How adjective ordering preferences develop in children

Rachael Lee
June 2018
Department of Cognitive Sciences
University of California, Irvine

Gregory Scontras
Department of Language Science

Lisa Pearl
Department of Language Science

Abstract

In English, it sounds much better to say "the small grey kitten" than "the grey small kitten." This is because adult English speakers have robust preferences for ordering adjectives. Recent research by Scontras, Degen, and Goodman (2017) found that these preferences are predicted by the speaker's understanding of adjective subjectivity, with less subjective adjectives tending to be closer to the noun. Past research has been inconclusive in determining the development of these preferences in children, or whether/when children use subjectivity for adjective ordering. This paper compares the likelihood of three hypotheses for representing children's knowledge of adjective ordering: (i) word-level input frequency, (ii) adjective lexical semantic class, and (iii) adjective subjectivity. To analyze the hypotheses, multi-adjective strings were extracted from 2- to 4-year-old English data from the CHILDES database, and adjectives were placed into a lexical class and given a crowd-sourced subjectivity score. The likelihood of the child-produced data by age was calculated under each hypothesis, given the child-directed data. Input frequency was found to be the best performing hypothesis at ages 2 and 3, with lexical class performance overtaking input frequency at age 4, and the subjectivity hypothesis never outperforming the others at these ages. These findings suggest that abstract knowledge does not emerge until age 4.

1 Introduction

In English, it sounds much better to say "the small grey kitten" than "the grey small kitten." Switching the order of these two adjectives completely changes the naturalness of the phrase. We would expect both phrases to communicate the same information, but English speakers seem to have intuitions that the second one is deviant while the other is natural. Adjective ordering preferences like these have received much attention in the language science research community, as it appears there are robust preferences for ordering adjectives not only in English, but also in many other languages around the world, such as Hungarian, Telugu, Mandarin Chinese, Dutch, Selepet, and Mokilese (Scontras, Degen, and Goodman, 2017).

For a while, it was unclear what was driving those robust preferences until a recent study by Scontras, Degen, and Goodman (2017) found that adjective ordering preferences are based on speakers' understanding of adjective subjectivity. What they found was that an adjective's distance from a noun could be predicted by the subjectivity of that adjective as perceived by adult speakers, with less subjective adjectives tending to be closer to the noun. For example, two people can faultlessly disagree about whether something is small more easily than whether something is grey, making *small* more subjective than *grey*. Thus, *grey* would be closer to the noun than *small* would be, resulting in "the small grey kitten" sounding more natural than "the grey small kitten."

However, because Scontras et al. only looked at adjective ordering preferences in adult speech, the development of children's adjective ordering preferences was left yet to be understood. While a few studies (Bever, 1970; Hare and Otto, 1978; Martin and Molfese, 1972) have attempted to look at how adjective ordering preferences develop in children, they were largely inconclusive aside from finding

that younger children tend to produce adjective orders that are less adult-like, and, as children get older, adult-like preferences tend to be seen more often. In other words, the results of these studies suggest that ordering preferences do develop in children. However, these studies did not examine the types of knowledge children were using to represent their ordering preferences, relying only on lexical semantic class for their analysis and never delving into the possibility of subjectivity-based knowledge. We do know that there is some kind of developmental trajectory for adjective ordering, but when or how these adult-like preferences develop has yet to be determined.

Several different hypotheses for adjective ordering preferences have been proposed, but it is not clear which theory best models children's speech patterns. This paper investigates (i) if/when stable preferences emerge in children, (ii) what the time course of this development is (i.e., whether preferences are present very early in children or if they instead develop over a longer period of time), and (iii) which hypothesis represents the adjective ordering knowledge children possess. In this paper, we look at three possible hypotheses: input frequency (the null hypothesis), lexical semantic class, and perceived subjectivity. The null hypothesis would hold that children are simply repeating back phrases and adjective orderings heard from their input. However, this would limit the amount of novel adjective phrases possible to only those that have been heard. Prior to Scontras et al.'s subjectivity hypothesis, one of the prevailing hypotheses held that adjectives were ordered based on which lexical semantic class (e.g., VALUE, DIMENSION, COLOR, MATERIAL) they belonged to (Dixon, 1982; Cinque, 1994, 2014). According to this theory, adjectives are sorted into semantic categories that fall into a hierarchy. Dixon (1982) suggests that adjectives would be ordered according to the hierarchy's order: VALUE > DIMENSION > PHYSICAL PROPERTY > SPEED >

HUMAN PROPENSITY > AGE > COLOR. Thus, an adjective that fell into the VALUE category would appear farthest from the noun, while an adjective in the COLOR category would appear closest to the noun.

Although it is known that there are stable adjective ordering preferences in English adult speakers, and that subjectivity predicts this, it has not been determined whether children follow the same pattern. The findings of this study will provide more insight into the development of linguistic preferences (i.e., adjective ordering) that rely on cognitive representations (i.e., subjectivity).

2 Background Literature

2.1 Adults: Mature knowledge

2.1.1 Lexical Semantic Class Hypothesis

Lexical class has been the prevalent theory for representing adjective ordering knowledge syntactically (Dixon, 1982; Cinque, 2014). According to this theory, adjectives are grouped into classes according to their semantic properties. For example, *blue* and *green* would be sorted into the COLOR class, while *big* and *small* would fall into the DIMENSION class. Adjective phrases are formed by following a syntactic hierarchy. According to the hierarchy proposed by Dixon (1982), as seen in Figure 1, adjectives would be ordered accordingly: VALUE > DIMENSION > PHYSICAL PROPERTY > SPEED > HUMAN PROPENSITY > AGE > COLOR, with VALUE being the highest up in the hierarchy and COLOR being the lowest in the hierarchy. Thus, adjectives belonging to a higher-up lexical class will be placed farther away from the noun, while adjectives in a lower lexical class will be placed closer to the adjective.

