Skewed distributions facilitate infants’ word segmentation

Citation for published version:
Wolters, L, Ota, M & Arnon, I 2025, 'Skewed distributions facilitate infants’ word segmentation', *Cognition*, vol. 263, 106221. https://doi.org/10.1016/j.cognition.2025.106221

Digital Object Identifier (DOI):
10.1016/j.cognition.2025.106221

Skewed Distributions Facilitate Infants’ Word Segmentation

Lucie Wolters[a], Mitsuhiko Ota[b], Inbal Arnon[c]

[a] **Corressponding author.** Department of Cognitive and Brain Sciences, The Hebrew University of Jerusalem. Mt Scopus, Israel. luciagesine@gmail.com

[b] School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh EH8 9AD, UK. mota@ed.ac.uk

[c] Department of Psychology, The Hebrew University of Jerusalem. Mt Scopus, Israel. inbal.arnon@gmail.com

Abstract: Infants can use statistical patterns to segment continuous speech into words, a crucial task in language acquisition. Experimental studies typically investigate this ability using artificial languages with a uniform frequency distribution, where all words occur equally often. However, words in natural language follow a highly skewed distribution conforming to a Zipfian power law, in which few words occur frequently while many occur infrequently. Prior work shows that such skewed distributions facilitate word segmentation, but the experimental evidence for this has been limited to individuals aged ten years or older, leaving unclear whether this effect arises from accumulated linguistic experience or is already present in the early stages of language learning. To address this, we conducted a word segmentation study with 7- to 9-month-old infants. Infants were exposed to a continuous speech stream containing four artificial words, presented either in a uniform or skewed frequency distribution. We found that infants exposed to the skewed distribution showed a greater looking time difference between familiar and unfamiliar words compared to those in the uniform condition. These findings suggest that skewed distributions facilitate learning during early linguistic development, highlighting the impact of such distributions on language acquisition. Moreover, these findings suggest that the widespread use of uniform distributions in lab-based studies may underestimate infants’ segmentation abilities.

Keywords: skewed distribution, Zipfian distribution, statistical learning, language acquisition, word segmentation, infants

1. Introduction

One fundamental feature of language is that word frequencies follow a highly skewed distribution. This distribution conforms to a Zipfian or near-Zipfian power law, where few words occur very frequently and many occur infrequently (Piantadosi, 2014; Zipf, 1949). Since Zipf's initial observation, this pattern has been found across different languages, linguistic modalities, speech registers, and parts of speech (Bentz et al., 2017; Ferrer i Cancho, 2005; Kimchi et al., 2023; Lavi-Rotbain & Arnon, 2023; Mehri & Jamaati, 2017). Notably, such distributions also characterize the learning environment infants are exposed to. Both the words they hear (Lavi-Rotbain & Arnon, 2023) and the objects they see (Clerkin et al., 2017) follow a Zipfian distribution. In a large corpus study, Lavi-Rotbain and Arnon (2023) found that child-directed speech (CDS) follows a Zipfian distribution across fifteen languages from seven language families, both overall and for different parts of speech. Moreover, they found that CDS exhibits Zipfian distributions when directed to infants as young as six months old, with this pattern remaining stable across development. These findings suggest that even though CDS differs from adult-directed speech in many respects, its word frequency distribution is consistently skewed, forming a prominent characteristic of the linguistic input infants learn from.

The prevalence of Zipfian distributions in language has motivated research into the effects of such distributions on learning. Experimental studies show that Zipfian distributions confer a learnability advantage in various language learning tasks (Goldberg et al., 2007; Hendrickson & Perfors, 2019; Lavi-Rotbain & Arnon, 2022; Wolters et al., 2024). Much of this research has focused on word segmentation, the process of identifying word boundaries in speech. Since spoken language is continuous—often lacking clear acoustic boundaries between words—word segmentation is one of the first challenges infants face during language acquisition. A large number of studies have explored how infants might succeed at this task by relying on statistical patterns in speech. In these studies, infants are typically exposed to a continuous speech stream where the only cues to word boundaries are the statistical regularities with which syllables co-occur (e.g. Saffran et al., 1996). Recent work has investigated the impact of skewed frequency distributions, like the Zipfian distribution, on this process. Findings indicate that both children and adults are better at word segmentation when word frequencies follow a skewed distribution as opposed to a uniform one, where all words appear equally often (Kurumada et al., 2013; Lavi-Rotbain & Arnon, 2019, 2022). Similar facilitative effects of skewed distributions have been reported for visual statistical learning (Lavi-Rotbain & Arnon, 2021), cross-situational word learning (Hendrickson & Perfors, 2019), word-referent mapping (Wolters et al., 2024), and learning novel grammatical categories (Goldberg et al., 2007; Wonnacott et al., 2017).

