Word Co-occurrence in Child-Directed Speech Predicts Children's Free Word Associations

Abdellah Fourtassi
Aix-Marseille Univ, Universite de Toulon, CNRS, LIS, ILCB, Marseille, France
abdellah.fourtassi@gmail.com

Abstract

The free association task has been very influential both in cognitive science and in computational linguistics. However, little research has been done to study how free associations develop in childhood. The current work focuses on the developmental hypothesis according to which free word associations emerge by mirroring the co-occurrence distribution of children's linguistic environment. I trained a distributional semantic model on a large corpus of child language and I tested if it could predict children's responses. The results largely supported the hypothesis: Co-occurrence-based similarity was a strong predictor of children's associative behavior even controlling for other possible predictors such as phonological similarity, word frequency, and word length. I discuss the findings in the light of theories of conceptual development.

1 Introduction

The mental lexicon is organized into a structure such that exposure to a given word, e.g. "cat", tends to activate semantically similar words such as "milk" or "dog" (Collins and Loftus, 1975; McNamara, 2005). In order to characterize this structure, researchers in cognitive science have often relied on the free association task where people are given a list of cue words and asked to provide the first words that come to mind. Data from this task — especially the word association norms collected by Nelson et al. (2004)— have proven successful in accounting for a variety of psycholinguistic phenomena (De Deyne and Storms, 2015). In addition, they have often been used as ground truth in evaluating the ability of NLP models to approximate human lexico-semantic organization (Silberer et al., 2013; Fourtassi and Dupoux, 2013; Vulić et al., 2017) .

While adult word associations have been extensively studied, it is still poorly understood how these associations develop in childhood. This is a significant gap in the developmental literature: To the extent that word free associations help us understand conceptual organization, the study of how these associations develop can inform our theories of conceptual development (Wojcik and Kandhadai, 2019).

Previous work has shown that adults' word associations can be predicted by patterns of co-occurrence in language use (Griffiths et al., 2007) (but see De Deyne et al. 2016), that is, pairs of words that tend to appear as cue-response in the free association data are also those that co-occur frequently in the language people are exposed to. Developmentally, thus, a plausible scenario is that word associations first originate in childhood by mirroring the co-occurrence distribution in the language children hear around them. Previous research has shown that word co-occurrence can support linguistic and conceptual development (e.g., Fourtassi et al. 2014, 2019), here I investigate how co-occurrence in child-directed speech can predict children's free associations.

The paper is organized as follows. First, I briefly present the data and methods. Second, I 1) quantify the variability in children's associations compared to that of adults, 2) explore if, despite individual variability, co-occurrence probabilities predict word associations, and 3) test if co-occurrence similarity remains predictive when controlling for other factors such as phonological similarity, response frequency, and response length. Finally, I discuss the results in the light of theories of early conceptual development.

2 Data and Methods

2.1 Word Association Data

I rely on a new dataset of children's free associations collected by Wojcik and Kandhadai (2019).

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 49–53
Online Event, November 19, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

For a detailed description of the data collection process, please refer to that paper. In brief, the authors used age-appropriate stimuli to collect word associations from both children ($N = 60$; age range 3 – 8 years) and adults ($N = 60$). Participants were instructed to respond to a cue word with the first word that came to mind. The list of cue words consisted of 65 of the most frequent words in a large corpus of child language. It contained 25 nouns, 17 adjectives, 12 verbs, and 6 others (e.g., "yes").[1] Following Wojcik and Kandhadai (2019), and in order to maximize the statistical power of the analyses, participants were categorized by age group as "Adult", "Older children" (6 – 8 years), or "Younger children" (3–5 years).

2.2 Methods

2.2.1 Normalized Entropy

In order to quantify individual variability, I followed (Dubossarsky et al., 2017) in measuring, for each cue word y, the normalized entropy $H(y)$ defined as:

$$H(y) = \sum_{i=1}^{N} \frac{p(x_i) * log_2(p(x_i))}{log_2(N)}$$

Where $p(x_i)$ is the probability of a response x_i, which I obtain, for each cue y, by averaging across responses for that cue. N is the total number of different responses given by all participants for a given cue. H has values between 0 (total agreement) and 1 (no agreement).