Scott (2002) proposed that each lexical class had its own separate functional projection in the syntax of a multi-adjective phrase, with adjectives in higher up classes of the hierarchy dominating those in the lower classes. However, Scott also proposed more and different classes than Dixon did, following this order: SUBJECTIVE COMMENT > SIZE > LENGTH > HEIGHT > SPEED > WIDTH > WEIGHT > TEMPERATURE > AGE > SHAPE > COLOR > NATIONALITY/ORIGIN > MATERIAL. Furthermore, Sproat and Shih's (1991) hierarchy consisted of QUALITY > SIZE > SHAPE > COLOR > PROVENANCE. Many other hierarchies have also been proposed, raising the question of which one is correct, and how specific the lexical classes ought to be. Additionally, what is the reason that the classes should be ordered in this way and not some other way? Is there some deeper reason for this specific ordering?

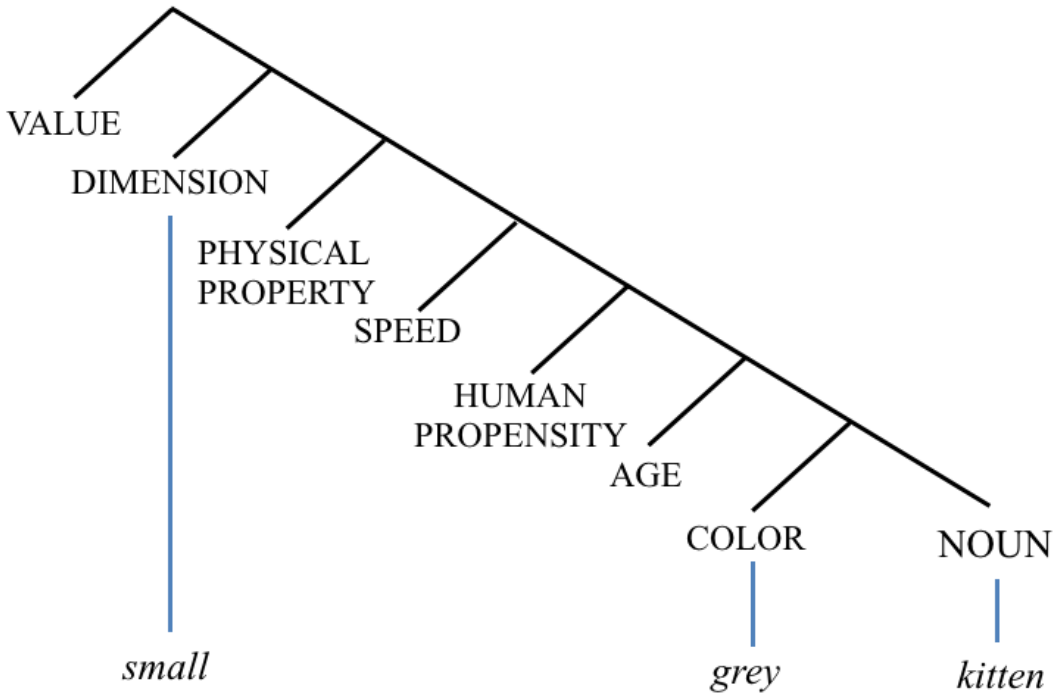

Figure 1: Dixon's (1982) lexical class hierarchy.

2.1.2 Subjectivity

Scontras et al. (2017) investigated another possible hypothesis for representing adjective ordering preferences: subjectivity. This hypothesis proposed that adjective ordering preferences in adults were determined by the subjectivity of the adjective, with less subjective adjectives occurring closer to the noun in a multi-adjective string. In order to test their hypothesis, a number of experiments were conducted to look at how correlated adjective ordering is with subjectivity.

The first experiment focused on confirming the existence of adjective ordering preferences and determining what the preferences were. This experiment used 26 adjectives sorted into seven different classes, as well as 10 nouns. They first examined naturalness by having 50 participants recruited through Amazon Mechanical Turk indicate which of two multi-adjective descriptions sounded more natural using a slider. For example, they might be asked to choose between "the red small chair" and "the small red chair." After collecting this data, the mean naturalness score of each adjective was calculated according to the average slider rating. Additionally, a corpus validation study was conducted to validate the behavioral measure. Phrases including a noun preceded by two unique adjectives were extracted from the British National Corpus and the Penn Treebank subset of the Switchboard corpus of telephone dialogues. The results of the corpus study were found to be highly correlated ($r^2 = 0.83$, 95% CI [0.63, 0.90]) with those from the naturalness ratings experiment, indicating that the naturalness ratings represent adult speakers' preferences in natural speech as well.

With ordering preferences now established, the authors conducted an experiment to measure subjectivity. 30 participants were recruited through Amazon Mechanical Turk. They were shown adjectives and were then asked to

adjust a slider to indicate how subjective they felt each adjective was, ranging from completely objective to completely subjective. Similar to the naturalness experiment and its corpus study validation, a faultless disagreement task was used to assess the validity of the findings of the subjectivity task. In this task, 40 participants were presented with a scenario in which two speakers are looking at the same item. The first speaker says, for example, "That apple is old." The second speaker then says, "That apple is not old." The participant's job is to rate with a slider the extent to which the statements of both speakers could be right, or if one speaker must be wrong. If both speakers can be right, their disagreeing faultlessly implies that old is subjective and its meaning can change according to the viewer's opinion. If one speaker must be wrong, then old must be less subjective or not at all. The results were found to be highly correlated with the subjectivity scores ($r^2 = 0.91$, 95% CI [0.86, 0.94]), indicating that the subjectivity task was a good measure, or that adults are accurate in determining the subjectivity of an adjective when asked explicitly.