However, until now, experimental studies demonstrating the facilitative effect of skewed distributions on learning have been conducted with children ten years and older, who have already had extensive experience with language. Therefore, it is not clear whether this effect arises from accumulated linguistic experience with such distributions or whether it is a cognitive disposition that is already present in the early stages of language acquisition. The current study addresses this gap by investigating whether word segmentation by infants is also facilitated by a skewed frequency distribution. If facilitation from skewed distributions is a driver of language acquisition, rather than a product of learners' linguistic experience, then we would expect infants to show improved segmentation when exposed to such distributions.

To investigate this, we compare infants' word segmentation performance when exposed to a skewed frequency distribution versus a uniform one. We use a word segmentation paradigm (Saffran et al., 1996) where infants are first familiarized with a novel continuous speech stream and then tested on their ability to discriminate familiar words from unfamiliar ones using a central fixation procedure (Cooper & Aslin, 1990). If skewed distributions facilitate word segmentation in infants, we expect those exposed to a skewed distribution to show stronger evidence for discriminating familiar words from unfamiliar words compared to those exposed to a uniform distribution.

2. Method

2.1. Participants

We recruited typically developing infants aged 7 to 9 months (range: 210 days to 301 days) through the browser-based platform Lookit, which enables remote participation via webcam. Participants were compensated with a $5 Amazon gift card. In total, 95 participants with a valid consent video (confirmed by the research team) completed the study on Lookit. Data from infants with duplicate participation was excluded (N=4). The remaining data was reviewed independently by three members of the research team (who were naive to the trial types). Participants whose data were deemed problematic by at least two members were excluded from the analysis (N=30, 17 in the skewed condition; 13 in the uniform condition). Reasons for exclusion were: caregivers or siblings interfering with the infant's looking behavior (N=11), failure to complete all trials (N=5), fussiness/crying (N=5), technical failure (N=4), and poor video quality (N=5). The final dataset included 61 infants, 29 in the skewed condition and 32 in the uniform condition (age range: 210–290 days; mean age: 236 days; 32 females and 29 males; ethnically, they were 56.7% white, 23.3% mixed, 10% Asian, 3.3% Hispanic, Latino or

Spanish origin, and 1.7% black or African American; 85% of the caregivers had an undergraduate degree; a further 60% had a graduate degree).

2.2. Experimental conditions: familiarization speech stream

Infants were exposed to a synthesized speech stream consisting of multiple repetitions of four artificial tri-syllabic words: 'kibeto', 'dukame', 'genodi', and 'nalubi'. They heard one of two exposure streams, featuring either a uniform or a skewed frequency distribution (see Table 1). The speech streams for both experimental conditions were taken from a previous study testing the facilitative effect of Zipfian distributions on word segmentation in adults and children by Lavi-Rotbain & Arnon (2022). The syllables were synthesized using PRAAT (Boersma, 2001), and were matched on pitch (~76 Hz), amplitude (~60 dB), and duration (250–350 ms) (for more details see Lavi-Rotbain. & Arnon, 2019). Lavi-Rotbain & Arnon (2022) compared two skewed distributions with different entropy levels: a distribution with an entropy level comparable to that found in natural language, referred to as the "language-like" condition, and a distribution with entropy levels higher than what is found in natural language, and therefore less predictable, referred to as the "reduced-entropy" condition. In the current study, the skewed distribution corresponds to the "reduced-entropy condition" in Lavi-Rotbain & Arnon (2022), rather than the language-like one. This was done to ensure that infants receive sufficient exposure to low-frequency items, which appear only nine times in the language-like condition but 19 times in the reduced condition.

