BUCLD 42 Proceedings
To be published in 2018 by Cascadilla Press
Rights forms signed by all authors

Little lexical learners: Quantitatively assessing the development of adjective ordering preferences

Galia Bar-Sever, Rachael Lee, Gregory Scontras, and Lisa Pearl*

1. Introduction

Adults have robust ordering preferences that determine the relative order of adjectives in multi-adjective strings: this is why *small grey kitten* and *nice round penny* are preferable to *grey small kitten* and *round nice penny*. Adults are reliably and robustly uncomfortable with the latter options, yet are typically unable to pinpoint why they have this reaction. Notably, these preferences surface for any multi-adjective string, even ones never before encountered: English adults would probably prefer *tiny green magical mouse-riding gnomes* to *mouse-riding magical green tiny gnomes*, even though it is unlikely they have encountered these adjectives strung together before. Even more remarkable than the robustness and productivity of these preferences in English is the fact that these ordering preferences surface in a variety of unrelated languages, both those with pre-nominal adjectives (like English) and those with post-nominal adjectives that follow the modified noun (for discussion, see Dixon 1982, Sproat and Shih 1991).

When it comes to the source of these preferences, there have been a number of hypotheses. The null hypothesis for adults would hold that they simply repeat back what they hear when forming multi-adjective strings, reflecting the item-level statistics of their input. However, an input frequency strategy is limited in its productivity (if you haven't heard it, you don't have a preference about it) and adults are not limited this way. Importantly, because of their productivity, these preferences appear to be based on abstract representations, rather than simply reflecting the positioning of specific adjectives in the input.

But how exactly do adults represent these ordering preferences? Prevalent approaches in linguistics advance the idea that adult adjective ordering is determined by abstract syntax, with adjectives grouped into lexical semantic classes that are hierarchically ordered (Dixon, 1982, Cinque, 1994). These lexical classes and their hierarchical ordering would then serve as primitives in the representation of

*Galia Bar-Sever, University of California, Irvine, gbarseve@uci.edu. Rachael Lee, University of California, Irvine, rachaejl@uci.edu. Gregory Scontras, University of California, Irvine, g.scontras@uci.edu. Lisa Pearl, University of California, Irvine. lpearl@uci.edu. Thanks to the audiences and organizers of BUCLD 2017 and CAMP 2017. Anything we got wrong isn't their fault.

the preferences. Yet why should these classes be ordered the way they are, and how do we handle adjectives that do not fit neatly into a clear lexical class?

Recently, Scontras, Degen and Goodman (2017) identified adjective subjectivity as a robust predictor of adult ordering preferences, with less subjective adjectives preferred closer to the modified noun; they advanced the hypothesis that ordering preferences—and the lexical class ordering observed cross-linguistically—derive from the perceived subjectivity of the adjectives. Thus, perceived subjectivity would serve as a primitive of the adult representation of adjective ordering.

Still, little is known about the development of these adjective-ordering preferences in children, other than that these preferences do in fact develop (Bever, 1970, Martin and Molfese, 1972, Hare and Otto, 1978). The current paper assesses when more abstract knowledge about adjective ordering emerges, how that knowledge gets represented, and whether the knowledge representation matches what we believe to be active in adults. To perform this assessment, we use corpus analysis and quantitative metrics to connect children's linguistic input, potential underlying representations regarding adjective ordering, and linguistic output, thereby arriving at a clearer picture of children's knowledge in this domain.

2. Previous accounts of adjective order

We start by reviewing relevant background for the competing hypotheses surrounding adult knowledge of adjective ordering. We then review behavioral studies aimed at understanding children's preferences.

2.1. The lexical class hypothesis

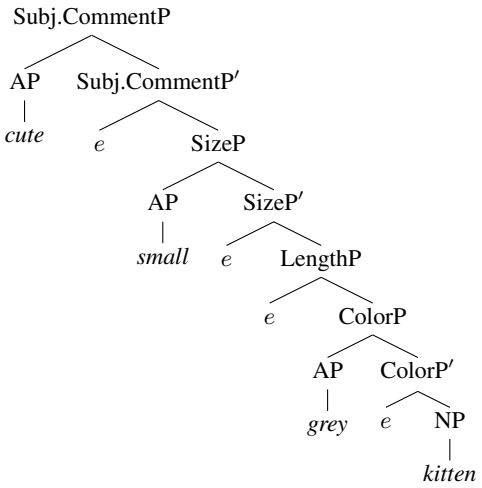
The lexical class hypothesis begins with the assumption that adjectives come pre-sorted into classes according to their semantic properties: COLOR adjectives group together, SIZE adjectives group together, etc. To account for adjective ordering, these classes correspond to a deterministic hierarchy that maps adjective strings to their linear order, as in (1); higher positioning in the hierarchy leads to greater distance from the modified noun.

