OPEN MIND
Discoveries in Cognitive Science

an open access journal

REPORT

Greater Early Disambiguating Information for Less-Probable Words: The Lexicon Is Shaped by Incremental Processing

Adam King[1] and Andrew Wedel[1]

[1]Department of Linguistics, University of Arizona

Keywords: language efficiency, Zipf's law of abbreviation, incremental processing, language evolution, information theory

ABSTRACT

There has been much work over the last century on optimization of the lexicon for efficient communication, with a particular focus on the form of words as an evolving balance between production ease and communicative accuracy. Zipf's law of abbreviation, the cross-linguistic trend for less-probable words to be longer, represents some of the strongest evidence the lexicon is shaped by a pressure for communicative efficiency. However, the various sounds that make up words do not all contribute the same amount of disambiguating information to a listener. Rather, the information a sound contributes depends in part on what specific lexical competitors exist in the lexicon. In addition, because the speech stream is perceived incrementally, early sounds in a word contribute on average more information than later sounds. Using a dataset of diverse languages, we demonstrate that, above and beyond containing more sounds, less-probable words contain sounds that convey more disambiguating information overall. We show further that this pattern tends to be strongest at word-beginnings, where sounds can contribute the most information.

BACKGROUND

Human languages are characterized by hierarchically organized, nested structure: utterances are composed of structured sequences of words, and words in turn are composed of structured sequences of sounds. Many of the ways these structures are organized in language have been argued to result in more efficient transmission of information than would occur otherwise (e.g., Fedzechkina, Jaeger, & Newport, 2012; Ferrer i Cancho, 2017; Futrell, Mahowald, & Gibson, 2015; Genzel & Charniak, 2002; Gibson et al., 2019; Gildea & Jaeger, 2015; Hale, 2003, 2006; Hawkins, 2010; Jaeger, 2010; Jaeger & Tily, 2011; Levy, 2008), suggesting that the details of language structures evolve under pressure to optimize communicative efficiency. The lexicon—roughly, the set of words in a language—is one of these possible loci of optimization, and can be conceptualized as a code that maps meaningful lexical units (referred to here as *words*) to word-forms, for example, a sequence of sounds, or *segments*. The relationship between words and their word-forms is not fixed a priori, but can evolve over the course of language change, such as when the original compound *electronic mail* shortened to *email* with increasing use. Because the lexicon is a constantly evolving system, and because many lexical

Citation: King, A., & Wedel, A. (2020). Greater Early Disambiguating Information for Less-Probable Words: The Lexicon Is Shaped by Incremental Processing. *Open Mind: Discoveries in Cognitive Science*, *4*, 1–12. https://doi.org/10.1162/opmi_a_00030

DOI:
https://doi.org/10.1162/opmi_a_00030

Supplemental Materials:
https://www.mitpressjournals.org/doi/suppl/10.1162/opmi_a_00030

Received: 31 May 2019
Accepted: 19 December 2019

Competing Interests: The authors declare no conflict of interest.

Corresponding Author:
Adam King
adamking@email.arizona.edu

The MIT Press

properties of interest—such as word length—can be straightforwardly measured, the lexicon has been a focus for much prior research on the role of biases toward efficient communication in shaping language patterns. Many of these studies conclude that patterns in the lexicon support the hypothesis that communicative efficiency is a driving pressure in the evolution of word to form mappings (Ferrer i Cancho & Solé, 2003; Kanwal, Smith, Culbertson, & Kirby, 2017; Piantadosi, Tily, & Gibson, 2009, 2012; Zipf, 1949).

One of the most cross-linguistically robust observations in this domain is Zipf's law of abbreviation: more-probable words tend to be shorter, while words that are less probable tend to be longer (Bentz & Ferrer i Cancho, 2016; Piantadosi, Tily, & Gibson, 2011; Zipf, 1935). In his *Principle of Least Effort*, Zipf (1949) proposed that this pattern arises as a trade-off between a pressure for accuracy on the one hand, and lower effort on the other. As it stands, there is robust evidence that the segment composition of word-forms is shaped for lower production-effort beyond the effect of short length (Dautriche, 2015; Dautriche, Mahowald, Gibson, & Piantadosi, 2017; Mahowald, Dautriche, Gibson, & Piantadosi, 2018; Meylan & Griffiths, 2017). Here we investigate whether segment composition may also be optimized to provide listeners greater disambiguating information as they identify words in the speech stream.

We can conceptually divide the information available to listeners in word identification into two sources: (a) the listener's prior expectation that the word will occur, and (b) the information provided by the word-form itself (reviewed in Hall, Hume, Jaeger, & Wedel, 2016). If word-forms evolve under pressure to balance accuracy and effort, the amount of information from these two sources should tend to trade-off: words that are on average more probable should evolve word-forms that contain less informative material because they can do so without compromising accuracy, and conversely, words that are less probable should evolve word-forms that convey relatively more information.