Exposure streams for both conditions contained 128 word tokens and lasted 1 minute and 50 seconds. In the uniform condition, all four words occurred 32 times. In contrast, in the skewed condition, one word, 'nalubi', occurred with a high frequency of 71, and each of the remaining three words occurred with a low frequency of 19. Note that unlike word frequency distributions in natural language, all low frequency words appeared the same number of times. This was done to control for item frequency effects during testing. Importantly, there were no acoustic breaks between words, nor any prosodic or co-articulation cues to indicate word boundaries. The only available cue was the transitional probabilities (TPs) from one syllable to the next: TPs between words were lower compared to TPs within words. The words were presented in a randomized order in the familiarization streams, with the only constraint being that the words in the uniform condition and the infrequent words in the skewed condition could not repeat themselves.

Table 1: Experimental conditions.

	Uniform condition	Skewed condition
Exposure length [seconds]	110	110
Number of word tokens	128	128
Tokens per word type	32 (all 4 words)	Frequent (1 word): 71 Infrequent (3 words): 19
Unigram entropy [bits]	2	1.7
TP's within words	1	1
TP's between words	0.33	Frequent: 0.64 Infrequent: 0.36

2.3. Test stimuli

Infants were tested on their recognition of two familiar words, 'dukame' and 'genodi', which had appeared in the familiarization speech stream, and two unfamiliar words, 'kinome' and 'gekato', which had not. The familiar words were selected from the low-frequency words in the skewed condition, meaning infants in this condition heard them only 19 times during familiarization, compared to 32 times in the uniform condition. The unfamiliar words were constructed by taking one syllable from each infrequent word of the skewed condition ('dukame', 'genodi', and 'kibeto'), preserving their original position in the word (beginning, middle, or end). This ensured that infants were paying attention to whole words rather than syllable positions in the sequence. Many word segmentation studies compare recognition of familiar words to part-words, composed of two syllables from a familiar word, and one syllable from another word. Such part-words are useful for investigating learners' sensitivity to the transitional probabilities in their input, since they do appear during exposure, but with lower transitional probabilities (e.g., Saffran et al. 1996). Here, we use unfamiliar words instead of part-words because we are primarily interested in how the frequency distribution impacts whole unit recognition. Prior studies looking at the effect of skewed distributions on word segmentation also used unfamiliar words, for similar reasons (e.g., Lavi-Rotbain & Arnon, 2022). To ensure that the familiar and unfamiliar test words were equally similar to English words, their 'wordlikeness' was judged by adult native speakers of English on a 7-point Likert scale. The familiar and unfamiliar words had similar word-likeness ratings (unfamiliar mean = 3.10; sd by word = 1.7; familiar mean score = 3.05; sd by word = 1.7). See Appendix A for more details.

2.4. Procedure

The experiment was conducted on the browser-based platform Lookit. Caregivers signed up for our experiment via the platform and completed the experiment with their children from a preferred location while data was collected with the webcam, usually the built-in camera of a laptop or desktop owned by the caregiver. Caregivers were instructed to position their baby on their lap, ensure their baby's entire face was visible, and prevent shadows from covering the baby's eyes. They were also asked to keep their eyes closed throughout the experiment and to avoid directing their baby's attention in any way.

The experiment began with a familiarization phase, during which infants were played one of the familiarization speech streams described in Section 2.2, while a multicolored checkerboard was displayed at the center of the screen. Next, infants were tested on their recognition of the four test words using a central fixation procedure. During each test trial, the infant heard ten repetitions of one of these words with an interstimulus interval of 500 ms while a spinning color wheel was displayed in the center of the screen. Each trial lasted approximately 15 seconds.

The testing phase consisted of two blocks, each containing trials for the two familiar and two unfamiliar words (four trials in total). Familiar and unfamiliar words were presented in a semi-randomized order, alternating familiar and unfamiliar words across the two blocks. To grab the infant's attention, each test trial was preceded by a spinning image of Elmo accompanied by an audio of giggling children.

The webcam turned itself on at the onset of the exposure trial and each of the test trials, creating a video recording in .mp4 format, which was used to analyze the infant's gaze fixation.

3. Results

Video data of the familiarization phase and the eight test trials was manually annotated using ELAN (2024) by five trained coders who determined frame-by-frame whether the infant's gaze was fixated on the visual display. The coded data was analysed using linear regression models, implemented using the lme4 library (Bates et al., 2015) in R (R Core Team, 2023).