2.2.2 Co-occurrence-based Similarity

I derived co-occurrence similarity in children's linguistic environment using Word2vec (Mikolov et al., 2013), a widely used distributional semantic model where pairs of words are assigned a similarity score based on the patterns of their co-occurrence in similar contexts (the context being the set of neighboring words). I trained the model on a large corpus of child language (CHILDES, MacWhinney 2000). Since I evaluate free association data of children aged 3 years and older, I chose to restrict the input data to the first three years (the common amount of exposure across age groups), allowing us to make comparison between age groups based on similar data. For comparison, I also used pre-trained vectors from Google Freebase trained on about 100 billion words from various news articles.[2]

2.2.3 Phonological Distance

First, I converted the orthographic transcription of word pairs into their phonological forms using the CMU pronouncing dictionary. [3] Then, I measured the Levenshtein distance (also known as edit distance) of each pair. This measure counts the minimum number of operations (insertions, deletions, substitutions) required to change one member of the pair into the other. Based on previous research that reported children tend to give phonologically similar responses (e.g., "house" - "mouse") (Cronin, 2002), I defined a binary variable that distinguished between small values of edit distance (edit ≤ 1) and larger values (edit > 1).

2.2.4 Frequency and Length

Previous research has shown that both frequency and length play an important role in children's expressive language (Braginsky et al., 2019; Fourtassi et al., 2020). I test the extent to which they also constrain children's expressive free word associations. Note that while co-occurrence and phonological similarity characterize the relationship between the cue and the response, frequency and length characterize only the response: They capture possible response tendencies regardless of the identity of the cue word. I obtained word frequency based on CHILDES data and I defined word length as the number of phones in the phonological transcription.

3 Analyses

3.1 The Development of Individual Variability

In Figure 1, the normalized entropy by age group shows how heterogeneity in responses develops. Children had higher entropy than adults, showing a greater diversity in their responses. This result replicates findings by Wojcik and Kandhadai (2019) who used different measures such as idiosyncrasy (percent of responses given by only one participant) and the number of common responses. I did not find a difference between the younger and older children, however: Both had as much variability in their responses on average, suggesting that

[1] The data is publically available at osf.io/qat7w

[2] the pre-trained vectors can be found at https://code.google.com/archive/p/word2vec/

[3] Retrieved from http://www.speech.cs.cmu.edu/cgi-bin/cmudict

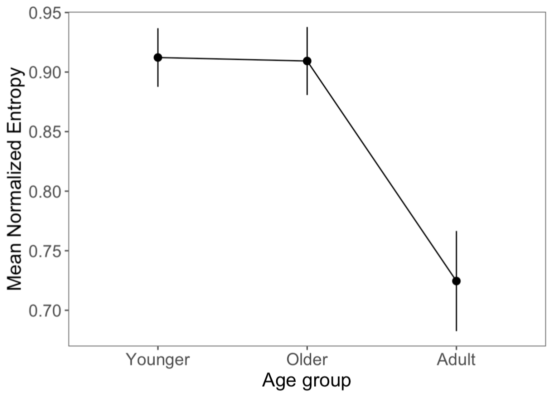

Figure 1: Mean normalized entropy by age group. The error bars indicate 95% confidence intervals.

children in this age range have not really started developing a shared lexical organization with one another, at least not to the same extent that adults do.

3.2 The Role of Co-occurrence in the Linguistic Environment

Figure 2 shows the average co-occurrence-based similarity of cue-response pairs in each age group using data from both CHILDES and Google Freebase. I found several interesting results. First, even Younger children were well above chance (chance being the average co-occurrence-based similarity between two random words), meaning that their associations already begins to reflect the co-occurrence distribution in the linguistic environment. Second, we see a clear developmental trend: On the one hand, older children had a higher correlation score than Younger children, suggesting that children mirror more and more the distribution of the input between 3 and 8 years old. On the other hand, adults had a higher score than older children, showing that development continues beyond this age range. Finally, I obtained similar developmental patterns whether I used CHILDES or Freebase, suggesting that children's associations reflect the distributional structure of their native language beyond the idiosyncrasies of child-directed speech.

3.3 Comparison to Other Developmental Factors

Here I examine how the role of co-occurrence compares to other possible predictors. I ran a mixed-effects model where the predicted variable was the probability of a response given a cue (obtained by summing over similar responses and dividing by all responses for a given cue). The predictors were the co-occurrence-based similarity from CHILDES (`co-occurrence`), phonological distance (`Phono`), log-frequency (`LogFreq`), and length (`Length`) (all were centered and scaled). To study change across time, I added an interaction term with age (`Age`) to each predictor. Finally, to take into account the possibly correlated data for the same cue, I specified `Cue` as a random effect in the model. The results of this regression are shown in Table 1.