Comparing the findings of the two experiments indicates that subjectivity is highly correlated with an adjective's position in relation to a noun, where adjectives considered the most subjective were found on average to be the farthest from the noun, and adjectives considered the least subjective were found on average to be closest to the noun. Adjective subjectivity scores accounted for 85% of the variance in the naturalness ratings ($r^2 = 0.85$, 95% CI [0.75, 0.90]). Similarly, faultless disagreement scores accounted for 88% of the variance in naturalness ratings ($r^2 = 0.88$, 95% CI [0.77, 0.95]).

The authors compared the performance of the subjectivity hypothesis in this analysis with the performance of other hypotheses, such as Whorf's (1945) adjective inherentness hypothesis and the concept-formability hypothesis, and

found subjectivity to do much better than the other hypotheses in predicting adjective ordering. The inherentness hypothesis accounted for 0% of the variance in naturalness ratings ($r^2 = 0.00$, 95% CI [0.00, 0.02]), and the concept-formability hypothesis also accounted for 0% of the variance ($r^2 = 0.00$, 95% CI [0.00, 0.00]).

Experiment 2 was carried out by Scontras et al. in order to further generalize what they had found in Experiment 1. As the first experiment only included 26 unique adjectives, this experiment expanded the adjectives used to 78, the semantic classes from 7 to 13, and the nouns from 10 to 166. 495 participants were recruited again through Amazon Mechanical Turk. These experiments were identical in methodology to Experiment 1, aside from the materials presented, which now included a wider range of adjectives, nouns, and semantic classes.

The first task assessed adult ordering preferences. Similar to Experiment 1's naturalness task, participants had to indicate which multi-adjective phrases sounded more natural. Participants again used a slider to indicate their choice. Unlike Experiment 1, this experiment included an option for "Neither option makes sense" if the participant felt that they could not use the slider to complete the trial. This was a result of descriptions possibly being nonsensical due to the adjectives being chosen at random (for example, "the wooden thick dream" would be nonsensical). Once again, naturalness scores were averaged. However, participants tended to have no stable adjective order preferences for the nonsense trials, and thus these trials were excluded from the analyses.

The second part of Experiment 2 evaluated subjectivity of the new adjectives with the same task from Experiment 1. 198 participants were recruited from Amazon Mechanical Turk. Again, the same methodology from Experiment 1's subjectivity task was used. Participants had to use a slider to determine how subjective 30 adjectives were. Upon averaging the subjectivity scores for the

adjectives, subjectivity scores were once again compared with naturalness ratings gained from the first part of Experiment 2.

Interestingly, the paper's analysis did find a subset of outliers with respect to subjectivity. Including the outliers, the statistical analysis found that subjectivity accounted for 51% of the variance in the naturalness ratings ($r^2 = 0.51$, 95% CI [0.32, 0.66]). The four superlatives, *last, closest, biggest*, and *best*, were also found to be outliers, due to the fact that they do not enter into adjective ordering preferences since they are always placed farthest away. After removing the superlatives from the analysis, subjectivity then accounts for 61% of the variance in the naturalness ratings ($r^2 = 0.61$, 95% CI [0.46, 0.71]). Removing both the outlier adjectives and superlatives increased subjectivity performance, accounting for 70% of the variance in the naturalness ratings ($r^2 = 0.70$, 95

Overall, this paper finds strong evidence supporting subjectivity being highly correlated to naturalness (mean distance from the noun), suggesting that subjectivity is the underlying representation for adult adjective ordering preferences.

2.2 Children: Developing knowledge

Scontras et al.'s study was limited to adult behavior, so the question of when and how these strong preferences emerge remained unanswered. There have been several studies examining adjective ordering preferences in children, but they were not successful in answering this question or in establishing the connection to subjectivity. Bever (1970) proposed that adjective order was determined by how closely lexically related an adjective is to a noun, with the more noun-like adjectives closer to the noun they are modifying. For example, the sentence "large

is my favorite size" is ungrammatical, while "red is my favorite color" is acceptable. In these sentences, both *large* and *red* are being used as nouns, although the use of *large* as a noun sounds unnatural. This is because *red* can often be used as both a noun and an adjective, while *large* can only be used as an adjective. According to Bever's noun-likeness hypothesis, *red* is more related to a noun (and thus would be closer to the noun) than *large* is, so the phrase "the large red house" should be more natural than "the red large house." Bever also conducted an experiment with children between 2 and 5 years of age, and found that younger children performed better on repeating unnatural adjective orders, such as "the plastic large pencil. . .". This would seem to indicate that younger children do not yet have stable adjective ordering preferences.

Martin and Molfese (1972) attempted to recreate Bever's experiment, but were unable to replicate his findings. They identified serious flaws with the repetition task, as Martin and Molfese's attempt yielded significantly different results. This led them to conclude that the repetition task was not a good measure for adjective ordering preferences. They instead used a production task, finding that 3- and 4-year-olds produced phrases with adjectives denoting CLEANLINESS closer to the noun than COLOR adjectives, while the adult preference is for adjectives denoting COLOR to be closer (i.e., "small clean yellow house"). This result provides early evidence that child preferences differ from adult preferences, but only with respect to CLEANLINESS and COLOR adjectives. Another study (Hare and Otto, 1978) had children in grades one through five arrange three adjectives of SIZE, COLOR, and MATERIAL with a noun to create adjective phrases that they thought were correct. The authors discovered that children in each succeeding grade level chose the adult preferred order of SIZE-COLOR-MATERIAL-noun more often than children in the preceding grade level. This seems to provide yet more evidence for some kind of

developmental trajectory, but with a focus on lexical semantic class, rather than on subjectivity, as was determined to be the predictor for adjective ordering by Scontras et al.