The average looking time during familiarization was 43.8 seconds (sd =19.1) overall, with 47.9 seconds (sd = 20.1) in the skewed condition and 40.0 seconds (sd =17.5) in the uniform condition. We fitted a model with total looking time during familiarization as the outcome variable, a fixed effect of condition (factor, two levels, dummy coded with skewed condition as reference), and a random intercept for participants. We found no significant difference in the total looking time during familiarization between the two conditions (β = -7.8, SE = 4.8, t = -1.6, p = .11).

The critical variable was the total duration of central fixation in each test trial. Figure 1 displays the mean total looking time for individuals by trial type (familiar or unfamiliar) in the two conditions. On average, the total looking time was longer for unfamiliar word trials (mean = 10.02 sec, sd = 3.5) than for the familiar word trials (mean = 9.73 sec, sd = 3.5). This pattern was found individually in 38 out of the 61 infants (20 in the skewed condition and 18 in the uniform condition).

To ensure that our results were not disproportionately influenced by extreme values, we used the interquartile range (IQR) to check whether there were any statistical outliers in terms of the difference between the two trial types, operationalized as the mean total looking time of unfamiliar trials minus that of familiar trials. The IQR was calculated as the difference between the 25th percentile (Q1) and the 75th percentile (Q3) of this measure. Infants with a difference falling outside the interquartile range defined by Q1 – 1.5 x IQR (–3.0 s) and Q3 + 1.5 x IQR (3.7 s) were classified as outliers and their data was excluded from further analysis. One infant was identified as an outlier, with a difference of –3.62 seconds, and their data was excluded from the analyses[1].

To test the effect of trial type and condition on the total looking time during test trials, we fitted a model with total looking time on trial as the outcome variable and fixed effects of trial type (factorial, two levels, dummy coded with familiar trials as reference), condition (factorial, two levels, dummy coded with skewed condition as reference), the interaction between trial type and condition, and trial number (numerical 1-8). The model included by-participant random slopes for trial number and trial type[2]. Infants exhibited significantly longer looking times for unfamiliar words compared to familiar words (β = .80, SE = 0.24, t = 3.3, $p < .01$). Importantly, infants who were exposed to a skewed distribution showed a greater looking time difference between familiar and unfamiliar words (mean = 0.8 s, sd = 1.1) compared to those in the uniform condition (mean = -0.02 s, sd = 1.4), indicated by a significant interaction between condition and trial type (β = .82, SE = 0.33, t = 2.5, $p < .05$). In addition, we found a significant effect of trial number (β = -.86, SE = 0.05, t = -18.7, $p < .001$), indicating that infants' total looking times decreased over trials. These findings suggest that infants are better at discriminating between familiar and unfamiliar words when exposed to a skewed distribution.

To determine whether infants in both conditions show a significant looking time difference between familiar and unfamiliar words, we analyzed the data for each condition separately,

[1] Analyses including outliers can be accessed on the OSF page of the study: https://osf.io/fk2s7/?view_only=1a6978ac0fe347a1b85335de9b41a53c

[2] We did not include a random effect for stimuli (factor, 4 levels) because there are only two stimuli per trial type, resulting in high collinearity between the random intercepts for stimuli and the fixed effect of condition.

excluding the fixed effect of condition. We found no significant effect of trial type in the uniform condition (**β** = -.02, SE = 0.24, t = -0.095, p = .93), whereas the effect was significant in the skewed condition (**β** = .80, SE = 0.24, t = 3.36, $p < .001$). That is, despite the familiar words occurring far less frequently in the skewed condition than in the uniform condition (19 vs. 32 occurrences), infants exposed to the skewed condition successfully discriminated between familiar and unfamiliar words, whereas those in the uniform condition did not.

In order to further explore the effect of skewed distributions on the behaviour of infants, we carried out a follow-up analysis with age (in days) and sex (male/female) included in the model. The results showed no effects of age or sex while trial type, condition and the interaction between them remained significant[3].

