. Cascadilla Proceedings Project, 2018.

Emma Nguyen, Diane Lillo-Martin, and William Snyder. Actionality speaks louder than felicity: Children's comprehension of long passives. In *Proceedings from the Generative Approaches to Language Acquisition*, pages 232–246, 2016.

J.M Pine and E.V. Lieven. Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, 18(02):123–138, 1997.

Julian M Pine, Daniel Freudenthal, Grzegorz Krajewski, and Fernand Gobet. Do young children have adult-like syntactic categories? Zipf's law and the case of the determiner. *Cognition*, 127(3):345–360, 2013.

Steven Pinker. *Language learnability and language development.* Harvard University Press, Cambridge, MA, 1984.

Caroline F Rowland and Anna L Theakston. The acquisition of auxiliary syntax: A longitudinal elicitation study. Part 2: The modals and auxiliary DO. *Journal of Speech, Language, and Hearing Research*, 52(6):1471–1492, 2009.

Yu Kyoung Shin. A New Look at Determiners in Early Grammar: Phrasal Quantifiers. *Language Research*, 48(3):573–608, 2012.

Anna L Theakston and Caroline F Rowland. The acquisition of auxiliary syntax: A longitudinal elicitation study. part 1: Auxiliary be. *Journal of Speech, Language, and Hearing Research*, 52(6):1449–1470, 2009.

Anna L Theakston, Paul Ibbotson, Daniel Freudenthal, Elena VM Lieven, and Michael Tomasello. Productivity of noun slots in verb frames. *Cognitive Science*, 39(6):1369–1395, 2015.

Michael Tomasello. What kind of evidence could refute the UG hypothesis? *Studies in Language*, 28(3):642–645, 2004.

Michael Tomasello and Silke Brandt. Flexibility in the semantics and syntax of children's early verb use. *Monographs of the Society for Research in Child Development*, 74(2): 113–126, 2009.

Michael Tomasello and William E Merriman. *Beyond names for things: Young children's acquisition of verbs*. Psychology Press, 1995.

Virginia Valian. Syntactic categories in the speech of young children. *Developmental Psychology*, 22(4):562, 1986.

Charles Yang. Who's Afraid of George Kingsley Zipf. Unpublished Manuscript, 2010.

Charles Yang. A statistical test for grammar. In *Proceedings of the 2nd workshop on Cognitive Modeling and Computational Linguistics*, pages 30–38. Association for Computational Linguistics, 2011.

Supplementary material for Bates, Pearl, & Braunwald 2018

A Calculation of Expected overlap for the different representation types

Below we provide a more detailed walk-through of the calculation of the Expected lexical overlap for the three representational types: NOT, SEMI, and fully PRODuctive.

For the NOT productive representation (e.g., *don't go* → *don't+go*), the speaker generates her multi-word combinations as memorized amalgams from her input. Using this representation, she will produce a given amalgam with about the same frequency she heard it in her input. To simulate this process, we generate multi-word combination data samples $S_{Exp_{Not}}$ that are the same size as the observed speaker multi-word combination sample S_{obs}; these samples are drawn from the speaker's input. That is, if there are 100 auxiliary+verb combinations in the speaker's output, we generate 100 auxiliary+verb combinations, based on the auxiliary+verb distribution in the speaker's input. This is shown in the top portion of equation (6).

The combinations that specific word w_{unk} from the category whose status is $Unknown$ is generated with depend on the combinations from the speaker's input that w_{unk} appeared with. To continue with our auxiliary example from above, if aux_j appeared with verb vb_k for 10% of the speaker's input, about 10% of the generated auxiliary+verb combinations $S_{Exp_{Not}}$ will be $aux_j vb_k$ combinations. That is, the probability of sample s_i involving word w_{unk} combined with w_{comb} depends on how often w_{unk}+w_{comb} appeared in the speaker's auxiliary+verb input. Once the sample using the NOT productive representation has been generated, we can calculate the lexical overlap for this sample and use that as the Expected overlap for a child using the NOT productive representation. This is shown in the bottom part of (6).

$$\begin{aligned} |S_{Exp_{Not}}| &= |S_{obs}| \\ s_i &\in S_{Exp_{Not}} \\ s_i &= w_{unk}w_{comb} \;\propto p_{w_{unk}w_{comb_{Input}}} \\ w_{unk} &\in Unknown,\; w_{comb} \in Combine \\ Expected_{Not} &= overlap_{Combine}(S_{Exp_{Not}}) \end{aligned} \tag{6}$$

Because we are generating samples of data produced by a child using the NOT productive representation, we repeat this process 1000 times (i.e., generate 1000 expected multi-word combination samples and calculate the Expected overlap). We then average these expected overlap scores to get the Expected overlap for the NOT productive representation.