These developmental studies indicate that adjective ordering preferences develop or strengthen over time. However, there is disagreement among these studies on the age of acquisition, and where the preferences come from. Additionally, previous work focused on lexical semantic class being the underlying representation rather than subjectivity. There is not much to be concluded about child adjective ordering preferences from the literature, except that children do not start out matching adult preferences, but over time slowly become more adult-like. Considering that subjectivity is what adults are using to order adjectives, we should analyze when children begin to do the same. With this in mind, what kind of representations are children using and when do they develop abstract knowledge?

3 Hypotheses and Predictions

3.1 The hypotheses

We will be examining three possible hypotheses for representing the knowledge children possess about adjective order: input frequency, lexical semantic class, and subjectivity.

(i) **Input frequency:** This hypothesis simply states that children are noticing how often an adjective appears closest to the noun, or farthest from the noun. This approach involves word-by-word frequencies, or tracking statistics of individual adjectives, rather than something abstract (Goldberg, 2006;

Tomasello, 2000). Much research has shown that children do take into account frequency and positional statistics in other areas of language acquisition, so it seems plausible that children are utilizing this strategy to learn adjective orders (Saffran, Aslin, and Newport, 1996; Maye, Werker, and Gerken, 2002; Gerken, 2006; Mintz, 2003; 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). Under this hypothesis, the output is limited to the information from the input that a child receives, meaning that no additional abstraction occurs. The output would represent the word-level statistics observed from the input. This hypothesis is the null hypothesis, as there are no underlying representations for adjective order, just item-based knowledge. However, adults have preferences for adjective phrases they have never heard before, implying that adults are doing more than just tracking statistics in the input. Previous research suggests that adults must have some underlying representation for their preferences, namely, subjectivity (Scontras et al., 2017). While it is possible that children may use input frequency at a very young age, once they develop and begin to have abstract knowledge it is likely that they will no longer use this strategy as their adjective ordering preferences begin to look more like adult preferences.

(ii) **Lexical semantic class:** The lexical semantic class hypothesis was the prominent theory for adjective ordering preferences in adults and children. This hypothesis comes from the idea that adjectives are sorted into classes based on their semantic properties. These semantic classes create a hierarchy, and thus determine the order that adjectives should follow according to this

hypothesis. The higher up in the hierarchy an adjective's class is, the farther from the noun that adjective should be. This paper uses the classes and hierarchy from Scontras et al. 2017, which are ranked from highest to lowest as follows: VALUE > DIMENSION > SPEED > PHYSICAL > AGE > HUMAN LOCATION > TEMPORAL > COLOR > SHAPE > MATERIAL > NATIONALITY. However, there continues to remain the problem of what kind of classes and how many exist within our internal grammar, especially considering that many other proponents of this theory have differed on which classes to include when considering adjective ordering. Regardless, it is also known that children do begin forming categories for other aspects of language acquisition, so this hypothesis is a plausible representation for children (Mintz, 2003; Mintz, 2006; Booth and Waxman, 2002; Waxman and Booth, 2003).

(iii) **Subjectivity:** Scontras et al. propose their hypothesis that adjective order is predicted by subjectivity, where adjectives that are less subjective will appear closer to the noun. Considering that Scontras et al. found that subjectivity was indeed the best predictor for adult adjective ordering preferences, we might expect to observe this hypothesis in children's preferences. However, it has also been found that subjectivity is a concept that is too complex for young children to understand, so while it is the predictor for adult preferences, it may not be used by children until a later age (Foushee and Srinivasan, 2017).

3.2 Implementing the hypotheses

Each of these three hypotheses represents an option that children may utilize to order adjectives. In order to determine which hypothesis is the best predictor of

how children do so, we need to figure out which representation most closely matches children's behavior. This paper takes the approach that each hypothesis acts as a multi-adjective-string-generating machine that takes input and transforms it into output based on the premises of a given hypothesis. We compare the predictions that each hypothesis produces to the actual observed output from children, and the hypothesis under which the child-produced dataset has the highest likelihood will be the most likely underlying representation for children's adjective ordering preferences. For example, the input frequency hypothesis predicts that an adjective's position is based on how frequently it is observed in that position in the input. If an adjective tends to appear in the farthest position in the input, but the output tends to place that adjective in the closest position, then that would decrease the likelihood for the input frequency hypothesis. On that same vein, the lexical semantic class and subjectivity hypotheses predict an adjective's position based on its lexical class or subjectivity score, respectively. Observing adjective positions in the output that do not match where the hypotheses predict them to appear in would also decrease the hypotheses' probabilities.

To analyze the hypotheses, both child-produced data (output) and child-directed data (input) are needed. The child-produced data is the set of observed multi-adjective strings that have actually been produced by children. The child-directed data is the input that children are receiving from adults on which they base their preferences. This data was obtained from corpora containing a large quantity of utterances by adults and children. Additionally, in order to test the lexical semantic class and subjectivity hypotheses, each unique adjective in the corpus needed to be sorted into a lexical semantic class and scored for its perceived subjectivity.

4 Collecting Data

4.1 Corpus data

The corpus study was conducted on all the available English data on the CHILDES database (MacWhinney, 2000) from the North American and United Kingdom corpora (1,927,582 child-produced utterances; 2,741,507 child-directed utterances; ages 0 months to 16 years). However, we found that a majority of the data including adjective-adjective-noun phrases occured within the data from ages 2 through 4, so we decided to focus our analysis on the data from children at these ages. The data from those ages contained 1,069,406 child-produced utterances and 688,428 child-directed utterances. The corpus study methodology is similar to that from Scontras et al. (2017). Phrases with two adjectives in front of a noun (i.e., "adjective-adjective-noun") were extracted from the corpora.

In order to ensure that children's productions were not merely repetitions from the adult input, we calculated the percentage of child-produced phrases that were immediate repetitions following an adult utterance. If the children were doing this a significant amount of time, this would indicate that the data used would not be useful for analysis of children's multi-adjective string production through some underlying representation, namely subjectivity, lexical semantic class, or input frequency. Our repetition analysis found that in the data from ages 2 through 4, only 0.5% of the multi-adjective strings were an immediate repetition produced by a child following an adult utterance. This low percentage indicates that 2-, 3-, and 4-year-old children are in fact generating their own multi-adjective phrases, with some underlying representation determining their preferred adjective order.