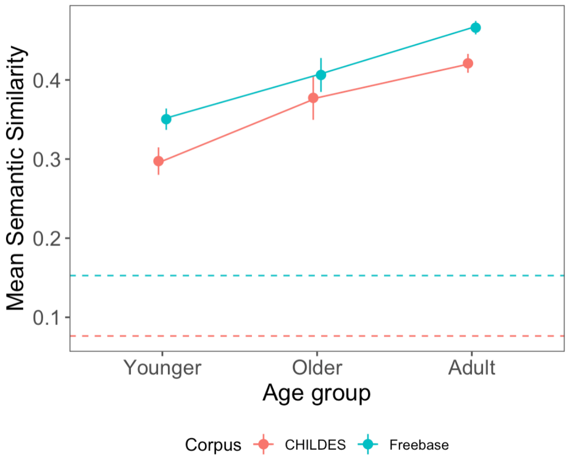

Figure 2: Average cosine similarity of cue-target pairs in each age group using data from both CHILDES and Freebase. The dotted lines represent chance similarity values in each training dataset. The error bars indicate 95% confidence intervals.

Table 1: Estimates of a regression model predicting the probability of responses in free association data. The model was specified as `Response ∼ (co-occurrence + Phono + LogFreq + Length)*Age + (1 | Cue)`

Predictors	*Estimates (95% CI)*
(Intercept)	7.997*** (6.497, 9.497)
Co-occurrence	4.562*** (3.924, 5.199)
LogFreq	2.022*** (1.314, 2.730)
Length	0.203 (−0.510, 0.916)
Phono	3.428*** (2.026, 4.831)
Age	−1.520 (−3.040, 0.0001)
Co-occurrence:Age	1.695*** (0.997, 2.393)
LogFreq:Age	1.004* (0.233, 1.774)
Length:Age	0.204 (−0.582, 0.989)
Phono:Age	2.045** (0.521, 3.570)
Note:	$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Confirming the previous analysis, the response's degree of co-occurrence with the cue strongly predicted the response's probability, even controlling for other predictors. The response's phonological distance with the cue and the response's frequency in the linguistic environment were also strong predictors of the response's probability. That is, participants were overall more likely to provide phonologically dissimilar and frequent responses. The length of the response, however, was not a significant predictor.

Concerning developmental change, co-occurrence strongly interacted with age, meaning that the effect of co-occurrence-based similarity becomes greater over development. The same can be said about phonological distance: Participants provide less and less phonologically similar response as they grow older. Frequency also varied with age, although to a lesser extent, meaning that responses get slightly more frequent over development.

in order to compare younger and older groups in a more direct fashion, I ran a second regression that was identical to the first one, but with adult data removed. In this second regression (estimates not shown here), co-occurrence-based similarity also interacted with age, indicating that a developmental change occurs between 3 and 8 years whereby children's responses mirror more and more the distribution of the language they are exposed to.

4 Discussion

Despite the fact that free associations varied greatly from child to child, they were highly predicted by their co-occurrence-based similarity in the linguistic environment. This prediction remained strong even controlling for other factors that may influence children's responses such as phonological similarity, word frequency, and word length. I found an interesting developmental change that appears to take place between 3 and 8 years old and whereby children's responses reflect more and more the co-occurrence structure of their native language, while becoming less tied to the lower-level phonological similarity. That said, I also found a difference between older children and adults, suggesting this development continues well into late childhood.

Since free associations have been used to study conceptual organization, the current study contributes to the literature on conceptual development. A big challenge in this literature is to understand how taxonomic categories (e.g., animal vs. artifact) are formed by children despite the fact that members of such categories do not necessarily look similar (e.g., fish and bird). Using free association data (the same I use here), Wojcik and Kandhadai (2019) showed that children's free associations become more paradigmatic/taxonomic in nature between 3 and 8 years old.