(1) *Lexical semantic class hierarchy from Dixon (1982)*:
VALUE > DIMENSION > PHYSICAL PROPERTY > SPEED
> HUMAN PROPENSITY > AGE > COLOR

As proposed by Dixon (1982), these hierarchical lexical semantic classes form part of a speaker's internal grammar and the lexical classes themselves are universal, existing in all human languages regardless of differences in the morpho-syntactic expression of these semantic types. In an attempt to formalize the linear ordering of these lexical semantic class hierarchies, Cinque (1994) proposed a fully syntactic account of ordering preferences whereby the individual classes project their own phrasal structure, with one phrase hierarchically dominating an-

other. Under a syntactic account, in *small grey kitten* the COLOR adjective appears closer to the noun than the SIZE adjective because the adjective phrase projected by *small* hierarchically dominates the adjective phrase projected by *grey*. This hierarchical ordering gets expressed as the linear order of adjectives modifying a noun. The proposal has been elaborated on since its initial formulation, with recent authors proposing even richer structure, as in (2) (see also Laenzlinger 2005).

(2) *Phrase structure proposed by Scott (2002)*



Throughout this work on lexical classes, authors have disagreed about the precise specification of the classes themselves. Dixon's classes in (1) gave way to Cinque's (POSSESSIVE, SPEAKER-ORIENTED, SUBJECT-ORIENTED, MANNER/THEMATIC), which depart from the classes proposed by Sproat and Shih (1991) (QUALITY, SIZE, SHAPE, COLOR, PROVENANCE). Worse, authors often disagree about the fine details of the class orderings. Still, despite the fact that it is hard to settle on *the* universal adjective classes, it has been shown that a certain ordering of adjective classes goes some way in explaining the patterns observed. What this collection of research does not address is where the hierarchy comes from in the first place: supposing SIZE adjectives do syntactically dominate COLOR adjectives, why should this be the case and not the reverse? This approach also relies on an ability to identify the appropriate lexical semantic class for any given adjective, enforcing a sorting into discrete bins based on a static meaning. What about adjectives that fail to fall into a semantic class?

2.2. The subjectivity hypothesis

In an attempt to address the concerns facing the lexical class hypothesis head-on, recent work by Scontras et al. (2017) advances the hypothesis that aspects of an adjective's meaning determine its relative position in a multi-adjective string.

Scontras et al. propose that the perceived subjectivity of the property an adjective names influences its ordering. This subjectivity hypothesis states that less subjective adjectives are preferred closer to the modified noun than adjectives that are more subjective (see also Hetzron 1978, Hill 2012).

Scontras et al. operationalized subjectivity as the potential for faultless disagreement between two speakers about whether an adjective applies to some object (Barker, 2013, Kennedy, 2013, Kölbel, 2004). In a test of faultless disagreement, two speakers are presented with an object (say, a kitten); the speakers then disagree about whether the object holds some property (say, being small). To the extent that both speakers can be right while they disagree, the property (and the adjective that names it) admits that degree of faultless disagreement. In other words, an adjective's subjectivity is defined by how much disagreement speakers can have about that adjective without one of the speakers necessarily being wrong. An adjective like *small* admits a relatively high degree of faultless disagreement (two people can disagree about whether they consider an object small), and is therefore relatively subjective. In contrast, an adjective like *grey* is relatively objective: when two people disagree about whether something is grey, one of those people is likely to be wrong. Surprisingly, Scontras et al. found that participants' estimates of faultless disagreement matched their ratings for adjective "subjectivity" ($r^2 = .91, 95\%$ CI $[.86, .94]$). So, simply asking adults how "subjective" they believe an adjective to be (a metalinguistic task) can serve as a proxy for the potentially more ecologically valid faultless disagreement task.

To get a clearer picture of the English ordering preferences that need to be accounted for, Scontras et al. measured ordering preferences in a behavioral experiment; participants indicated the preferred ordering for multi-adjective strings (e.g., *small grey kitten* vs. *grey small kitten*). To ensure that the behavioral measure captured the implicit knowledge that speakers use when forming multi-adjective strings, Scontras et al. compared their measure against naturalistic multi-adjective strings from corpora. Finding a high correlation between the behavioral measure and corpus statistics ($r^2 = .83, 95\%$ CI $[.63, .90]$), Scontras et al. concluded that the preferences that were measured accurately capture speaker knowledge.