We can use a similar approach when calculating the Expected overlap for the SEMI-productive representation (e.g., *don't go* → AUX+*go* or *don't*+VERB). Using this representation, a speaker generates her multi-word combinations by relying on her internal category representation for one word and her input distributions for combinations with the other word. More specifically, let's consider the case where the word from $Unknown$, w_{unk}, comes from a category while the word from $Combine$, w_{comb}, doesn't. To generate combination $w_{unk}w_{comb}$, the child relies on her internal category representation to generate word w_{unk} and then looks to her input to see how often words from this category combine with word w_{comb}. So, she would generate combination $w_{unk}w_{comb}$ with about the same frequency she heard examples of $Unknown$ w_{comb} in her input. To simulate this process, we again can generate multi-word combination data samples $S_{Exp_{Semi}}$ that are the same size as the observed speaker multi-word combination sample S_{obs}. Because this representation assumes that all words $w_{unk} \in Unknown$ in the speaker's output were generated from her internal category, they will appear as often as they appeared in her observed output. For example, if AUX is $Unknown$ and auxiliary $aux_j \in$ AUX appears 10 out of 100 times in the speaker's output, the generated sample will include a combination with aux_j about $\frac{10}{100} * 100 = 10\%$ of the time. In particular, category $Unknown$ generates words w_{unk} with some probability, and this is the probability we see these words in the speaker's output. So, the SEMI expected samples involve word w_{unk} proportional to how often they appeared in the speaker's observed productions. This is shown in the top part of (7).

The combinations w_{unk} is generated with depend on the combinations from the speaker's input that words of category $Unknown$ appeared with. Returning to our auxiliary example from before, if AUX is being assessed in combination with individual verbs, and auxiliaries appear with verb vb_j 5 out of 100 times, the generated sample will include auxiliaries in combination with vb_j 5% of the time. That is, the probability of multi-word sample $s_i \in S_{Exp_{Semi}}$ involving a specific word $w_{unk} \in Unknown$ combined with w_{comb} depends on how often any word in $Unknown$ combines with w_{comb} in the speaker's input. This is equivalent to how often w_{comb} appeared in the multi-word combinations involving words of category $Unknown$ in the speaker's input $w_{comb_{input}}$, as shown in (7).

$$\begin{aligned} |S_{Exp_{Semi}}| &= |S_{Obs}| \\ s_i &\in S_{Exp_{Semi}} \\ w_{unk} &\in s_i \propto p_{w_{unk_{Obs}}} \\ s_i &= w_{unk}w_{comb} \ \propto p_{Unknown}\ p_{w_{comb_{Input}}} \\ w_{unk} &\in Unknown,\ w_{comb} \in Combine \\ Expected_{Semi} &= overlap_{Combine}(S_{Exp_{Semi}}) \end{aligned} \tag{7}$$

A similar process can be used when $Unknown$ isn't a category while $Combine$ is. To generate combination $w_{unk}w_{comb}$, the child relies on her internal category representation to generate word w_{comb} and then looks to her input to see how often words from this category combine with word w_{unk}. So, she would generate combination $w_{unk}w_{comb}$ with about the same frequency she heard examples of w_{unk} $Combine$ in her input. To simulate this process, we again can generate multi-word combination data samples $S_{Exp_{Semi}}$ that are the same size as the observed speaker multi-word combination sample S_{obs}. Because this representation assumes that all words $w_{comb} \in Combine$ in the speaker's output were generated from her internal category, they will appear as often as they appeared in her observed output. For example, if VERB is $Combine$ and verb $v_j \in$ VERB appears 30 out of 100 times in the speaker's output, the generated sample will include a combination with v_j about $\frac{30}{100} * 100 = 30\%$ of the time. In particular, category $Combine$ generates words w_{comb} with some probability, and this is the probability we see these words in the speaker's output. So, the SEMI expected samples involve word w_{comb} proportional to how often they appeared in the speaker's observed productions. This is shown in the top part of (8).