For the input frequency hypothesis, we calculated how often adjectives appeared in the 1-away position (closest to the noun) and in 2-away position

(farthest from the noun).

For the lexical semantic class hypothesis, we used the categories from Scontras et al.'s expanded experiment in this order: VALUE > DIMENSION > SPEED > PHYSICAL > AGE > HUMAN LOCATION > TEMPORAL > COLOR > SHAPE > MATERIAL > NATIONALITY. We had 307 unique adjectives in total from both the child-directed and child-produced data. Adjectives that coincided with those used in Scontras et al.'s experiment remained in their already-determined lexical classes, and the remaining unclassified adjectives were sorted by hand into the twelve classes. This involved looking at each individual adjective's meaning and best determining which lexical class it belonged to by coming to a consensus among the four collaborators. Some adjectives (82) were either too ambiguous or failed to fall neatly into one of the lexical classes, so they were classified as OTHER.

For subjectivity, we ran an experiment to obtain subjectivity scores for the 307 unique adjectives from the CHILDES corpus.

4.2 Measuring subjectivity

Participants. 108 participants were recruited for this experiment through Amazon.com's Mechanical Turk crowdsourcing service. Participants were compensated with $0.30 for their involvement in the experiment. All participants were native English speakers.

Stimuli. 307 unique adjectives were used. A list of the adjectives used and the lexical classes they were classified into is available in the appendix. An example of one trial is shown in Figure 2.

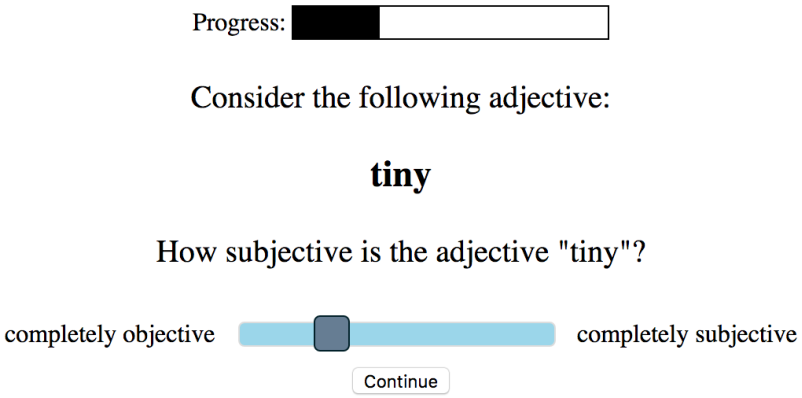

Figure 2: One trial from the subjectivity experiment.

Procedure. Participants were instructed that they would see 30 adjectives and that their task would be to determine how subjective they are. Each trial consisted of a random adjective from the stimuli; participants were asked to adjust a slider to indicate how subjective they felt each adjective was, ranging from completely objective (coded as 0) to completely subjective (coded as 1). There were 30 trials in total for each participant.

Results. The scores for each adjective were averaged across all the participants, which resulted in the final subjectivity score for an adjective. These scores were used in the analyses performed in the next section.

5 Metrics

The knowledge children are using to determine adjective ordering is represented by three possible hypotheses: input frequency, lexical class, and subjectivity. The hypothesis that is most likely to have generated the observed child-produced data

is considered the hypothesis that best represents the children's knowledge. This is determined by calculating the likelihood of a given hypothesis generating the data.

In order to determine the likelihood of the child-produced data under each hypothesis, each individual adjective's likelihood will be calculated by observing how often it appears in the 1-away or 2-away position in the set of observed child-produced data. Multiplying these likelihoods under a hypothesis gives the likelihood for the entire dataset under that hypothesis. A likelihood for an adjective looks at the number of times that adjective appears in the output and in the 2-away position, and the probability that it should appear in the 2-away position given the hypothesis and the input.

The probability of an adjective appearing in the 2-away position under the input frequency hypothesis ($p_2exp(adj_x|h_i = h_{freq}$ in Equation 1) depends on how often it appears in the 2-away position in the input. This frequency count ($f_{2input}(adj_x)$ in Equation 1) is then divided by the total number of strings in which the adjective appeared in any position ($N_{input}(adj_x)$ in Equation 1), giving us the probability of the adjective appearing in the 2-away position in the output under the input frequency hypothesis. A smoothing factor, represented as α, is added to ensure that the overall probability cannot end up as 0. In this equation, α equals 0.5, and this small value is added to every adjective observed. α is also multiplied by 2 in the denominator to represent the two possible positions for the adjective to appear in the string.

$$p_2exp(adj_x|h_i = h_{freq}) = \frac{f_{2input}(adj_x) + \alpha}{N_{input}(adj_x) + 2 * \alpha} \tag{1}$$

The probability of an adjective appearing in the 2-away position under the lexical class hypothesis ($p_2exp(adj_x|h_i = h_{lex}$) in Equation 2) depends on the

lexical class of the adjective it appears with in a multi-adjective string. For example, if the adjective appears with another adjective that is hierarchically-closer to the noun, then that adjective would be expected to appear in the 2-away position. If the adjective appears with an adjective in the same lexical class, then the adjective has a 50% chance of appearing in the 2-away position. Thus, having more hierarchically-closer adjectives (relative to the adjective in question) in the input means a higher chance of the adjective being in the 2-away position in the output. The adjective's likelihood under the lexical class hypothesis can be determined by calculating the number of adjective tokens in a lexically-closer class, ($f_{input}(< adj_x|h_i)$ in Equation 2). This count is multiplied by 1 to represent the expected appearance in the 2-away position. The number of adjective tokens in the same lexical class ($f_{input}(= adj_x|h_i)$ in Equation 2) is multiplied by 0.5 to represent the 50% chance of appearing in the 2-away position. These two numbers are added together and then divided by the total number of adjective tokens in the input, ($N_{input}(Adj)$ in Equation 2), giving us the probability of the adjective appearing in the 2-away position under the lexical class hypothesis. The smoothing factor α is also added to ensure that the overall probability cannot end up as 0. This small value is added to every adjective type observed. Additionally, the denominator includes α as well by multiplying it by the number of adjective types ($|Adj|$ in Equation 2).