To test the subjectivity hypothesis, Scontras et al. used their estimates of adjective subjectivity to predict the preferred adjective orderings. They found that adjective subjectivity accounts for between 51% and 88% of the variance in the ordering preferences. In other words, subjectivity does predict adjective ordering, thus offering a cognitive explanation for a linguistic universal. Recent work by Hahn, Futrell and Degen (2017) shows that perceived subjectivity influences adjective ordering even in an artificial language paradigm with novel adjectives. This finding lends support to the idea that adults attend to subjectivity (and not simply to the lexical class statistics of their input) as they form multi-adjective strings. Given its promise in accounting for adult knowledge of adjective ordering, we might reasonably wonder about the source of this subjectivity-based knowledge.

2.3. The development of adjective ordering preferences

The cross-linguistic robustness of ordering preferences has led many researchers to conclude that the knowledge underlying these preferences is innate, pre-specified as part of the Universal Grammar that shapes human language. Part of the appeal of the subjectivity hypothesis is that it allows us to move away from claims of innateness (and the puzzle of genetically specifying linguistic structure). Instead, the subjectivity hypothesis favors an account where subjectivity awareness develops as we use language to communicate; after all, the potential for faultless disagreement is a problem all speakers must attend to. To better understand the role of subjectivity in ordering preferences and the pressures that lead to it, we must therefore ask whether this knowledge is present from the start, or whether it develops—perhaps in stages—into what we observe as the adult state.

There have been several studies examining adjective ordering preferences in children, but they have not been successful in answering this question. Still, the existing evidence at least suggests that the preferences do in fact develop.

Bever (1970) found that with children between two and five years of age, the younger children were more likely to repeat unnatural adjective orderings such as *the plastic large pencil*; older children corrected the phrase to *the large plastic pencil*. We might therefore conclude that the younger children fail to demonstrate stable adjective ordering preferences. However, Martin and Molfese (1972) attempted to recreate Bever's experiment but were unable to replicate his findings. This replication failure led Martin and Molfese to suggest that the original repetition task is not a reliable measure of adjective ordering preferences. In its place, they used a production task, finding that three- and four-year-olds produced phrases with adjectives denoting CLEANLINESS closer to the noun than COLOR adjectives (e.g., *yellow clean house*), while the adult preference is for COLOR adjectives to appear closer (i.e., *clean yellow house*). This result provides evidence that children's preferences differ from adult preferences, but only with respect to adjectives of CLEANLINESS and COLOR. A later study by Hare and Otto (1978) had children in grades one through five arrange three adjectives of SIZE, COLOR, and MATERIAL to create natural adjective phrases; children in each succeeding grade level chose the adult preferred order of SIZE–COLOR–MATERIAL (e.g., *little yellow rubber duck*) more often than children in the preceding grade level.

These developmental studies leave much unsettled, but they do suggest that adjective ordering preferences develop or strengthen over time. However, there is disagreement among these studies on the age of acquisition, and what the developmental trajectory looks like. Moreover, none of these studies attempt to tie children's knowledge to adjective subjectivity. If in fact the perceived subjectivity of adjectives is what adults are using to inform their adjective ordering preferences, we ought to wonder when children begin to deploy this strategy.

Notably, this question becomes more complicated in light of recent work showing that children may not have reliable estimates of subjectivity until around the age of seven or eight (Foushee and Srinivasan, 2017). If subjectivity is not

available but children still demonstrate clear ordering preferences, how are these preferences acquired from the input children receive and represented with their available cognitive resources? It may be possible (indeed, likely) that children evolve through various stages of knowledge representation for their adjective ordering preferences. To investigate this knowledge and its stages of development, we examine children's production of multi-adjective strings in light of the input they are receiving at different ages.

3. Quantitatively assessing representational hypotheses

3.1. Corpus data

To identify the representations underlying the development of adjective ordering preferences, we assess naturalistic child input in the form of child-directed speech and naturalistic child output in the form of child-produced speech; data come from the CHILDES database MacWhinney (2000). We focus on the morphologically annotated corpora in the North American datasets for children between the ages of two and four, yielding 688,428 child-directed and 1,069,406 child-produced utterances. After extracting all instances of adjective-adjective-noun (**AdjAdjN**) strings, we arrived at the counts in Table 1.