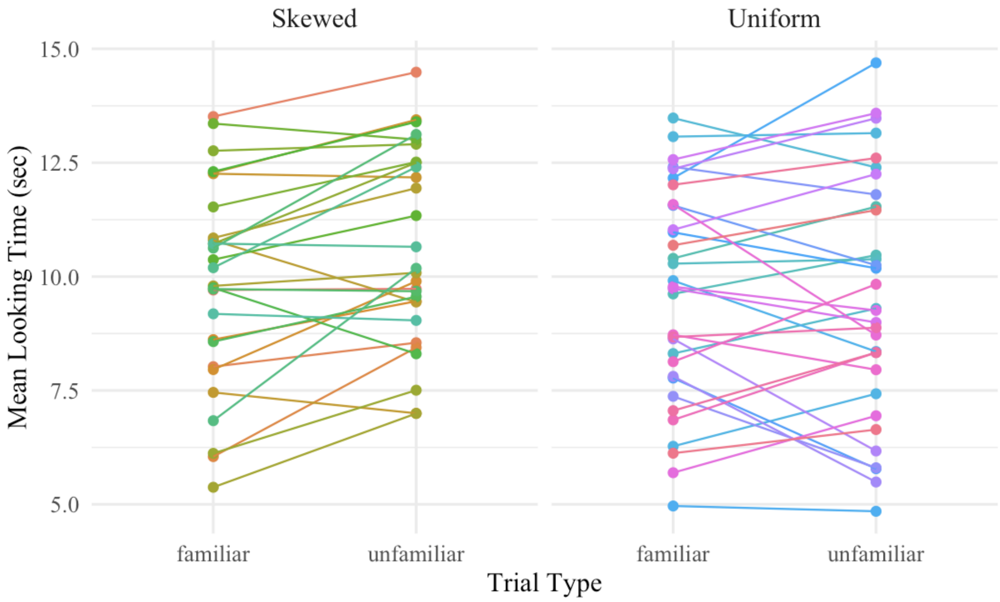

Figure 1: Mean looking time (seconds) over trials. The lines connect individual participants' mean looking times for familiar and unfamiliar trials.

4. Discussion

The words infants hear when learning to segment speech follow a highly skewed frequency distribution. Previous research has demonstrated that such distributions facilitate word segmentation in school-aged children and adults (Lavi-Rotbain & Arnon, 2019, 2022). In this study, we asked whether infants also benefit from skewed distributions during word segmentation. Our findings indicate that infants who were exposed to a skewed word frequency showed a significantly larger looking time difference between familiar and unfamiliar words,

[3] Full analysis and results can be accessed on the OSF page of the study: https://osf.io/fk2s7/?view_only=1a6978ac0fe347a1b85335de9b41a53c

compared to those who were exposed to a uniform distribution. This effect occurs despite infants in the skewed condition hearing the critical test words only 19 times during exposure, compared to 32 times in the uniform condition. This suggests that, like older children and adults, infants' word segmentation is facilitated when they are exposed to a skewed frequency distribution, and that the characteristic frequency distribution of natural language facilitates language learning during early linguistic development.

Separate analyses of the looking behaviour in the two conditions indicate there was no significant difference between the looking time for familiar and unfamiliar items in the uniform condition, whereas there was in the skewed condition. These results contrast with previous segmentation studies using uniform distributions with infants at this age, and finding evidence for learning (for a review see Saffran & Kirkham, 2018). One possible explanation is that the frequency of word occurrences in the uniform condition is lower than in previous infant segmentation studies. In the current study, each word appeared 32 times, whereas prior studies typically have each word appear 45 times or more (e.g. Saffran et al., 1996; Thiessen et al., 2005). If additional exposure is necessary for learning in a uniform distribution, this would make the successful learning of words appearing only 19 times in the skewed distribution even more striking. Another possible explanation is the use of a browser-based experimental procedure, where caregivers and infants participate from their homes. Compared to lab experiments, browser-based procedures offer less control over the testing environment, introducing more distractions and more noise in the data. The additional distractions could be more detrimental in more challenging learning conditions, like the uniform distribution. Future work is needed to determine the extent to which the experimental setting (browser-based vs. in-lab) impacts learning from skewed and uniform distributions.