All things being equal, a word-form with more segments is likely to possess more information overall. However, segments can differ in how much information they contribute to disambiguating a word-form from others: a segment in a word-form that disambiguates from many other forms in the lexicon provides more information than one that disambiguates from few. Further, earlier segments in a word-form tend to contribute more disambiguating information in word identification than later segments because listeners process word-forms incrementally, progressively updating inferences about the intended word as the segment sequence unfolds in time (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Magnuson, Dixon, Tanenhaus, & Aslin, 2007; Marslen-Wilson, 1987; see Dahan & Magnuson, 2006, and Weber & Scharenborg, 2012, for review). For example, consider the word *vacuum* /vækjum/. Perception of the word-initial [v] is highly informative as it allows a listener to begin to discount the large set of lexical items that begin with other segments. The final [m] contributes less information because the previous segments [vækju...] already indicate *vacuum* as the most likely word. Correspondingly, psycholinguistic studies show that listeners preferentially allocate attention to word-beginnings, which has been attributed to the greater information provided by early segments (Connine, Blasko, & Titone, 1993; Grosjean, 1996; Marslen-Wilson & Zwitserlood, 1989; Nooteboom, 1981; Salasoo & Pisoni, 1985).

Two related predictions for efficient lexical structure arise from the fact that different segments can convey different amounts of disambiguating information. First, words that are on average less probable should tend to not only have *longer* forms, but to have forms with relatively *higher information* segments. Second, if the lexicon is structured to capitalize on incremental word processing, this association between segmental information and word probability should be strongest early in word-forms and decay at later positions. Early segments more strongly narrow the range of lexical possibilities, and in parallel, narrow the prior contextual differences in word probability (for more on the benefit of early informativeness, see Hawkins, 2010). A useful way to think of this second prediction is that segments should become distributed throughout the lexicon such that the probability mass of competing words drops more steeply for less-probable words during processing. This can potentially be achieved with two conceptually distinct strategies, one focusing on the segmental *network structure* of the lexicon, that is, the specific sequences of segments that distinguish words, and the other on the relative word probabilities within the competing groups of words that exist.

In the first strategy, the lexicon evolves such that the segments in less-probable words act to disambiguate from a greater *number* of competing words early on. As an example, the form for the less-probable word *sphinx* begins with a nearly unique cluster, [sf], which immediately disambiguates it from most of the lexicon. In the second, the lexicon evolves such that less-probable words have segments that disambiguate from relatively high-probability competitors. Both of these strategies have the effect of reducing the probability mass of competitors faster for less-probable words.

Here, we explore these predictions, showing evidence that the lexicons of a diverse set of languages are in fact structured to be an efficient code given incremental processing, both in terms of the structure of the lexicon and the relative probabilities of competitors in that structure.

METHODS

We investigated the relationship between segment information and word probability in phonemically transcribed corpora for 20 languages (see Table 1 in the Supplemental Materials; King & Wedel, 2020). The dataset is reasonably typologically diverse, drawn from 10 different language families from four continents, where 13 of the 20 languages are non-Indo-European. All corpora except Hausa, Kaqchikel, Malay and Tagalog were morphologically annotated, allowing us to focus the analysis on uninflected word stems. For each language, we limited our investigation to the 10,000 most frequent word-types.

We used a context-free measure of word probability (see Equation 1) which allowed us to include languages with fewer and less detailed linguistic resources.[1]

$$p(word) = \frac{count(word)}{\sum_{word'} count(word')} \tag{1}$$

Prior work on the effects of incremental word processing has measured *segment information* as the -log$_2$ conditional probability of a segment given the current *cohort* (Marslen-Wilson & Welsh, 1978), that is, the set of word-forms in the lexicon that share identical segments until that point. For example, the information of [f] in *sphinx* is determined by dividing the token-count of words beginning with [sf] by the token-count of words beginning with [s], and then taking the base-2 logarithm of the resulting quantity. This token-based measure (Equation 2) has been previously shown to predict variation in information in the speech signal: lower

[1] Though the use of larger contextual windows has been shown to provide a potentially better fit between word probability and word length (Piantadosi et al., 2011), context-free probability is strongly correlated with probability measured over larger context window sizes (see Cohen Priva & Jaeger, 2018, for more).

token-based segment information correlates with shorter segment duration and less-distinct articulation (Tang & Bennett, 2018; van Son & Pols, 2003; van Son & van Santen, 2005), and lower average token-based segment information correlates with a greater probability of segment deletion in casual speech (Cohen Priva, 2015, 2017). Below, we will show that a parallel type-based measure provides similar results.

$$h(seg_n) = -\log_2 \frac{count(seg_1 ... seg_n)}{count(seg_1 ... seg_{n-1})} \tag{2}$$

We emphasize that this general approach for estimating the segmental information available to a listener is coarse-grained relative to how the speech stream is actually processed. For one, this measure treats segments as equivalently distinctive, abstract symbols, rather than as phonetic signals, which are differentially perceptually distinctive from one another and differentially robust to noise (Mielke, 2012; Smits, Warner, McQueen, & Cutler, 2003). We anticipate that future work will benefit from using measures that take perceptual distance into account (e.g., Gahl & Strand, 2016; Strand, 2014). Our current method also implies that there is no uncertainty in perception, assuming listeners immediately discard alternatives not compatible with each successive segment. Instead, there is evidence that lexical access is moderately tolerant of segmental mis-ordering (Toscano, Anderson, & McMurray, 2013) and that listeners can backtrack to some degree when new segmental information is incompatible with previous information (Gwilliams, Linzen, Poeppel, & Marantz, 2018; McMurray, Tanenhaus, & Aslin, 2009). Nonetheless, the method we use here should capture some portion of the information flow during lexical processing and has the advantage of being broadly applicable. We anticipate that more fine-grained, perceptually sophisticated measures will provide yet clearer outcomes.