	Child-directed data			**Child-produced data**		
age	**# AdjAdjN**	**# tokens**	**# types**	**# AdjAdjN**	**# tokens**	**# types**
2	1440	2880	131	466	932	79
3	881	1762	128	274	584	72
4	745	1490	124	235	470	81

Table 1: Number of AdjAdjN strings and both the adjective tokens and adjective types comprising these strings per age in the morphologically-tagged North American CHILDES corpora.

3.2. Analysis of direct repetitions

We first asked whether children's AdjAdjN productions are simply direct repetitions of AdjAdjN strings they had heard. For example, is a child simply repeating *big bad wolf* because she just heard it in her input? If most child productions are due to direct repetition, it is unlikely that these productions reflect any sophisticated generative process on the part of the child that transforms the child's input into the child's output AdjAdjN strings. In other words, it is unlikely that the child possesses knowledge of adjective ordering preferences at all.

Fortunately, our analysis revealed that only 0.50% of all child-produced Adj AdjN strings in our corpus were direct repetitions of an AdjAdjN string heard

immediately prior from an adult.[1] The absence of repetitions in children's productions suggests that children are generating the AdjAdjN strings they produce based on some transformation of the input they hear. That is, they are internalizing some representation based on their input with AdjAdjN strings and using that representation to generate the AdjAdjN strings they produce as output. The question the becomes which representations best fit children's observed output at ages two, three, and four, based on their input at these ages.

3.3. The representational hypotheses

We consider three representational hypotheses that could underlie children's adjective ordering preferences. The first two correspond to the two potential adult representations discussed in Section 2: representations based on (i) hierarchically ordered adjective **lexical classes** and (ii) perceived **subjectivity** of adjectives. Both hypotheses require children to create some abstraction across individual adjective lexical items (i.e., in terms of lexical class or perceived subjectivity), and then order adjectives with respect to this abstraction. In contrast, the third representational hypothesis we consider is a simpler item-based approach, and does not require additional abstraction. This hypothesis states that children track the **input frequency** of adjectives appearing in certain positions in multi-adjective strings, and their productions mirror the frequencies observed in the input. In particular, for each adjective, children would pay attention to how often it appears in the ***1-away*** position closest to the noun vs. the ***2-away*** position farther from the noun (e.g., *small*$_{2-away}$ *grey*$_{1-away}$ *kitten*). This input frequency approach corresponds to the null hypothesis discussed in Section 1, and serves as one of the simplest approaches to adjective ordering preferences if young children are able to track the statistical distributions of adjectives in their input. This statistical learning ability seems cognitively plausible, given evidence from many areas of language development demonstrating very young children's statistical learning abilities (e.g., Saffran, Aslin and Newport 1996, Maye, Werker and Gerken 2002, Gerken 2006, Mintz 2006, Xu and Tenenbaum 2007, Maye, Weiss and Aslin 2008, Smith and Yu 2008, Dewar and Xu 2010, Feldman, Myers, White, Griffiths and Morgan 2013, Gerken and Knight 2015, Gerken and Quam 2017, among others).

3.4. Empirical grounding of the representational hypotheses

Each potential representation requires certain information to be known about an adjective: lexical class, perceived subjectivity, or positional input frequency. For lexical class, we first relied on the 13 lexical classes and adjective assignments reported in Scontras et al. (2017), which derived from a synthesis of previous literature (Dixon, 1982, Sproat and Shih, 1991); we inferred a hierarchical ordering

[1] Adults produced more direct repetitions in child-directed speech, although the amount was still minimal: 2.96% of child-directed AdjAdjN strings were direct repetitions.

of these classes on the basis of the behavioral data reported by Scontras et al. If an adjective had no lexical class entry in Scontras et al. (2017), we attempted to analogize it to an existing entry based on similar meaning (e.g., *teeny* is similar in meaning to *small* and so was assigned to the DIMENSION class). If there was no clear analogy to an existing entry (e.g., *ripe*), we manually assigned it to a lexical class via collective agreement by all four authors. Some of the adjectives (62.6% of adjective types but only 4.7% of adjective tokens) wound up in the X "elsewhere" class as defined in Scontras et al.; these adjectives did not neatly fit into any of the other class categories. Because the elsewhere class is so heterogeneous, its adjectives fail to cohere on the basis of meaning. As a result, this collection of adjectives does not stand as a lexical *semantic* class, so we excluded its adjectives from the representational analyses described below.