An additional difference between our study and other word segmentation studies (e.g., Saffran et al. 1996) is that we compared the discrimination between familiar and unfamiliar words, rather than familiar words and part-words (a sequence where the first two syllables belong to one word, and the last syllable to another). The unfamiliar words consisted of syllables that were part of the input but had never appeared as a sequence during exposure. The use of unfamiliar words means that we cannot draw any conclusions about infants' sensitivity to transitional probabilities, but only about their ability to extract whole word forms from the input. Interestingly, infants are capable of distinguishing between familiar words and part-words based on their TP differences, even when those differences are learnt from a non-uniform distribution: Aslin, Saffran & Newport (1998) exposed infants to an artificial language where two of the words appeared 90 times, and the other two appeared only 45 times. This was done to ensure that the part-words used in testing were as frequent as the real words. Infants showed learning, indicating they are sensitive to the TPs when learning from a non-uniform distribution. However, we do not yet know how TP sensitivity is impacted by distribution shape, a question that can be tested using the same conditions we used, but with part-words as foils.

Our finding of an effect in the skewed condition, but not in the uniform condition, highlights methodological implications for infant research more generally. Experimental designs typically use uniform frequency distributions to control for frequency effects on learning. However, this approach may underestimate infants' segmentation abilities, as their performance improves when exposed to a skewed distribution that more closely resembles natural language. This insight extends beyond word segmentation to other areas of language learning, such as cross-situational learning and word learning, where frequency distributions play a crucial role. For instance, the objects infants encounter during learning also follow highly skewed distributions (Clerkin et al., 2017; Lavi-Rotbain & Arnon, 2021). These non-uniform distributions of words and objects may affect cross-situational word learning, where infants track word-referent pairings across multiple exposures (Hendrickson & Perfors, 2019; Kachergis et al., 2017). Similarly, in word learning paradigms, the probability of encountering words in a Zipfian distribution may affect word recall or generalization. More broadly, our findings align with work illustrating how the statistical structure of infants' learning environment changes throughout development (Smith et al. 2018), and highlight the importance of using more ecologically valid frequency distributions to gain a more accurate understanding of infants' language learning capabilities.

Why do skewed distributions facilitate word segmentation in infants? Several hypotheses have been proposed. One suggestion is that the high frequency words are identified quickly, allowing them to provide contextual cues for segmenting adjacent low frequency words (Kurumada et al., 2013). Indeed, infants show improved segmentation of novel words when they are presented adjacent to familiar ones (Bortfeld et al., 2005), suggesting that once infants are familiar with the high frequency items in the Zipfian distribution, they can use these items as contextual cues. The facilitative effect has also been attributed to the lower entropy of skewed distributions, which make them more predictable than uniform distributions, and help learners form predictions about upcoming words in the speech stream (Lavi-Rotbain & Arnon, 2022; Wolters et al., 2024). While no studies have directly tested the effect of entropy on segmentation in infants, 7- to 8-month-old infants were found to preferentially attend to auditory stimuli with intermediate levels of predictability, rather than stimuli that are either highly predictable or highly unpredictable, within the same auditory sequence (Kidd et al., 2014). Since all items in the uniform distribution appear with equal frequency, they are all equally unpredictable. In contrast, in the skewed distribution, one item appears more frequently than the others, resulting in a larger range of predictability, including items of intermediate predictability. In this context, the lower entropy of the skewed distribution (compared to uniform one) may have led infants to attend to it more closely during familiarization, thereby enhancing segmentation. Consistent with this idea, infants in the skewed condition looked on average eight seconds longer during familiarization, even though this difference was not statistically reliable. If skewed distributions enhance attention to the stream in general, or to the low frequency words, we may see this reflected in the online looking times for the different items during familiarization.

Prior work with children and adults has investigated the independent contribution of several features of Zipfian distributions, by comparing segmentation across multiple distributions with different shapes and entropy levels (Lavi-Rotbain & Arnon, 2022). Distribution shape did not affect performance: segmentation was similar in distributions where each infrequent word occurred with equal frequency (as in the current study) compared to ones where word frequencies followed a power law distribution. However, entropy did impact performance: learners segmented words more effectively when the distribution entropy matched the entropy levels found in natural language but not when entropy levels were higher than those of natural language, as in the skewed condition of the present study. That is, unlike children and adults, infants showed facilitation in distributions with higher entropy than natural language. Further work is needed to see whether infants also benefit more from language-like entropy levels.