The combinations w_{comb} is generated with depend on the combinations from the speaker's input that words of category $Combine$ appeared with. Returning to our verb example from before, if VERB is being assessed in combination with individual auxiliaries, and verbs appear with auxiliary aux_j 2 out of 100 times, the generated sample will include verbs in combination with aux_j 2% of the time. That is, the probability of multi-word sample

$s_i \in S_{Exp_{Semi}}$ involving a specific word $w_{comb} \in Combined$ combined with w_{unk} depends on how often any word in $Combine$ combines with w_{unk} in the speaker's input. This is equivalent to how often w_{unk} appeared in the multi-word combinations involving words of category $Combine$ in the speaker's input $w_{unk_{input}}$, as shown in (8).

$$\begin{gathered} |S_{Exp_{Semi}}| = |S_{Obs}| \\ s_i \in S_{Exp_{Semi}} \\ w_{comb} \in s_i \propto p_{w_{comb_{Obs}}} \\ s_i = w_{unk}w_{comb} \ \propto p_{w_{unk_{Input}}} \, p_{Combine} \\ w_{unk} \in Unknown, \ w_{comb} \in Combine \\ Expected_{Semi} = overlap_{Combine}(S_{Exp_{Semi}}) \end{gathered} \tag{8}$$

As before, once the sample using the SEMI representation has been generated, we can calculate the lexical overlap for this sample and use that for a child using a SEMI representation. This is shown in the bottom part of (7) and (8). We then do this process 1000 times to get 1000 SEMI samples, compute the lexical overlap for each, and take the average as the Expected SEMI overlap score.

For the fully PRODuctive representation (e.g., *don't go* → AUX+VERB), the speaker generates her multi-word combinations by relying on internal category representations for both words. Yang (2010, 2011) describes an analytical solution for the expected lexical overlap when both categories exist (shown in (9)). We can use this here, rather than generating expected samples and calculating lexical overlap for those samples. The key intuition involves the definition of lexical overlap, where a word w_{comb} shows lexical overlap if more than one word $w_{unk} \in Unknown$ combines with w_{comb}. So, we can calculate this analytically as 1 minus the probability that w_{comb} will (i) never appear with any word in $Unknown$, or (ii) only appear with a single word in $Unknown$.

For w_{comb} to never appear with any word in $Unknown$, this means that for all multi-word combination samples S_{obs} involving words from $Unknown$, w_{comb} was never selected. The probability of w_{comb} can be represented as $p_{w_{comb}}$, and so the probability of not choosing w_{comb} to combine with a word from $Unknown$ S_{obs} times is (1-$p_{w_{comb}}$)$^{S_{obs}}$.

For w_{comb} to appear with only a single word w_{unk} in $Unknown$, this means that for every multi-word combination UNKNOWN+COMBINE, either w_{comb} was selected and combined with w_{unk} (which occurs with probability $p_{w_{comb}} * p_{w_{unk}}$) or some other word – and not w_{comb} – was selected (which occurs with probability 1-$p_{w_{comb}}$). Any given sample with w_{comb} only ever appearing with w_{unk} will have some split between these two options, for all S_{obs} samples (i.e., k samples will have w_{comb} with w_{unk} and $S_{obs} - k$ samples will have some other $Combine$ word). So, if we sum up all these possibilities (shown in (9)), this is the probability of w_{comb} only ever appearing with a single w_{unk}.

Some algebraic rearrangement yields the formula at the bottom of (9) for the Expected overlap for word w_{comb} from Yang (2010; 2011). Note that all word probabilities are estimated based on the speaker's productions of w_{comb} and w_{unk} (i.e., $p_{w_{comb}} = p_{w_{comb_{Obs}}}$, $p_{w_{unk}} = p_{w_{unk_{Obs}}}$). This is because all words in these combinations are generated from an underlying internal category, and so don't rely on the speaker's input. As with the original calculation of lexical overlap, these individual word overlaps are averaged to get the Expected overlap.

$$\begin{aligned}
overlapprod_{w_{comb}} &= \\
&= 1 - P(no\ w_{comb}) - P(only\ 1\ w_{comb}) \\
&= 1 - (1 - p_{w_{comb}})^{S_{obs}} - \sum_{k=1}^{S_{obs}} \binom{S_{obs}}{k} (p_{w_{comb}} * p_{w_{unk}})^k (1 - p_{w_{comb}})^{S_{obs}-k} \\
&= 1 + (|Unknown| - 1)(1 - p_{w_{comb}})^{S_{obs}} \\
&- \sum_{w_{unk} \in Unknown} (p_{w_{comb}} * p_{w_{unk}} + 1 - p_{w_{comb}})^{S_{obs}} \\
& p_{w_{unk}} = p_{w_{unk_{Obs}}}, p_{w_{comb}} = p_{w_{comb_{Obs}}} \\
Expected_{Prod} &= \frac{\sum_{w_{comb} \in Combine} overlapprod_{w_{comb}}}{|Combine|}
\end{aligned} \tag{9}$$