For perceived subjectivity, we obtained subjectivity scores from 108 adult participants on Amazon.com's Mechanical Turk crowdsourcing service, replicating the methodology of Scontras et al. (2017). Participants were presented with 30 adjectives total (one at a time) in a random order and asked to indicate how "subjective" a given adjective was on a sliding scale; endpoints were labeled "completely objective" and "completely subjective." To arrive at the perceived subjectivity score for a given adjective, responses were averaged across participants.

For positional input frequency, we derived both 1-away and 2-away frequencies from the child-directed speech AdjAdjN strings in our corpus by calculating how often an adjective appeared in the 1-away vs. 2-away position in the input.

3.5. Quantitatively linking input to output

Recall that producing an AdjAdjN string requires transforming the input according to the underlying knowledge representation, and using that representation to generate the AdjAdjN string. For each representational hypothesis, we can define how this process would occur, thereby linking the child-directed AdjAdjN input to child-produced AdjAdjN output. We focus on how a given representational hypothesis would generate an adjective in the 2-away vs. the 1-away position when combined with another adjective in an AdjAdjN string.

We consider the collection of child-produced AdjAdjN output at a particular age as a dataset D that is produced according to any of the three representational hypotheses $h_i \in H$, where $H = \{h_{lex}, h_{subj}, h_{freq}\}$. We select the hypothesis that is most likely to have generated the data in D (i.e., the child productions) by calculating the likelihood of a given hypothesis h generating the data D, $p(D|h)$. The representational hypothesis with the largest probability of generating D (i.e., the highest likelihood) is the hypothesis that best matches children's output.

We can conceive of D as the set of AdjAdjN strings involving different combinations of all the adjectives *Adj* observed in the corpus. For example, D might be the set {*grey furry kitten*, *small grey kitten*, *small grey kitten*, *small furry kitten*}, where *Adj* is {*grey*, *furry*, *small*}. To account for the portion of the AdjAdjN strings involving a particular adjective $adj_x \in Adj$, we can calculate the likeli-

hood of the data involving that adjective, $p(D_{adj_x}|h)$. Continuing the example from above, the *small* strings D_{small} would be the set {*small grey kitten*, *small grey kitten*, *small furry kitten*}; in this example, *small* occurs in the 2-away position with probability 1.0. The *grey* strings would form the set D_{grey}: {*grey furry kitten*, *small grey kitten*, *small grey kitten*}; here, *grey* occurs in the 2-away position with probability 0.33. Having calculated the likelihood of the data involving each individual adjective for a specific representational hypothesis h_i, we then multiply these individual adjective likelihoods to yield the likelihood for the whole dataset D under that hypothesis, as shown in equation (3).

$$p(D|h_i) = \prod_{adj_x \in Adj} p(D_{adj_x}|h_i) \tag{3}$$

We define the likelihood for an individual adjective adj_x for a given hypothesis h_i as in equation (4), which considers the number of times N that adj_x appeared in an AdjAdjN string in the output, the number of times t that adj_x appeared in the 2-away position, and the probability that adj_x would appear in the 2-away position given the representational hypothesis h_i, $p_2exp(adj_x|h_i)$.

$$p(D(adj_x)|h_i) = \binom{N}{t}(p_2exp(adj_x|h_i))^t(1 - p_2exp(adj_x|h_i))^{N-t} \tag{4}$$

To see how this equation works, consider D_{grey} from above: {*grey furry kitten*, *small grey kitten*, *small grey kitten*}. Suppose a given representational hypothesis h_i predicts that *grey* should appear in the 2-away position with a certain probability $p_2exp(adj_x|h_i)$. We compare this expected probability with the actual frequency of *grey* occurring in the 2-away position to calculate the likelihood of D_{grey} under h_i, $p(D(adj_x)|h_i)$; if the expected probability matches the actual frequency, the hypothesis does an excellent job of accounting for the child output.

To calculate the likelihood, we need to determine the number of ways of generating the pattern in D_{grey} (i.e., *grey* in the 2-away position twice and in the 1-away position once). This corresponds to $\binom{N}{t}$, the number of ways of generating N AdjAdjN strings with *grey* in the 2-away position t times. Having determined the number of ways to generate the observed pattern, we then calculate the probability of generating the observed pattern given a specific representational hypothesis h_i. We first need to calculate the probability that *grey* would appear in the 2-away position two times, $(p_2exp(adj_x|h_i))^t$. To capture the full pattern, we also need to calculate the probability that *grey* would appear in the 1-away position once, $(p_2exp(adj_x|h_i))^{N-t}$. By multiplying the probability of generating the observed pattern together with the number of ways we could have generated it, we arrive at the likelihood in equation (4).