Another open question is whether the facilitative effect of skewed distributions in infants results from accumulated exposure to such distributions in early life or whether it is driven by a more general cognitive disposition for such distributions (Shufaniya & Arnon, 2021). Since 7- to 9-month-old infants have not yet fully segmented the speech they encounter in their environment, it is unlikely that their improved segmentation in skewed distributions is solely driven by experience with natural language. However, infants are exposed to other skewed distributions in their learning environment. The objects children see tend to follow a Zipfian distribution (Clerkin et al., 2017; Lavi-Rotbain & Arnon, 2021), and power-law distributions are common in the physical world (Newman, 2005). One possibility is that infants attune to these distributional patterns in non-linguistic domains and subsequently transfer this knowledge to language learning. However, while prior studies have shown that infants acquire language-dependent segmentation strategies over development (e.g. Mersad et al., 2011; Onnis & Thiessen, 2013), we are not aware of any studies that show infants transfer (distributional) knowledge from non-linguistic domains to language learning. Furthermore, although Zipfian distributions are common, they are not the only type of distribution present in infants' learning environments. This gap in research can be addressed by exploring whether the facilitative effect of skewed distributions in infants extends to other learning domains (e.g., visual learning) and different learning tasks (e.g., visual expectation/prediction).

Finally, our findings have implications for broader theories on the emergence of Zipfian distributions in language. Regardless of its exact origins, our findings show that infants' language acquisition benefits from skewed distributions early on. It has been proposed that language evolves to be learnable by being culturally transmitted from one generation of learners to the next (Griffiths et al., 2008; Kirby et al., 2008, 2014), and that this process can drive the emergence of Zipfian frequency distributions, as they facilitate segmentation (Arnon et al., 2025; Arnon & Kirby, 2024). Our findings from young infants support this learnability hypothesis by showing that this facilitative effect of skewed distributions is already observable in young infants at the cusp of word learning.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

Arnon, I., & Kirby, S. (2024). Cultural evolution creates the statistical structure of language. *Scientific Reports*, *14*(1), 5255.

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

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324. https://doi.org/10.1111/1467-9280.00063

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), Article 6. https://doi.org/10.3390/e19060275

Boersma, Paul (2001). Praat, a system for doing phonetics by computer. Glot International, 5, 341-345.

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

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372:20160055*. https://doi.org/10.1098/rstb.2016.0055

Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development*, *61*(5), 1584–1595. https://doi.org/10.1111/j.1467-8624.1990.tb02885.x

Ferrer i Cancho, R. (2005). The variation of Zipf's law in human language. *The European Physical Journal B - Condensed Matter and Complex Systems*, *44*, 249–257. https://doi.org/10.1140/epjb/e2005-00121-8

Goldberg, A. E., Casenhiser, D., & White, T. R. (2007). Constructions as categories of language. *New Ideas in Psychology*, *25*(2), 70–86. https://doi.org/10.1016/j.newideapsych.2007.02.004

Griffiths, T. L., Kalish, M. L., & Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1509), 3503–3514. https://doi.org/10.1098/rstb.2008.0146

Hendrickson, A. T., & Perfors, A. (2019). Cross-situational learning in a Zipfian environment. *Cognition*, *189*, 11–22. https://doi.org/10.1016/j.cognition.2019.03.005

Kachergis, G., Yu, C., & Shiffrin, R. M. (2017). A bootstrapping model of frequency and context effects in word learning. *Cognitive Science*, *41*(3), 590–622. https://doi.org/10.1111/cogs.12353

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2014). The Goldilocks effect in infant auditory attention. *Child Development*, *85*(5), 1795–1804. https://doi.org/10.1111/cdev.12263

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

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686. https://doi.org/10.1073/pnas.0707835105

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114. https://doi.org/10.1016/j.conb.2014.07.014

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

Lavi-Rotbain, O., & Arnon, I. (2019). Children learn words better in low entropy. *Cognition*, *223*, 1–15.

Lavi-Rotbain, O., & Arnon, I. (2021). Visual statistical learning is facilitated in Zipfian distributions. *Cognition*, *206*, 104492. https://doi.org/10.1016/j.cognition.2020.104492

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

Lavi-Rotbain, O., & Arnon, I. (2023). Zipfian distributions in child-directed speech. *Open Mind*, *7*, 1–30.

Mehri, A., & Jamaati, M. (2017). Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations. *Physics Letters A*, *381*(31), 2470–2477. https://doi.org/10.1016/j.physleta.2017.05.061

Mersad, K., Goyet, L., & Nazzi, T. (2011). Cross-linguistic differences in early word form segmentation: A rhythmic-based account. *Journal of Portuguese Linguistics*, *10*(1), Article 1. https://doi.org/10.5334/jpl.100

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351. https://doi.org/10.1080/00107510500052444

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition*, *126*(2), 268–284. https://doi.org/10.1016/j.cognition.2012.10.008

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

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*(1), 181–203.

