Anchoring and the learnability advantage of Zipfian distributions

Tamar Matas*[1], Inbal Arnon[2], and Noam Siegelman[1,2]

*Corresponding Author: tamar.matas@mail.huji.ac.il

[1]Department of Cognitive and Brain Sciences, Hebrew University of Jerusalem, Jerusalem, Israel

[2]Department of Psychology, Hebrew University of Jerusalem, Jerusalem, Israel

Word frequencies across languages follow a Zipfian distribution (Zipf, 1949), where a few words are highly frequent ("the") while the majority are rare ("pancreas"). Studies have found that this Zipfian pattern is not limited to word frequencies, but also applies to letters, phrases, and sign-language signs (Ha et al., 2002; Kimchi et al., 2023; Pande & Dhami, 2010; Piantadosi, 2014), as well as units in other species' communication systems, such as whale songs and dolphin whistles (Arnon et al., 2025; McCowan et al., 1999). What makes this pattern so common across languages and linguistic units, and present in other communication systems in nature?

One explanation previously raised is that Zipfian distributions *facilitate learning*: that is, some properties of this distribution make learning easier. However, the exact mechanism for this facilitation remains debated (Bentz et al., 2017; Ferrer- i-Cancho et al., 2022; Lavi-Rotbain & Arnon, 2022). A possible explanation is the *anchoring hypothesis*, which posits that frequent words are learned first and then serve as anchors for segmenting and learning about the properties of adjacent less frequent words (Valian & Coulson, 1988). In a Zipfian environment, the frequent words provide more anchoring opportunities compared to less skewed distributions, where infrequent words are more likely to appear next to other infrequent words (Kurumada et al., 2013). To date, however, most evidence for the anchoring hypothesis has originated from studies using artificial languages with uniform word distributions, or from studies using very limited samples of natural language. (Bortfeld et al., 2005; Cunillera et al., 2010; Lany, 2014; Shi & Lepage, 2008). Therefore, whether anchoring plays a role in natural language processing remains an open question.

To bridge this gap, in the current study we examine whether the extent of a word's *anchoring* level when it was initially learned (i.e., its extent of proximity to frequent elements in child-directed speech) predicts its future processing ease. We predicted that words that are more anchored during learning should form more robust representations and, therefore, be more easily processed. To test this prediction, we developed an **"**anchoring index**"**, a measure that quantifies a word's occurrence after high-frequency words, which we apply to input from child-directed speech. We then test the effect of the anchoring index on adults' auditory lexical decision reaction times (RTs) as a proxy for the strength of the mental representation. The *anchoring index* was calculated as shown in formula (1), based on CHILDES English corpora (MacWhinney, 2000). Thus, it captures the ratio between a word's number of occurrences after frequent words (i.e., anchors) and its total frequency in the corpus.

$$anchoring\ index = \frac{occurrences\ after\ an\ anchor}{total\ word\ occurrences} \quad (1)$$

In the absence of an a priori frequency threshold for words to count as "anchors," we ran models with anchor parameterizations ranging from the top 0.1% to 10% of the most frequent words, to determine which definition had the highest explanatory power. We then tested each model's predictive power on adults' lexical decision RTs from the Massive Auditory Lexical Decision Database (MALD, Tucker et al., 2019). Using a linear regression model, we examined whether the anchoring index predicts lexical decision RTs while controlling for potential covariates known to affect auditory lexical decision, including frequency, word duration, age of acquisition and others (correlations between anchoring and covariates were low; $|r| < .08$).