RESULTS

Mean Segmental Information

Less-probable word-forms tend to have more overall segmental information just by virtue of having more segments (Piantadosi et al., 2011; Zipf, 1935). If all segments contributed equivalent information to word identification, addition or subtraction of segments would be the only way to change the information carried by a word-form. If this were the case, the *mean* information contributed by each segment across a word-form would not correlate with word probability in words of the same length. Conversely, if we find that less-probable words have higher mean segment information when controlling for length, it suggests these words, in addition to having more segments, have more disambiguating information packed into those segments.

We begin with the token-based measure of segment information described above because it is sensitive to both categorical cohort structure and word frequencies in those cohorts (see below for parallel tests using a type-based measure). However, the token-based measure carries a built-in correlation between word probability and segment information, because the frequency of a word contributes to the calculation of information of its segments. To eliminate this source of correlation, for the results presented in this section we used a modified form of the equation in which we subtract a word's frequency from the calculation of the information of its own segments (Equation 3).

$$h^*(seg_n) = -\log_2 \frac{count(seg_1 ... seg_n) - count(word)}{count(seg_1 ... seg_{n-1}) - count(word)} \tag{3}$$

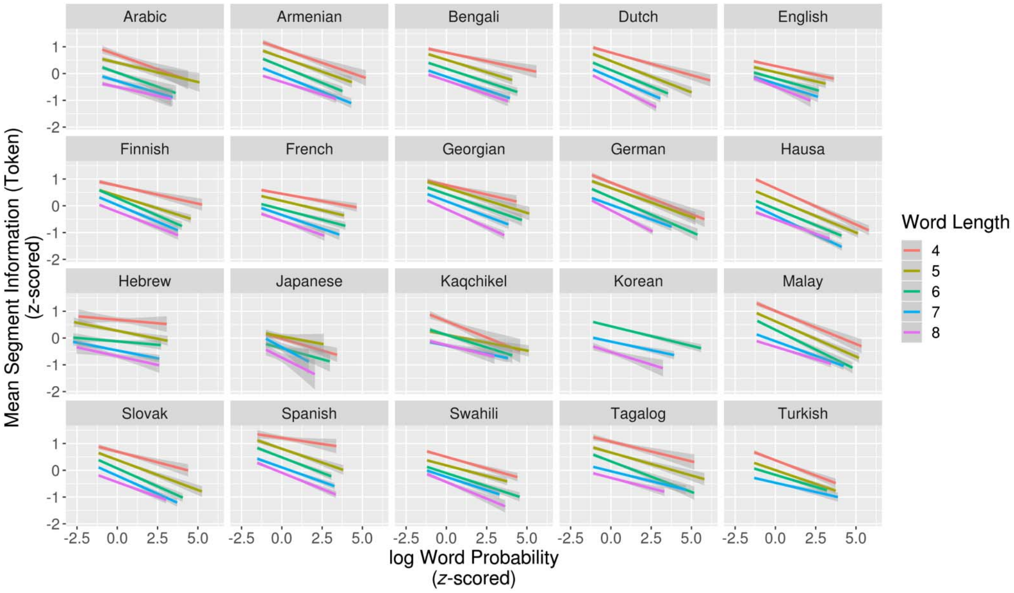

Figure 1. Relationship between log word probability and mean token-based segment information for words of length 4–8. Grayed area represents 95% confidence intervals. Less-probable words contain higher information segments.

We calculated the mean segment information for each word-form including only the segments before the *uniqueness-point* (cf. Marslen-Wilson & Welsh, 1978), that is, the point at which it is the only remaining word in the cohort.[2] We excluded post-uniqueness segments because, by this method, they contribute zero information. As a result, if segment information is averaged over the whole word, words with longer post-uniqueness-point sequences systematically show lower mean segment information values. Note, however, the relationship between word probability and mean segment information in this dataset remains significant when post-uniqueness-point segments are included (not shown), indicating that less-probable words have more incrementally informative segments on average across the entire word-form (cf. Mahowald et al., 2018; Meylan & Griffiths, 2017, who show that less-probable words are composed of phonotactically less-probable sequence types).

Figure 1 shows best-fit regression lines for mean token-based segment information by word probability for word lengths 4–8 in each language. As predicted, less-probable words contain more informative segments. Across all languages, all but three (94/97) of the by-length regression models show a significant correlation between word probability and mean segment information. When word length categories are pooled and word length is included as a separate factor, the models show a significant effect of word probability for all languages (Table 2 in the Supplemental Materials; King & Wedel, 2020).[3]

Using a token-based measure for segment information, the *sum* of the information contributed by each segment in a word-form is equal to context-free word information, $-\log_2$

[2] We added end-of-word boundary symbols to all words, setting the uniqueness-point to be the word length plus one for words that do not have a word-internal uniqueness-point, e.g., *cat*, given the existence of *catalog*.