$$p_2exp(adj_x|h_i = h_{lex}) = \frac{f_{input}(< adj_x|h_i) + 0.5 * f_{input}(= adj_x|h_i) + \alpha}{N_{input}(Adj) + \alpha * |Adj|} \quad (2)$$

The probability of an adjective appearing in the 2-away position under the subjectivity hypothesis ($p_2exp(adj_x|h_i = h_{subj})$ in Equation 3) is determined very similarly to the probability under the lexical class hypothesis. However, the

probability under the subjectivity hypothesis is dependent upon the subjectivity score of the adjective it appears with. If the adjective in question appears with a more subjective adjective, it will be expected to appear in the 2-away position. If the adjective appears with an equally subjective adjective, it will have a 50% chance of appearing in the 2-away position. The number of adjective tokens that are more subjective ($f_{input}(< adj_x|h_i)$ in Equation 3) or equally subjective are counted ($f_{input}(= adj_x|h_i)$ in Equation 3) and then multiplied by 1 and 0.5, respectively, to represent their chance of appearing in those positions. These numbers are also added and then divided by the total number of adjective tokens, resulting in the probability of the adjective appearing in the 2-away position under the subjectivity hypothesis. The smoothing factor α is again included in the subjectivity hypothesis calculations to avoid an observed 0 probability affecting the final probability calculation.

$$p_2exp(adj_x|h_i = h_{subj}) = \frac{f_{input}(< adj_x|h_i) + 0.5 * f_{input}(= adj_x|h_i) + \alpha}{N_{input}(Adj) + \alpha * |Adj|} \quad (3)$$

Having these probabilities for the different hypotheses allows us to determine the likelihood for an individual adjective appearing in the 2-away position in the output for a given hypothesis ($p(D(adj_x|h_i)$ in Equation 4). Equation 4 demonstrates how this likelihood is calculated. Since the data can appear in any order, we need to determine the number of possible ways for the observed dataset pattern to be generated ($\binom{N}{t}$ in Equation 4). For example, say we have a dataset that includes three strings, {*small grey kitten; small brown kitten; nice small kitten*}, where *small* appears in the 2-away position two times. We need to account for number of ways of generating the pattern seen in the dataset (*small* in the 2-away position twice and in the 1-away position once). To do this, we use the

the number of times that the adjective appeared in a multi-adjective string in the output (N in Equation 4), and the number of times the adjective appeared in the 2-away position (t in Equation 4). In our *small* example, there are three possible ways, or orders, that gives us this pattern of *small* in the 2-away position: this is equivalent to 3 (N) choose 2 (t). Once we have this number, it is multiplied by the probability of an adjective appearing in the 2-away position under a given hypothesis ($(p_2 exp(adj_x|h_i))^t$ in Equation 4) and the probability of that adjective appearing in the 1-away position under a given hypothesis is calculated ($(1 - p_2 exp(adj_x|h_i))^{N-t}$ in Equation 4). Multiplying these three parts gives us the overall likelihood of an individual adjective under the given hypothesis.

$$p(D(adj_x)|h_i) = \binom{N}{t}(p_2 exp(adj_x|h_i))^t(1 - p_2 exp(adj_x|h_i))^{N-t} \tag{4}$$

Finally, multiplying all the individual adjective likelihoods yields the likelihood for the entire dataset under the given hypothesis ($p(D|h_i)$ in Equation 5), after which the three hypotheses' performances can then be compared to determine which best matches the children's observed output.

$$p(D|h_i) = \prod_{adj_x \in Adj} p(D_{adj_x}|h_i) \tag{5}$$

The results of these multiplications will often yield very small numbers, and thus the products have been log transformed using log base e.

6 Results

The performance of the three hypotheses was calculated at ages 2, 3, and 4. Adjectives that were assigned both a lexical class and a subjectivity score were

included in the analysis. Thus, the 82 adjectives assigned to the OTHER category were excluded, as this class does not stand as a coherent semantic class. Because these are log likelihood scores, the less negative numbers (or larger values) represent the more probable hypothesis at that age. While the scores can be compared within an age across hypotheses, they cannot be compared across ages due to each age being a different dataset.

At age 2, we see that the input frequency hypothesis outperforms the other two hypotheses, with the subjectivity hypothesis performing intermediate and the lexical class hypothesis performing the worst. This suggests that at age 2, children are just tracking the statistics of individual words in the input. Table 1 shows the likelihood scores of each hypothesis at age 2.

Table 1. *Hypothesis Log Likelihood Scores at Age 2*

age	input frequency	lexical class	subjectivity
2	**-202.6**	-334.9	-274.6

Note: Scores range from 0 to negative infinity, with 0 being the most probable. The bolded score is the best performing hypothesis.

At age 3, the input frequency hypothesis continues to outperform the other two. However, the probability scores for the lexical class and subjectivity hypotheses begin catching up to the winning hypothesis. These scores can be seen in Table 2.

Table 2. *Hypothesis Log Likelihood Scores at Age 3*

age	input frequency	lexical class	subjectivity
3	**-125.1**	-164.0	-163.0

At age 4, the lexical class hypothesis surpasses the input frequency hypothesis, becoming the winning representation, suggesting an emergence of abstract knowledge at this age. Table 3 shows the scores for age 4.