As shown in Figure 1, across the vast majority of definitions for words to be considered as anchors, a higher anchoring index (i.e., more appearances next to anchors during learning) significantly predicted faster lexical decision RTs, above and beyond the known

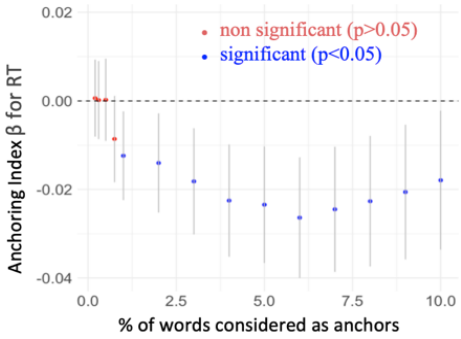

Figure 1. Coefficients for the Anchoring Index in Predicting RTs across anchor definitions ranging from 0.1% to 10%. Error bars represent 95% CI.

covariates. To the best of our knowledge, ours is the first study to link proximity to high-frequency words in naturalistic input to lexical processing. It is also the first study to link anchoring in child-directed input to long-term processing outcomes in adulthood. As such, these findings support the anchoring hypothesis, highlighting anchoring as a possible mechanism supporting the learnability advantage of Zipfian distributions and their prevalence across human languages.

References

Arnon, I., Kirby, S., Allen, J. A., Garrigue, C., Carroll, E. L., & Garland, E. C. (2025). Whale song shows language-like statistical structure. Science, 387(6734), 649–653. https://doi.org/10.1126/science.adq7055

Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i-Cancho, R. (2017). The Entropy of Words—Learnability and Expressivity across More than 1000 Languages. Entropy, 19(6), 275. https://doi.org/10.3390/e19060275

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and Me: Familiar Names Help Launch Babies Into Speech-Stream Segmentation. Psychological Science, 16(4), 298–304. https://doi.org/10.1111/j.0956-7976.2005.01531.x

Cunillera, T., Càmara, E., Laine, M., & Rodríguez-Fornells, A. (2010). Words as Anchors: Known Words Facilitate Statistical Learning. Experimental Psychology, 57(2), 134–141. https://doi.org/10.1027/1618-3169/a000017

Ferrer-i-Cancho, R., Bentz, C., & Seguin, C. (2022). Optimal Coding and the Origins of Zipfian Laws. Journal of Quantitative Linguistics, 29(2), 165–194. https://doi.org/10.1080/09296174.2020.1778387

Ha, L. Q., Sicilia-Garcia, E. I., Ming, J., & Smith, F. J. (2002). Extension of Zipf's law to words and phrases. Proceedings of the 19th International Conference on Computational Linguistics -, 1, 1–6. https://doi.org/10.3115/1072228.1072345

Kimchi, I., Wolters, L., Stamp, R., & Arnon, I. (2023). Evidence of Zipfian distributions in three sign languages. Gesture, 22(2), 154–188. https://doi.org/10.1075/gest.23014.kim

Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. Cognition, 127(3), 439–453. https://doi.org/10.1016/j.cognition.2013.02.002

Lany, J. (2014). Judging Words by Their Covers and the Company They Keep: Probabilistic Cues Support Word Learning. Child Development, 85(4), 1727– 1739. https://doi.org/10.1111/cdev.12199

Lavi-Rotbain, O., & Arnon, I. (2022). The learnability consequences of Zipfian distributions in language. Cognition, 223, 105038. https://doi.org/10.1016/j.cognition.2022.105038

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk: The database, Vol. 2, 3rd ed (pp. viii, 418). Lawrence Erlbaum Associates Publishers.

McCowan, B., Hanser, S. F., & Doyle, L. R. (1999). Quantitative tools for comparing animal communication systems: Information theory applied to bottlenose dolphin whistle repertoires. Animal Behaviour, 57(2), 409–419. https://doi.org/10.1006/anbe.1998.1000

Pande, H., & Dhami, H. S. (2010). Mathematical Modelling of Occurrence of Letters and Word's Initials in Texts of Hindi Language.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic Bulletin & Review, 21(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

Shi, R., & Lepage, M. (2008). The effect of functional morphemes on word segmentation in preverbal infants. Developmental Science, 11(3), 407–413. https://doi.org/10.1111/j.1467-7687.2008.00685.x

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. Behavior Research Methods, 51(3), 1187–1204. https://doi.org/10.3758/s13428-
018-1056-1

Valian, V., & Coulson, S. (1988). Anchor points in language learning: The role of marker frequency. Journal of Memory and Language, 27(1), 71–86. https://doi.org/10.1016/0749-596X(88)90049-6

Zipf, G. K. (1965). Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Ravenio Books.