[3] There is a significant, negative correlation between word length and mean segment information in each language. Variance inflation factor (VIF) scores were below an acceptable threshold (<2) in all languages (see O'brien, 2007, for discussion of VIF).

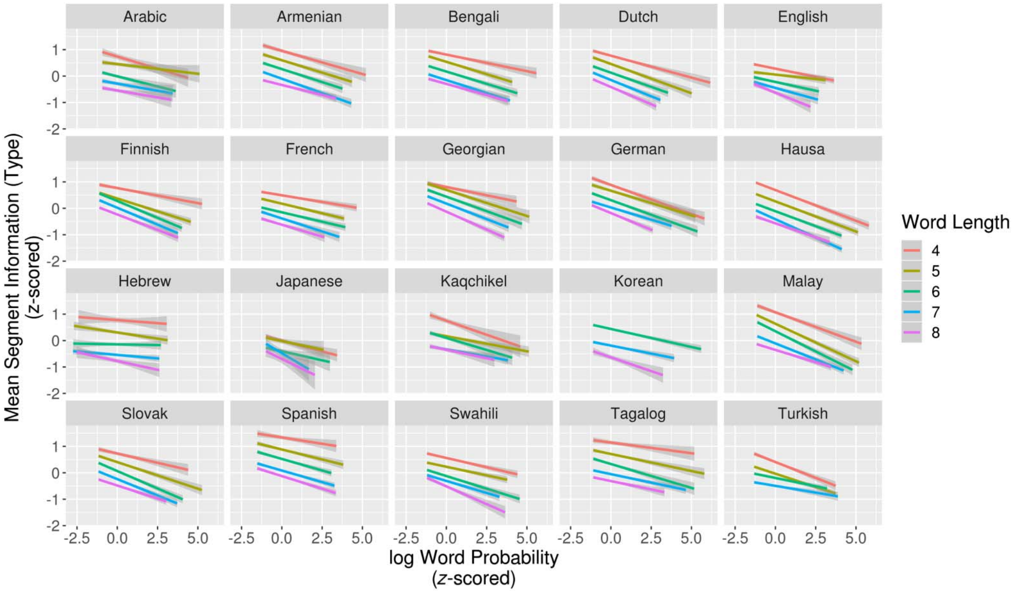

Figure 2. Relationship between mean type-based segment information and log word probability for words of length 4–8. Less-probable words more quickly reduce the cohorts of competing words.

$p(word)$.[4] As a consequence, the negative correlation that we find between *mean* segment information and word probability can only arise if the information in segments of less-probable words are concentrated in fewer segments. This has the effect of more rapidly reducing the probability mass of alternatives with each successive segment.

There are two conceptually distinct ways this more rapid reduction in probability mass could be accomplished: (a) successive segments could categorically reduce the number of competitors more quickly, or (b) successive segments could preferentially eliminate higher probability cohort members. In the following sections we show evidence that lexicons are optimized in both ways.

Segments in Less-Probable Words Reduce the Number of Competitors More Quickly

Cohort Size. To ask whether the segment sequences of less-probable words tend to more rapidly disambiguate them from a greater number of competing word-types early in processing, we used a type-based variant of the measure of segment information (Equation 2). For example, the type-based information of [f] in *sphinx* is equal to the number of word-types that begin with [sf] in a corpus divided by the number that begin with [s].

As above, we fit linear regression models to predict word probability given mean type-based segment information for words of length 4–8 separately in all languages (Figure 2). In all but two (95/97) cases, word probability showed significant negative correlation with mean type-based segment information. Again, when all lengths were pooled together, word probability showed a significant, negative correlation with mean type-based segment information in all languages (see Table 3 in the Supplemental Materials; King & Wedel, 2020). Because the

[4] However, here it is not precisely equivalent, because we subtract a word's frequency from the calculation of information for its own segments.

mean type-based segment information measure ignores word frequency, its significant correlation with word probability indicates that segments in less-probable words disambiguate from relatively more competitors, reducing cohort sizes more rapidly.

Position of Uniqueness-Point. If less-probable words contain segments that more quickly reduce cohort sizes, then the point at which their word-forms become unique should be earlier, relative to length. As an example, the words *thwart* and *story* both have five segments, but the uniqueness-point of the less-probable *thwart* comes in its second segment (i.e., no other word in our English corpus begins with [θw]), while the uniqueness-point of *story* falls in its last segment, where it disambiguates from *stork, storm, storage*, and so on.

Linear regression models predicting the relative position of a word's uniqueness-point (i.e., uniqueness-point divided by number of segments) by word probability for lengths 4–8 separately showed that less-probable words do in fact have significantly earlier uniqueness-points in all but four (93/97) by-length regressions (Figure 3). As above, when lengths are pooled, we found that less-probable words have significantly earlier relative uniqueness-points in all languages (Table 4 in the Supplemental Materials; King & Wedel, 2020). When included as a factor in models to predict the nonrelative uniqueness-point position, we found an independent, positive effect of word length, suggesting that longer words have later uniqueness-points, on average (Table 5 in the Supplemental Materials; King & Wedel, 2020; cf. Strauss & Magnuson, 2008).