The calculation of $p_2exp(adj_x|h_i)$, the probability that a particular adjective adj_x will appear in the 2-away position, depends on the hypothesis h_i under consideration, as well as the input the child has encountered via child-directed speech.

For both the lexical class hypothesis h_{lex} and the subjectivity hypothesis h_{subj}, the probability that adj_x surfaces in the 2-away position in an AdjAdjN string depends on the kind of adjective it appears with. For h_{lex}, if adj_x is combined with an adjective in a hierarchically-closer lexical semantic class, it should surface in the 2-away position 100% of the time ($p = 1.0$); if adj_x is combined with an adjective in the same lexical class, it should surface in the 2-away position with chance probability ($p = 0.5$). For h_{subj}, if adj_x is combined with an adjective perceived as more subjective, it should surface in the 2-away position 100% of the time ($p = 1.0$); if adj_x is combined with an adjective perceived as equally subjective, it should surface in the 2-away position with chance probability ($p = 0.5$).[2] These considerations represent the numerator in equation (5).

$$p_2exp(adj_x|h_i \in \{h_{lex}, h_{subj}\}) = \frac{f_{input}(< adj_x|h_i) + 0.5 * f_{input}(= adj_x|h_i) + \alpha}{N_{input}(Adj) + \alpha * |Adj|} \quad (5)$$

In particular, $f_{input}(< adj_x|h_i)$ represents the number of adjective tokens in the input (i.e., the child-directed speech) that are either from a lexically-closer class than adj_x (given h_{lex}) or are more subjective than adj_x (given h_{subj}); the greater this number, the more we would expect the child to produce adj_x in the 2-away position under the relevant hypothesis. Similarly, $f_{input}(= adj_x|h_i)$ represents the number of adjectives that are from the same lexical class as adj_x (h_{lex}) or are equally subjective as adj_x (h_{subj}); this number gets multiplied by 0.5 to represent the chance probability that adj_x would appear 2-away with adjectives of the same kind. We arrive at the probability of adj_x appearing in 2-away position once we divide these counts by the total number of adjective tokens appearing in AdjAdjN strings in the input, $N_{input}(Adj)$. Both the numerator and the denominator of equation (5) contain the smoothing factor $\alpha = 0.5$, which is added to handle adjectives for which there are no observations; in the denominator, α is multiplied by the number of adjective types $|Adj|$.[3]

A different calculation is used for p_2exp for the input frequency representational hypothesis h_{freq}, as shown in equation (6). The probability of adj_x appearing in the 2-away position given h_{freq} is a simple reflection of how often it appeared in the 2-away position in the input ($f_{2input}(adj_x)$) divided by the total number AdjAdjN strings in which adj_x appeared at all ($N_{input}(adj_x)$). Again, we add the smoothing factor α to avoid assigning zero probability for adjectives not

[2]We considered two adjectives to be perceived as equally subjective if their subjectivity scores were within 0.1 of each other; scores ranged from 0 to 1.

[3]To implement the idea that the target adjective adj_x cannot combine with tokens of itself (e.g., *small small kitten*), the number of adj_x tokens is subtracted from the counts of how many adjectives either are in the same lexical class or have the same subjectivity score in the numerator; this number is also subtracted from the total adjective token count in the denominator.

observed; in the denominator, α gets multiplied by 2, corresponding to the two positional options for adj_x: 2-away vs. 1-away.

$$p_2 exp(adj_x | h_i = h_{freq}) = \frac{f_{2input}(adj_x) + \alpha}{N_{input}(adj_x) + 2 * \alpha} \quad (6)$$

Using equations (3)-(6), we can evaluate how probable it is that children would have produced the AdjAdjN strings in the child-produced output for a certain age, given the input they heard during that age and a particular representational hypothesis: lexical class, subjectivity, and input frequency.[4]

4. Results

Table 2 reports log likelihood scores for each representational hypothesis at each age. Because likelihood probabilities of an entire set of AdjAdjN output strings can be very small, we take the log of the likelihood probabilities for easier comparison. Note that because these numbers are logged, the smallest negative number for each age corresponds to the most probable representational hypothesis for that age.

	representational hypotheses		
age	**lexical class**	**subjectivity**	**input frequency**
2	-145.4	-119.3	**-88.0**
3	-71.2	-70.8	**-54.3**
4	**-71.8**	-84.0	-79.4

Table 2: Log likelihood scores for each hypothesis. Scores range from 0 (best possible) to negative infinity (worst possible). The best score for each age is **bolded**. Each age should be looked at individually (i.e., numbers should only be compared across a row), since the datasets between ages differ.