Shufaniya, A., & Arnon, I. (2021). A cognitive bias for Zipfian distributions? Uniform distributions become more skewed via cultural transmission. *Proceedings of the*

Annual Meeting of the 43th Cognitive Science Society. https://escholarship.org/uc/item/66c40451

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences*, *22*(4), 325–336. https://doi.org/10.1016/j.tics.2018.02.004

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*(1), 53–71. https://doi.org/10.1207/s15327078in0701_5

Wolters, L., Lavi-Rotbain, O., & Arnon, I. (2024). Zipfian distributions facilitate children's learning of novel word-referent mappings. *Cognition*, *253*, 105932. https://doi.org/10.1016/j.cognition.2024.105932

Wonnacott, E., Brown, H., & Nation, K. (2017). Skewing the evidence: The effect of input structure on child and adult learning of lexically based patterns in an artificial language. *Journal of Memory and Language*, *95*, 36–48. https://doi.org/10.1016/j.jml.2017.01.005

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.

Appendix A: 'Wordlikeness' study of the test stimuli

Eleven artificial words were tested for their wordlikeness to English: the familiar words in the familiarization speech streams ('dukame', 'nalubi', 'kibeto' and 'genodi'), the unfamiliar words used for testing ('kinome' and 'gekato'), and six optional unfamiliar words ('dunobi', 'gekabi', 'geluto', 'kilume', 'kinome', 'nabedoi'). We wanted to test whether the words used for testing in the main experiment are considered possible English words to the same extent, so that we know no words in the experiment will stand out by this measure.

1. Participants

20 participants participated in the experiment on the browser-based platform Prolific. All participants were screened to have English as their first language. The experiment lasted five minutes and participants were paid at a nine pound hourly rate.

2. Method

Participants were told they would hear made-up words, and that for each word, they had to indicate how likely it is for the word to be a word of English. At each trial–one for each word–participants heard the audio of the word once, and had to indicate their answer on the 7-point Likerd scale displayed in figure 1.

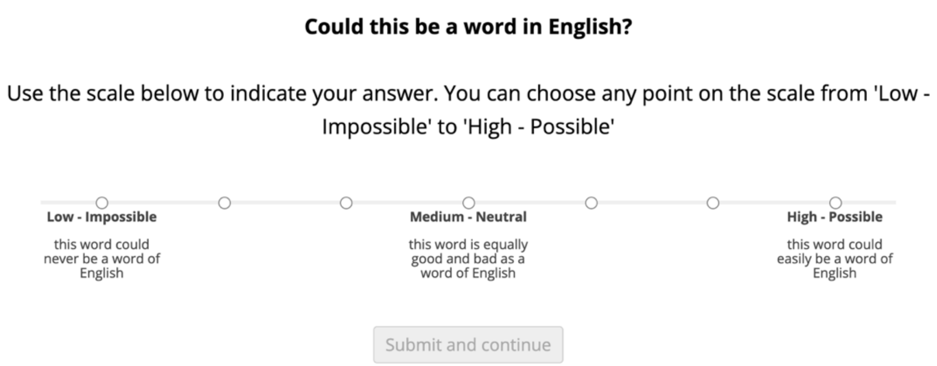

Figure 1: Example screen of one trial in the experiment.

3. Results

We only analyzed the words used in the test phase of the main experiment: 'dukame', 'genodi', 'kinome' and 'gekato'. To assess the effect of stimulus on the responses of the participants, we fitted a linear mixed effects model (Bates, Mächler, Bolker, & Walker, 2015) in R (R Core

Team, 2023) with response per stimulus as the outcome variable, a fixed effect for stimulus type (familiar or unfamiliar), and a random intercept for participants. Stimulus type did not significantly predict the responses of participants, indicating that familiar and unfamiliar words were judged to be equally similar to English (unfamiliar mean = 3.1; sd by word = 1.7; familiar mean score = 3.05; sd by word = 1.7) (**β** = -0.27, SE = 0.19, t = -1.4, p = .16).