Table 3. *Hypothesis Log Likelihood Scores at Age 4*

age	input frequency	lexical class	subjectivity
4	-182.9	**165.2**	-193.5

Now that we've examined the best performing hypothesis for each age dataset, we can also see how abstract knowledge emerges over time by comparing the abstract representational hypothesis probabilities against the best performing hypothesis. This is done by subtracting the scores of the lexical class and subjectivity hypotheses from the score of the best performing hypothesis. Smaller differences mean closer probabilities, while a 0, or no difference, means a hypothesis has taken over as the winning hypothesis. We can use these differences to compare how close a representational hypothesis is to the best performing hypothesis. Table 4 shows these differences.

Table 4. *Differences Between Worse-Performing and Best-Performing Hypotheses*

age	lexical class vs. best	subjectivity vs. best
2	-132.3	-72
3	-38.9	-37.9
4	**0**	-28.3

Over time, the gap between the representational hypotheses and the winning hypothesis narrows. The differences for lexical class decreases until it takes over input frequency as the winning hypothesis at age 4. For subjectivity, we see the same narrowing as children are getting older, although it never takes over as the winning hypothesis in the data we have, which stops at age 4.

7 Discussion

The results of this analysis tell us that children begin using input frequency as the underlying representation for adjective ordering at ages 2 and 3, and eventually switch to a more abstract representation, lexical semantic class, at age 4. This reveals that children are simply tracking word-level statistics until age 4, when children start classifying adjectives into lexical classes, suggesting an emergence of abstract lexical class knowledge.

These results provide insight into the developmental trajectory of children's acquisition of abstract knowledge, as there is a clear increase over time in the likelihood of abstract knowledge being the underlying representation for children's adjective ordering. However, in our dataset, we only see abstract lexical class knowledge, rather than subjectivity-based knowledge. As past research (Scontras et al.) has determined that subjectivity is the best predictor for adjective ordering preferences in adults, it would logically follow that the trajectory would continue to develop into subjectivity-based knowledge as the underlying representation. The performance of the subjectivity hypothesis did increase in comparison to the best-performing hypothesis, so we could also expect to see the gap narrow until it takes over as the winning hypothesis. Children may perhaps be using lexical class as a rudimentary method of ordering adjectives before they finally develop subjectivity-based knowledge and shift to subjectivity as the representation at some later age.

We never see subjectivity overtake input frequency or lexical class as the winning hypothesis, and so it is still uncertain when between 4 years old and adulthood that it does. However, recent research (Foushee and Srinivasan, 2017) has determined that children do not develop subjectivity awareness and

understanding until much later, around ages 8 to 9. Thus we could expect to see subjectivity become the winning representation at around these ages. However, further investigation would be required to determine whether this is indeed the point in the trajectory that subjectivity becomes the abstract representation underlying children's adjective ordering preferences.

Further research could look into the cross-linguistic development of adjective ordering preferences in children, as robust preferences do exist in adults in other languages. Future work also involves looking at which representations adults are using to form their output to children, or the input that the children are receiving. It is known that adults do adjust their speech when speaking to children, often emphasizing certain aspects of language (Kunert, Fernandez, and Zuidema, 2011; Ferguson, 1964; Fernald, Taeschner, Dunn, Papousek, de Boysson-Bardies, and Fukui, 1989; Grieser and Kuhl, 1988; Snow, 1977), and so it would be interesting to see whether adults are also basing their speech off of the same representations that children or adults use.

In conclusion, using quantitative approaches and corpus analysis gives us insight into the point in the developmental trajectory at which abstract underlying representations begin to emerge for adjective ordering preferences. Although we do not see subjectivity become the winning representation in our data, this study has found that lexical knowledge emerges at age 4, providing a start for further study of the trajectory of children's abstract knowledge development.

References

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

Booth, A. E., & Waxman, S. (2002). Object names and object functions serve as cues to categories for infants. *Developmental Psychology*, *38*(6), 948–957.

Cinque, G. (1994). On the evidence for partial N-movement in the Romance DP. In R. S. Kayne, G. Cinque, J. Koster, J.-Y. Pollock, L. Rizzi, & R. Zanuttini (Eds.), *Paths Towards Universal Grammar. Studies in Honor of Richard S. Kayne* (pp. 85–110). Washington DC: Georgetown University Press.

Cinque, G. (2004). The semantic classification of adjectives: A view from syntax. *Studies in Chinese Linguistics*, *35*(6), 1–30.

Dewar, K. M., & Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge: Evidence from 9-month-old infants. *Psychological Science*, *21*(12), 1871–1877.

Dixon, R. M. (1982). *Where have all the adjectives gone? and other essays in semantics and syntax* (Vol. 107). Walter de Gruyter.

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*(3), 427–438.

Ferguson, C. (1964). Baby talk in six languages. *American Anthropologist*, *66*, 103–113.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (n.d.). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501.

Foushee, R., & Srinivasan, M. (2017). Could both be right? Children's and adults' sensitivity to subjectivity in language. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 379–3384). London, UK: Cognitive Science Society.

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

Gerken, L., & Knight, S. (2015). Infants generalize from just (the right) four words. *Cognition*, *143*, 187–192.

Gerken, L., & Quam, C. (2017). Infant learning is influenced by local spurious generalizations. *Developmental Science*, *20*(3).

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Grammatical categories. (n.d.). *Language*, *21*(1), 1–11.

Grieser, D., & Kuhl, P. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology*, *24*(1), 14–20.

Hare, V. C., & Otto, W. (1978). Development of preferred adjective ordering in children, grades one through five. *The Journal of Educational Research*, *71*(4), 190–193.

The item-based nature of children's early syntactic development. (n.d.). *Trends in cognitive sciences*, *4*(4), 156–162.