We find that the input frequency hypothesis—in other words, simply tracking the word-level position statistics—best accounts for children's productions at age two; the subjectivity hypothesis performs less well, and the lexical class hypothesis performs worst of all. At age three, the input frequency hypothesis continues to outperform the other two hypotheses, but at this age the lexical class and subjectivity hypotheses seem to be performing on a par. It is only at age four that we see one of the abstract representational hypotheses best accounting for children's productions: the lexical class hypothesis.

[4]We only included AdjAdjN strings in both the input and output sets where both adjectives in the string had been assigned a lexical class and a subjectivity score. This excluded 870 AdjAdjN strings in the child-directed input (23.6%) and 178 AdjAdjN strings in the child-produced output (15.4%).

In addition to diagnosing the most likely hypothesis, we can use these probabilities to examine the emergence of abstract knowledge representations. In particular, we can compare hypothesis probabilities against each other to determine how much less probable one hypothesis is than another at accounting for child productions at a particular age. To perform this comparison on logged probabilities, we subtract them; the smaller the difference in log space, the closer the probabilities are in unlogged space. For our purposes, this method allows us to diagnose how close a worse-performing representational hypothesis is to the best-performing representational hypothesis at accounting for child-produced data. Table 3 reports these differences. Recall that at ages two and three, the best-performing hypothesis is the item-based input frequency representation, while the best-performing hypothesis at age four is the abstract lexical class hypothesis.

	representational hypotheses	
age	**lexical class**	**subjectivity**
2	-57.4	-31
3	-16.9	-16.5
4	0	-7.6

Table 3: How much less probable the abstract representational hypotheses are for the child productions at each age, compared against the best-performing hypothesis, and reported as log scores. Scores range from 0 (equally likely because it *is* the most probable hypothesis) to negative infinity (infinitely less likely).

We see that both abstract representational hypotheses perform better as children age, signaling the emergence of abstract knowledge about adjective ordering preferences. From age two to three, the lexical class hypothesis closes in on the lexical class hypothesis, eventually overtaking it by age four. We see a similar trajectory for the subjectivity hypothesis, although it never overtakes the best-performing hypothesis in our data, which ends at age four.

5. Discussion

Our quantitative assessment of the development of adjective ordering preferences demonstrates that abstract knowledge is likely to underlie children's preferences at age four (but not earlier), and that this abstract knowledge is lexical-class-based rather than subjectivity-based. Children initially track the word-level statistics of their input when determining adjective ordering, but by age four they shift to a more abstract (and compact) representation based on lexical semantic class. Given this developmental trajectory, it would appear that lexical semantic classes are a useful (and salient) tool for children as they move from simply tracking the statistics of their input to systematically organizing that knowledge.

To better understand the knowledge children are compressing into lexical-class-based representations, future computational work should examine the rep-

resentations adults are using to form the input children receive. Adults are known to adjust the complexity of their child-directed speech based on the child's age (e.g., Kunert, Fernández and Zuidema 2011), and so it may be that the representations underlying child-directed adjective orderings vary depending on the age of the child being addressed. If adults are providing children with input of a fundamentally different character from what they are providing other adults—for example, by hyperarticulating positional differences between adjectives—we ought to understand the pressures that lead to that divergence.

It remains unclear when subjectivity replaces lexical classes as the underlying representation for adjective ordering preferences—this timing no doubt depends on children's development of the conceptual underpinnings of subjectivity, which occurs remarkably late (Foushee and Srinivasan, 2017). Future behavioral work can assess children's perceived subjectivity of adjectives at different ages. The subjectivity scores used in our assessment derived from adult judgments, but children's estimates are likely to differ, given the sophisticated theory of mind involved in evaluating subjectivity. Whether these differences would better capture children's productive knowledge of adjective ordering remains an open question.

References

BARKER, CHRIS. 2013. Negotiating Taste. *Inquiry* 56(2-3).240–257.

BEVER, THOMAS G. 1970. The cognitive basis for linguistic structures. *Cognition and the development of language* 279(362).1–61.

CINQUE, GUGLIELMO. 1994. On the evidence for partial N-movement in the Romance DP. *Paths Towards Universal Grammar. Studies in Honor of Richard S. Kayne*, ed. by R S Kayne, G Cinque, J Koster, J.-Y. Pollock, Luigi Rizzi and R Zanuttini, 85–110. Washington DC: Georgetown University Press.