Segments in Less-Probable Words Eliminate More-Probable Competitors

Comparison to Word Probability-Shuffled Baselines. Here we ask whether the significant relationship between word probability and token-based segment information is in part because less-probable words tend to be grouped in cohorts with more-probable words, allowing early segments in less-probable words to eliminate a greater probability mass of competitors, independently of *how many* competitors they eliminate. To do this, we compared the real-world

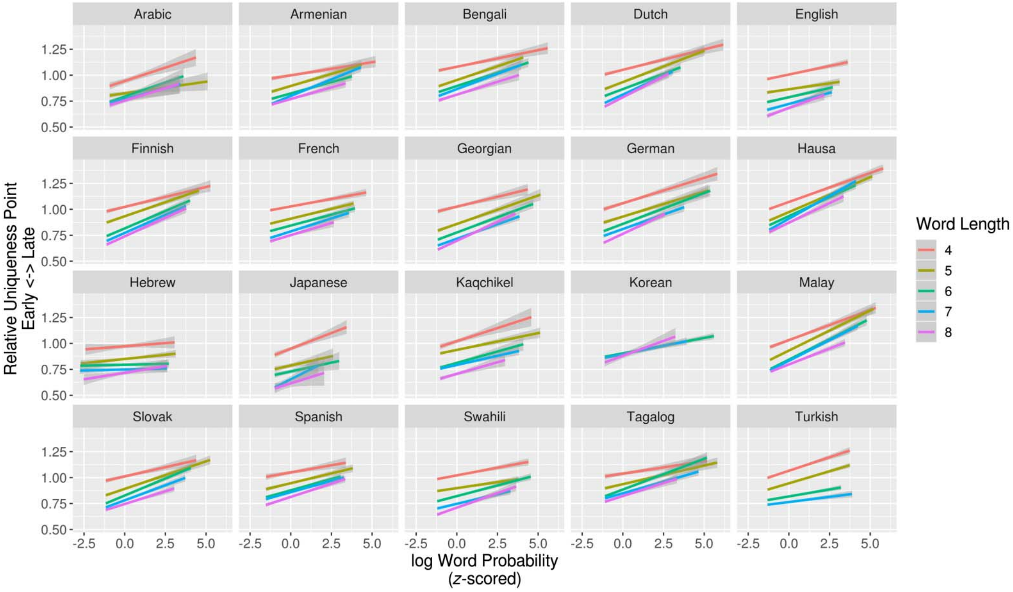

Figure 3. Relationship between log word probability and relative position of uniqueness-point for words of length 4–8. Less-probable words have relatively earlier uniqueness-points for all lengths.

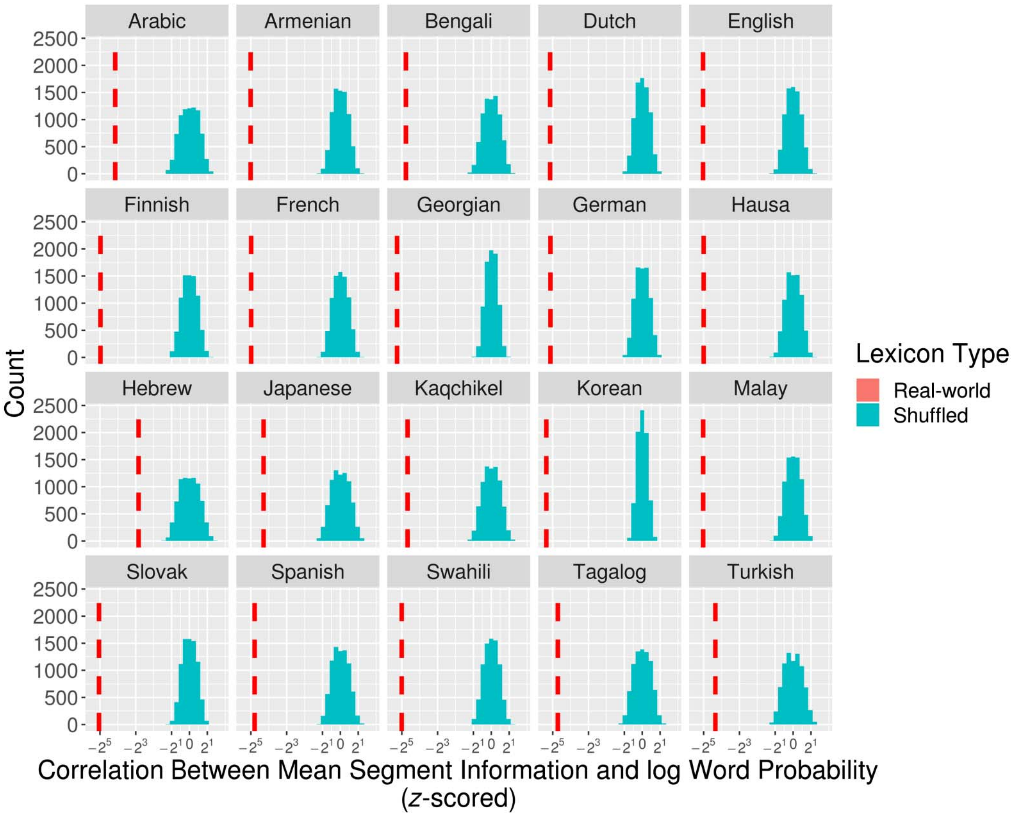

Figure 4. Distribution for Pearson's correlation between log word probability and mean token-based segmental information for 10,000 shuffled variants of the real-world lexicons. The x-axis shows the number of standard deviations from the mean correlation in frequency-shuffled variants (in $\log_2$ scale) and the red dashed lines indicate the correlation in the real-world lexicons. The real-world lexicons show a significantly stronger correlation relative to shuffled variants.