Kunert, R., Fernández, R., & Zuidema, W. (2011). Adaptation in child directed speech: Evidence from corpora. *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue*, *112–119*.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Martin, J. E., & Molfese, D. L. (1972). Preferred adjective ordering in very young children. *Journal of Verbal Learning and Verbal Behavior*, *11*(3), 287–292.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*(1), 122–134.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117.

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

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Scontras, G., Degen, J., & Goodman, N. D. (2017). Subjectivity predicts adjective ordering preferences. *Open Mind: Discoveries in Cognitive Science*, *1*(1), 53–65.

Scott, G.-J. (2002). Stacked adjectival modification and the structure of nominal phrases. In G. Cinque (Ed.), *The cartography of syntactic structures, volume 1: Functional structure in the dp and ip* (pp. 91–120). Oxford: Oxford University Press.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Snow, C. (1977). The development of conversation between mothers and babies. *Journal of Child Language*, *4*(1), 1–22.

Sproat, R., & Shih, C. (1991). The cross-linguistic distribution of adjective

ordering restrictions. In *Interdisciplinary approaches to language* (pp. 565–593). Springer.

Waxman, S., & Booth, A. (n.d.). The origins and evolution of links between word learning and conceptual organization: New evidence from 11-month-olds. *Developmental Science*, *6*(2), 128–135.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Appendix

Table A.1: List of all unique adjectives used

Adjective	Class	Adjective	Class	Adjective	Class
ancient	age	sorry	human	beautiful	value
fresh	age	stupid	human	bonnie	value
new	age	vain	human	brilliant	value
old	age	vicious	human	comfy	value
ripe	age	whiny	human	correct	value
teenage	age	wicked	human	cozy	value
young	age	wise	human	crummy	value
black	color	back	location	cute	value
blue	color	bottom	location	dandy	value
bright	color	central	location	dear	value
brown	color	close	location	delicious	value
clear	color	east	location	dulcet	value
dark	color	far	location	exciting	value
faint	color	front	location	fancy	value
gold	color	glottal	location	fantastic	value
golden	color	high	location	fun	value
gray	color	left	location	good	value
green	color	low	location	gorgeous	value
grey	color	meadow	location	grand	value
orange	color	polar	location	great	value
pale	color	side	location	handy	value
pink	color	south	location	horrible	value
purple	color	top	location	horrid	value
red	color	upstairs	location	illustrious	value
silver	color	western	location	junk	value
tan	color	concrete	material	just	value
tawny	color	haired	material	nasty	value
wan	color	iron	material	neat	value
white	color	plastic	material	nice	value
yellow	color	silk	material	nifty	value
big	dimension	wooden	material	odd	value
bitsy	dimension	wool	material	okay	value
chubby	dimension	American	nationality	pathetic	value
eensie	dimension	Chinese	nationality	perfect	value
fat	dimension	Dutch	nationality	plain	value
flat	dimension	English	nationality	poop	value
giant	dimension	French	nationality	poor	value
gigantic	dimension	Mexican	nationality	posh	value
huge	dimension	bald	physical	precious	value
itsy	dimension	bumpy	physical	pretty	value
itty	dimension	clean	physical	proper	value
large	dimension	cold	physical	regular	value
little	dimension	cool	physical	right	value
long	dimension	crisp	physical	rotten	value
narrow	dimension	crooked	physical	scary	value
open	dimension	crunchy	physical	special	value
podgy	dimension	damp	physical	strange	value
short	dimension	delicate	physical	super	value
skinny	dimension	dense	physical	tidy	value

Adjective	Class	Adjective	Class	Adjective	Class
small	dimension	dirty	physical	ugly	value
tall	dimension	dry	physical	weird	value
teensie	dimension	empty	physical	wild	value
teeny	dimension	fluffy	physical	wrong	value
thick	dimension	fragile	physical	yucky	value
thin	dimension	full	physical	yummy	value
tiny	dimension	fuzzy	physical	bubbly	other
trim	dimension	hard	physical	bumpity	other
tubby	dimension	heavy	physical	busy	other
wee	dimension	hollow	physical	choice	other
weensie	dimension	hot	physical	darned	other
weeny	dimension	light	physical	different	other
wide	dimension	loud	physical	double	other
angry	human	misshapen	physical	even	other
attentive	human	quiet	physical	exact	other
brave	human	rough	physical	fake	other
clever	human	salty	physical	fellow	other
cranky	human	sharp	physical	formal	other
crazy	human	shiny	physical	free	other
cross	human	slick	physical	glitzy	other
dead	human	slippery	physical	head	other
dizzy	human	smooth	physical	horned	other
dump	human	soft	physical	medical	other
fair	human	sparkly	physical	messy	other
famous	human	spicey	physical	multiple	other
fierce	human	squeaky	physical	musical	other
gentle	human	steep	physical	naked	other
gloomy	human	strong	physical	natural	other
goofy	human	sweet	physical	obvious	other
gracious	human	warm	physical	organic	other
greedy	human	waterproof	physical	own	other
grouchy	human	wet	physical	part	other
grumpy	human	oval	shape	personal	other
happy	human	round	shape	pokey	other
healthy	human	square	shape	pretend	other
human	human	squiggly	shape	ready	other
hungry	human	fast	speed	real	other
innocent	human	instant	speed	roundabout	other
jolly	human	quick	speed	same	other
kind	human	slow	speed	sick	other
lazy	human	speedy	speed	sore	other
live	human	early	temporal	spare	other
mad	human	final	temporal	steady	other
maternal	human	first	temporal	still	other
mean	human	last	temporal	stinky	other
merry	human	late	temporal	trick	other
naughty	human	next	temporal	true	other
royal	human	past	temporal	undercover	other
sad	human	second	temporal	verbal	other
silly	human	third	temporal	whole	other
sleepy	human	alright	value	wiggly	other
smart	human	awesome	value	wriggly	other
smiley	human	awful	value		
sneaky	human	bad	value		