Researchers have suggested children can learn such abstract categories (at least partly) through the language they hear around them. In fact, cues from language can provide children with information beyond what they can obtain through observation alone (Gelman, 2009; Harris, 2012; Csibra and Gergely, 2009). In particular, word co-occurrence in child-directed speech has been shown to be a reliable cue for several taxonomic categories (Huebner and Willits, 2018; Fourtassi et al., 2019). For example, though "fish" and "bird" do not look very similar, people talk about them in similar linguistic contexts, typically leading to a high co-occurrence-based similarity.

The current study provides (correlational) evidence for this proposal by showing that, at the same time children's associations become more taxonomic (Wojcik and Kandhadai, 2019), they also become more tuned with the word co-occurrence distribution of their native language. I suggest that these are not totally independent developments, and more precisely, that the later could (at least partly) influence the former. If this were to be true, then we could possibly explain a major high-level episode in conceptual development based on a simple mechanism of tracking statistical co-occurrence in language (which we know children are highly skilled at, see Saffran et al. 1996).

Future investigations should go beyond the limitations of the current work. For example, here I predicted data from one set of children (free word association) with data about the experience of a completely different set (CHILDES corpus). Such a research approach has been used before (e.g., Braginsky et al. 2019). It is cost-effective, allowing us to collect large data and average out differences between children and their input. However, this approach fundamentally limits the amount of variability we can capture in terms of how the input may influence uptake. One way to mitigate this limitation is through doing dense data analysis, correlating input and behavior for the same child (e.g., Roy et al. 2015). Another (complementary) approach is to explore if the phenomenon can be produced in controlled, albeit simplified, behavioral experiments (e.g., Unger et al. 2020).

5 Acknowledgement

This research was supported by the grant ANR16-CONV-0002 (ILCB)

References

Mika Braginsky, Daniel Yurovsky, Virginia A Marchman, and Michael C Frank. 2019. Consistency and variability in children's word learning across languages. *Open Mind*, 3:52–67.

Allan M Collins and Elizabeth F Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407.

Virginia S Cronin. 2002. The syntagmatic-paradigmatic shift and reading development. *Journal of Child Language*, 29(1):189.

Gergely Csibra and György Gergely. 2009. Natural pedagogy. *Trends in cognitive sciences*, 13(4).

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1861–1870.

Simon De Deyne and Gert Storms. 2015. Word associations. In *The Oxford Handbook of the Word*.

Haim Dubossarsky, Simon De Deyne, and Thomas T Hills. 2017. Quantifying the structure of free association networks across the life span. *Developmental psychology*, 53(8):1560.

Abdellah Fourtassi, Yuan Bian, and Michael C Frank. 2020. The growth of children's semantic and phonological networks: Insight from 10 languages. *Cognitive Science*, 44(7).

Abdellah Fourtassi, Ewan Dunbar, and Emmanuel Dupoux. 2014. Self-consistency as an inductive bias in early language acquisition. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Abdellah Fourtassi and Emmanuel Dupoux. 2013. A corpus-based evaluation method for distributional semantic models. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 165–171, Sofia, Bulgaria. Association for Computational Linguistics.

Abdellah Fourtassi, Isaac Scheinfeld, and Michael C Frank. 2019. The development of abstract concepts in children's early lexical networks. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 129–133.

Susan A Gelman. 2009. Learning from others: Children's construction of concepts. *Annual review of psychology*, 60.

Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. 2007. Topics in semantic representation. *Psychological review*, 114(2):211.

Paul L Harris. 2012. *Trusting what you're told: How children learn from others*. Harvard University Press.

Philip A Huebner and Jon A Willits. 2018. Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*, 9:133.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press.

Timothy P McNamara. 2005. *Semantic priming: Perspectives from memory and word recognition*. Psychology Press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Brandon C Roy, Michael C Frank, Philip DeCamp, Matthew Miller, and Deb Roy. 2015. Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, 112(41):12663–12668.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of semantic representation with visual attributes. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–582, Sofia, Bulgaria. Association for Computational Linguistics.

Layla Unger, Olivera Savic, and Vladimir M Sloutsky. 2020. Statistical regularities shape semantic organization throughout development. *Cognition*, 198:104190.

Ivan Vulić, Douwe Kiela, and Anna Korhonen. 2017. Evaluation by association: A systematic study of quantitative word association evaluation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 163–175.

Erica H Wojcik and Padmapriya Kandhadai. 2019. Paradigmatic associations and individual variability in early lexical–semantic networks: Evidence from a free association task. *Developmental Psychology*.