B Evaluating potential representations

For any multi-word combination, we calculate lexical overlap scores with respect to the *Unknown* category, and either word in the multi-word combination may be the *Unknown* category being assessed. For example, in VERB+NOUN multi-word combinations, either the VERB or the NOUN could be assessed as the *Unknown* category while the other category serves as the collection of words in *Combine* (i.e., *Unknown*=NOUN and *Combine*=VERB as shown in the first row of Table 3, or vice versa, as shown in the second row of Table 3). The Observed and Expected lexical overlap scores are calculated based on which set of words is *Unknown* and which set is *Combine*.

Importantly, the Expected calculation depends on the status of *Unknown* and *Combine*, and this is determined by the category representation itself. For example, in the completely NOT productive representation in the top half of Table 3, all the Expected calculations are done using $Expected_{Not}$. In contrast, the Expected calculation for fully PRODuctive combinations where both words come from categories would be calculated using $Expected_{Prod}$. An example of this shown in the bottom half of Table 3 for $\text{Rep}_{\text{SEMI}_7}$ when assessing multi-word combinations with VERB and NEG. Because both of these are categories according to $\text{Rep}_{\text{SEMI}_7}$, $Expected_{Prod}$ is used. Likewise, the Expected calculation for SEMI-productive combinations where one word is a category while one word isn't would be calculated using $Expected_{Semi}$ (e.g., the *Unknown*=VERB, *Combine*=NOUN combination for SEMI-productive representation $\text{Rep}_{\text{SEMI}_7}$ in Table 3).

Table 3: Lexical overlap scores for multi-word combinations involving words from the *Unknown* category and the words that combine with it (*Combine*). The Observed overlap is calculated based on the speaker's productions while the Expected overlap is calculated based on the representation for that multi-word combination (Exp Calc), which depends on the category representation being used. Example scores are shown for the completely NOT productive representation where there are no categories and for a SEMI-productive representation that has categories for VERB and NEG only.

<table>
<tr><th colspan="5">Rep$_{\text{NOT}}$ = (<i>No Categories</i>)</th></tr>
<tr><th><i>Unknown</i></th><th><i>Combine</i></th><th>Exp Calc</th><th><i>Expected</i></th><th><i>Observed</i></th></tr>
<tr><td>Noun</td><td>Verb</td><td>$Expected_{Not}$</td><td>0.79</td><td>0.83</td></tr>
<tr><td>Verb</td><td>Noun</td><td>$Expected_{Not}$</td><td>0.19</td><td>0.27</td></tr>
<tr><td>Auxiliary</td><td>Verb</td><td>$Expected_{Not}$</td><td>0.65</td><td>0.62</td></tr>
<tr><td>Verb</td><td>Auxiliary</td><td>$Expected_{Not}$</td><td>0.73</td><td>0.67</td></tr>
<tr><td>Negation</td><td>Verb</td><td>$Expected_{Not}$</td><td>0.63</td><td>0.70</td></tr>
<tr><td>Verb</td><td>Negation</td><td>$Expected_{Not}$</td><td>0.63</td><td>0.83</td></tr>
<tr><th colspan="5">Rep$_{\text{SEMI}_7}$= (VERB, NEG)</th></tr>
<tr><th><i>Unknown</i></th><th><i>Combine</i></th><th>Exp Calc</th><th><i>Expected</i></th><th><i>Observed</i></th></tr>
<tr><td>Noun</td><td>Verb</td><td>$Expected_{Semi}$</td><td>0.87</td><td>0.83</td></tr>
<tr><td>Verb</td><td>Noun</td><td>$Expected_{Semi}$</td><td>0.53</td><td>0.27</td></tr>
<tr><td>Auxiliary</td><td>Verb</td><td>$Expected_{Semi}$</td><td>0.71</td><td>0.62</td></tr>
<tr><td>Verb</td><td>Auxiliary</td><td>$Expected_{Semi}$</td><td>0.79</td><td>0.67</td></tr>
<tr><td>Negation</td><td>Verb</td><td>$Expected_{Prod}$</td><td>0.64</td><td>0.70</td></tr>
<tr><td>Verb</td><td>Negation</td><td>$Expected_{Prod}$</td><td>0.92</td><td>0.83</td></tr>
</table>