lexicons of each of our tested languages against probability-randomized but structurally identical variants of the lexicon, created by shuffling the context-free probability for words of the same length in each language and then recalculating segmental information.[5] Shuffling word probabilities within length classes creates variant lexicons in which the (potentially optimized) probability relationship between words in cohorts is severed, while maintaining both the original cohort structure as well as the original relationship between word probability and length. For example, in a shuffled variant of the English lexicon, *thwart* might take on a higher probability, which would slightly reduce the information of word-initial [θ].

For each language, we compared the Pearson's correlation between mean segmental information and log word probability in 10,000 probability-shuffled lexicons against the correlation found in the real-word lexicon. In all cases, the correlation in the real-world lexicon was significantly stronger (3+ standard deviations, $p < .001$) than in the shuffled lexicons (Figure 4), indicating that the strength of the real-world correlation is greater than would be expected by chance. This suggests that the real-world lexicons have evolved such that less-probable words

[5] Here we use the unmodified form of segment information (Equation 2) because the built-in correlation between word probability and token-based segment information is the same for the real-world and shuffled lexicons.

have segment sequences that preferentially eliminate higher probability competitors across the word-form.

Segmental Information Distribution Within Words

The patterns we have presented so far show that less-probable words contain relatively higher information segments. If the lexicon is structured for incremental processing, this bias should be greatest at word beginnings, where high-information segments can better offset lower word probability. If less-probable words evolve higher segment information early in the word, mean segment information should be relatively higher when calculated in the forward than reverse order for less-probable words. Likewise, less-probable words should have earlier relative uniqueness-points in the forward-order than in the reverse. As an example, the less-probable word *thwart* has a high mean segment information and an early uniqueness-point due to the rarity of its initial segments, while in the reverse lexicon, these values are less extreme because of the large set of words that end in [aɹt].

To test this, we compared the mean segment information and relative uniqueness-point for words in each actual lexicon against those in the reverse-order lexicon. Each word in the reverse lexicon is the same as its forward counterpart in length and segment composition, but the set of cohorts defined by each successive segment is different. For example, the initial cohort for *thwart* comprises the relatively few words beginning with [θ], while in the reverse lexicon it comprises the many words ending in [t]. In these studies, we computed mean segment information only up to the uniqueness-point, with the result that the same word can have different mean segment information when calculated in the forward or backward lexicon. Note that because our measure of segment information treats segments as abstract symbols defining a network, using a reverse lexicon is licit even though it creates segment sequences that may not be pronounceable. We fit linear mixed-effects models over a pooled dataset of all languages to predict the forward and reverse values by (a) word probability, (b) a binary factor for lexicon order (reverse-order vs. forward-order), and (c) their interaction, with random intercepts and slopes for word and language nested within family. In all models, the interaction between order and word probability was significant with the expected sign, supporting the prediction that less-probable words have higher early segmental information in the actual, as opposed to the reverse lexicons (Tables 6, 7, 8 in the Supplemental Materials; King & Wedel, 2020).

To confirm these differences at the word-level (as opposed to just across the lexicons as a whole), we constructed linear mixed-effects models to predict the *difference* in forward- and reverse-order measures within each word, by subtracting the reverse from the forward-lexicon value. These models provided the same outcomes, supporting the hypothesis that less-probable words tend to evolve higher segment information early (Tables 9, 10, 11 in the Supplemental Materials; King & Wedel, 2020). Using this approach, we additionally carried out regressions within each individual language (Figures 5, 6, 7 in the Supplemental Materials; King & Wedel, 2020). We found that the majority, but not all languages showed the significant effects apparent within the pooled dataset; see Discussion.

DISCUSSION

We have presented evidence that segments in less-probable words convey more disambiguating information in incremental processing. Further, in many languages this positive correlation is concentrated at word-beginnings, where the potential difference in segmental information is greatest. These findings contribute compelling evidence that lexicons are optimized efficient communication overall within the constraints of the language processing system (Dautriche et al., 2017; Ferrer i Cancho & Solé, 2003; Gibson et al., 2019; Mahowald et al., 2018; Meylan & Griffiths, 2017; Piantadosi et al., 2011, 2012).