DEWAR, KATHRYN M; and FEI XU. 2010. Induction, overhypothesis, and the origin of abstract knowledge: Evidence from 9-month-old infants. *Psychological Science* 21(12).1871–1877.

DIXON, ROBERT MW. 1982. *Where have all the adjectives gone? and other essays in semantics and syntax*, vol. 107. Walter de Gruyter.

FELDMAN, NAOMI H; EMILY B MYERS; KATHERINE S WHITE; THOMAS L GRIFFITHS; and JAMES L MORGAN. 2013. Word-level information influences phonetic learning in adults and infants. *Cognition* 127(3).427–438.

FOUSHEE, RUTHE; and MAHESH SRINIVASAN. 2017. Could both be right? Children's and adults' sensitivity to subjectivity in language. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 379–3384. London, UK: Cognitive Science Society.

GERKEN, LOUANN. 2006. Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition* 98(3).B67–B74.

GERKEN, LOUANN; and SARA KNIGHT. 2015. Infants generalize from just (the right) four words. *Cognition* 143.187–192.

GERKEN, LOUANN; and CAROLYN QUAM. 2017. Infant learning is influenced by local spurious generalizations. *Developmental Science* 20(3).

HAHN, MICHAEL; RICHARD FUTRELL; and JUDITH DEGEN. 2017. Exploring adjective ordering preferences via artificial language learning. California Meeting on Psycholinguistics.

HARE, VICTORIA CHOU; and WAYNE OTTO. 1978. Development of preferred adjective ordering in children, grades one through five. *The Journal of Educational Research* 190–193.

HETZRON, ROBERT. 1978. On the relative order of adjectives. *Language universals*, ed. by H Seiler, 165–184. Tübingen.

HILL, FELIX. 2012. Beauty before age?: applying subjectivity to automatic english adjective ordering. *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies: Student research workshop*, 11–16. Association for Computational Linguistics.

KENNEDY, CHRISTOPHER. 2013. Two Sources of Subjectivity: Qualitative Assessment and Dimensional Uncertainty. *Inquiry* 56(2-3).258–277.

KÖLBEL, MAX. 2004. Faultless Disagreement. *Proceedings of the Aristotelian Society* 104.53–73.

KUNERT, RICHARD; RAQUEL FERNÁNDEZ; and WILLEM ZUIDEMA. 2011. Adaptation in child directed speech: Evidence from corpora. *Proc. SemDial* 112119.

LAENZLINGER, CHRISTOPHER. 2005. French adjective ordering: perspectives on DP-internal movement types. *Lingua* 115.645–689.

MACWHINNEY, BRIAN. 2000. *The CHILDES project: The database*, vol. 2. Psychology Press.

MARTIN, JAMES E; and DENNIS L MOLFESE. 1972. Preferred adjective ordering in very young children. *Journal of Verbal Learning and Verbal Behavior* 11(3).287–292.

MAYE, JESSICA; DANIEL J WEISS; and RICHARD N ASLIN. 2008. Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental science* 11(1).122–134.

MAYE, JESSICA; JANET F WERKER; and LOUANN GERKEN. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82(3).B101–B111.

MINTZ, TOBEN H. 2006. Finding the verbs: Distributional cues to categories available to young learners. *Action meets word: How children learn verbs* 31–63.

SAFFRAN, JENNY R; RICHARD N ASLIN; and ELISSA L NEWPORT. 1996. Statistical learning by 8-month-old infants. *Science* 1926–1928.

SCONTRAS, GREGORY; JUDITH DEGEN; and NOAH D GOODMAN. 2017. Subjectivity predicts adjective ordering preferences. *Open Mind: Discoveries in Cognitive Science* 1.53–65.

SCOTT, GARY-JOHN. 2002. Stacked adjectival modification and the structure of nominal phrases. *The cartography of syntactic structures, volume 1: Functional structure in the dp and ip*, ed. by G Cinque, 91–120. Oxford: Oxford University Press.

SMITH, LINDA; and CHEN YU. 2008. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition* 106(3).1558–1568.

SPROAT, RICHARD; and CHILIN SHIH. 1991. The cross-linguistic distribution of adjective ordering restrictions. *Interdisciplinary approaches to language*, 565–593. Springer.

XU, FEI; and JOSHUA B TENENBAUM. 2007. Word learning as Bayesian inference. *Psychological Review* 114(2).245.