Evolution of Word-Forms in the Lexicon

How might these lexical patterns arise? Evidence from corpus studies suggest that less informative segments are more likely to be shortened or deleted in speech (Cohen Priva, 2015, 2017), and are more likely to be replaced by similar sounds over time (Wedel, Kaplan, & Jackson, 2013). Parallel evidence shows that more-probable words are more likely to shorten (Bybee & Hopper, 2001; Kanwal et al., 2017) and become more similar to other words (Frauenfelder, Baayen, & Hellwig, 1993; Mahowald et al., 2018). Conversely, segments that provide more information tend to be pronounced with greater clarity (Aylett & Turk, 2004, 2006; Buz, Jaeger, & Tanenhaus, 2014; Nelson & Wedel, 2017; Sano, 2017; Seyfarth, Buz, & Jaeger, 2016; van Son & Pols, 2003; van Son & van Santen, 2005; Wedel, Nelson, & Sharp, 2018), and are more likely to persist in a language over time (Wedel et al., 2013).

Because more-probable words require less segmental information to be accurately understood (reviewed in Hall et al., 2016), these considerations predict that segments in more-probable words should be more rapidly lost, or replaced over time with more-frequent segments (see discussion in Bybee & Hopper, 2001; Piantadosi et al., 2011). Because of this asymmetry, over long time periods more-probable words should drift into denser, phonotactically probable cohorts, while less-probable words should preferentially retain less-common segment sequences, leaving them in sparser regions of the lexicon's network structure.

All languages in the dataset show a significant correlation between lower word probability and greater incremental segment information. Why do some languages fail to show a skew toward higher segment information at beginnings of less-probable words? Many of those particular languages have constraints that enforce denser lexical networks: for example, word-forms in Hebrew and Arabic are based in tri-consonantal roots that constrain lexicon size (Ussishkin, 2005); words in Kaqchikel are based on single syllable roots (Bennett, 2016); words in Tagalog and Malay have simple phonotactics and tend to be bi-syllabic (Blust, 2007); Swahili, likewise, has simple phonotactics and a preference for bisyllabic word stems (Mohamed, 2001). These language-specific constraints on word-forms result in denser lexical networks, which should inhibit loss of information late in word-forms. Initial work indicates a significant link between denser lexical networks and maintenance of late segment information in less-probable words.

Broader Implications

Zipf's law of abbreviation is strikingly consistent across a wide range of tested languages (Bentz & Ferrer i Cancho, 2016). Likewise, we find a similar pattern of correlation between word probability and segment information across a diverse set of languages. The fact that we see similar correlations in each of these languages suggests that like Zipf's law of abbreviation, this may also be a robust, "statistically-universal" property of human languages (Dryer, 1998). Together with the evidence that word-forms are shaped for efficient production by speakers (Dautriche, 2015; Mahowald et al., 2018; Meylan & Griffiths, 2017), the findings here support a broader trend of linguistic evolution toward systems that benefit both speakers and listeners, in which modulation of segment number and segment composition in words are complementary parts of this larger process.

ACKNOWLEDGMENTS

The authors would like to thank Roger Levy and the two anonymous reviewers for their insightful comments and suggestions during the preparation of this article. The authors would also like to thank the attendees of CUNY 2016 and EvoLang XII for useful discussion and feedback.

FUNDING INFORMATION

AK, National Science Foundation (NSF), Award ID: Graduate Research Fellowship (2016233374).

AUTHOR CONTRIBUTIONS

AK: Conceptualization: Lead; Data curation: Lead; Formal analysis: Equal; Methodology: Equal; Resources: Lead; Visualization: Lead; Writing—Original Draft: Equal; Writing—Review & Editing: Equal. AW: Formal analysis: Equal; Methodology: Equal; Supervision: Equal; Writing—Original Draft: Equal; Writing—Review & Editing: Equal.

REFERENCES

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419–439.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America*, *119*, 3048–3058.

Bennett, R. (2016). Mayan phonology. *Language and Linguistics Compass*, *10*, 469–514.

Bentz, C., & Ferrer i Cancho, R. (2016). Zipf's law of abbreviation as a language universal. *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*, Leiden, Netherlands.

Blust, R. (2007). Disyllabic attractors and anti-antigemination in Austronesian sound change. *Phonology*, *24*(1), 1–36.

Buz, E., Jaeger, T. F., & Tanenhaus, M. K. (2014). Contextual confusability leads to targeted hyperarticulation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36. Retrieved from https://escholarship.org/uc/item/7ph8539f

Bybee, J. L., & Hopper, P. J. (2001). *Frequency and the emergence of linguistic structure* (Vol. 45). Amsterdam, Netherlands: John Benjamins.

Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, *6*, 243–278.

Cohen Priva, U. (2017). Informativity and the actuation of lenition. *Language*, *93*, 569–597.

Cohen Priva, U., & Jaeger, T. F. (2018). The interdependence of frequency, predictability, and informativity in the segmental domain. *Linguistics Vanguard*, *4*(s2).

Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language*, *32*, 193–210.

Dahan, D., & Magnuson, J. S. (2006). Spoken word recognition. In M. Traxler & M. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 249–283). Amsterdam, Netherlands: Elsevier.

Dautriche, I. (2015). *Weaving an ambiguous lexicon* (Unpublished doctoral dissertation). Université Sorbonne Paris Cité.

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, *41*, 2149–2169.

Dryer, M. S. (1998). Why statistical universals are better than absolute universals. In *Papers from the 33rd Regional Meeting of the Chicago Linguistic Society* (pp. 1–23). Chicago, IL.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*, 17897–17902.

Ferrer i Cancho, R. (2017). The placement of the head that maximizes predictability. An information theoretic approach. *arXiv preprint arXiv:1705.09932*.

Ferrer i Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, *100*, 788–791.

Frauenfelder, U. H., Baayen, R. H., & Hellwig, F. M. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language*, *32*, 781–804.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*, 10336–10341.

Gahl, S., & Strand, J. F. (2016). Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language*, *89*, 162–178.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 199–206). Stroudsburg, PA.

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., et al. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*. Retrieved from http://www.sciencedirect.com/science/article/pii/S1364661319300580 https://doi.org/10.1016/j.tics.2019.02.003

Gildea, D., & Jaeger, T. F. (2015). Human languages order information efficiently. *arXiv preprint arXiv:1510.02823*.

Grosjean, F. (1996). Gating. *Language and Cognitive Processes*, *11*, 597–604.

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, *38*, 7585–7599.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, *32*(2), 101–123.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*, 643–672.

Hall, K. C., Hume, E., Jaeger, F., & Wedel, A. (2016). The message shapes phonology. Manuscript, University of British Columbia, University of Canterbury, University of Rochester & Arizona University. Retrieved from http://psyarxiv.com/sbyqk

Hawkins, J. A. (2010). Processing efficiency and complexity in typological patterns. In J. J. Song (Eds.), *The Oxford handbook of linguistic typology* (pp. 206–226). New York: Oxford University Press.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jaeger, T. F., & Tily, H. (2011). On language "utility": Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*, 323–335.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52.

King, A., & Wedel, A. (2020). Supplemental material for "Greater early disambiguating information for less-probable words: The lexicon is shaped by incremental processing." *Open Mind: Discoveries in Cognitive Science*, *4*. doi:10.1162/opmi_a_00030

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *31*(1), 133–156.

Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. T. (2018). Word forms are structured for efficient use. *Cognitive Science*, *42*, 3116–3134.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1–2), 71–102.

Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63.

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 576–585.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category vot affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language*, *60*(1), 65–91.

Meylan, S. C., & Griffiths, T. L. (2017). Word forms—not just their lengths—are optimized for efficient communication. *arXiv preprint arXiv:1703.01694*.

Mielke, J. (2012). A phonetically based metric of sound similarity. *Lingua*, *122*(2), 145–163.

Mohamed, M. A. (2001). *Modern Swahili grammar*. Kampala, Uganda: East African Publishers.

Nelson, N. R., & Wedel, A. (2017). The phonetic specificity of competition: Contrastive hyperarticulation of voice onset time in conversational English. *Journal of Phonetics*, *64*, 51–70.

Nooteboom, S. G. (1981). Lexical retrieval from fragments of spoken words: Beginnings vs. endings. *Journal of Phonetics*, *9*, 407–424.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*, 673–690.

Piantadosi, S. T., Tily, H. J., & Gibson, E. (2009). The communicative lexicon hypothesis. In *The 31st Annual Meeting of the Cognitive Science Society (cogsci09)* (pp. 2582–2587). Austin, TX.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*, 3526–3529.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*, 280–291.

Salasoo, A., & Pisoni, D. B. (1985). Interaction of knowledge sources in spoken word identification. *Journal of Memory and Language*, *24*(2), 210–231.

Sano, S.-I. (2017, October). *Minimal pairs and hyperarticulation of singleton and geminate consonants as enhancement of lexical/pragmatic contrasts*. Paper presented at the 48th conference of North East Linguistic Society (NELS 48), University of Iceland, Reykjavík.

Seyfarth, S., Buz, E., & Jaeger, T. F. (2016). Dynamic hyperarticulation of coda voicing contrasts. *Journal of the Acoustical Society of America*, *139*(2), EL31–EL37.

Smits, R., Warner, N., McQueen, J. M., & Cutler, A. (2003). Unfolding of phonetic information over time: A database of Dutch diphone perception. *Journal of the Acoustical Society of America*, *113*, 563–574.

Strand, J. F. (2014). Phi-square lexical competition database (phi-lex): An online tool for quantifying auditory and visual lexical competition. *Behavior Research Methods*, *46*(1), 148–158.

Strauss, T., & Magnuson, J. S. (2008). Beyond monosyllables: Word length and spoken word recognition. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1306–1311). Washington, DC.

Tang, K., & Bennett, R. (2018). Contextual predictability influences word and morpheme duration in a morphologically complex language (Kaqchikel Mayan). *Journal of the Acoustical Society of America*, *144*, 997–1017.

Toscano, J. C., Anderson, N. D., & McMurray, B. (2013). Reconsidering the role of temporal order in spoken word recognition. *Psychonomic Bulletin & Review*, *20*, 981–987.

Ussishkin, A. (2005). A fixed prosodic theory of nonconcatenative templaticmorphology. *Natural Language & Linguistic Theory*, *23*(1), 169–218.

van Son, R., & Pols, L. C. (2003). How efficient is speech. *Proceedings of the Institute of Phonetic Sciences*, *25*, 171–184.

van Son, R., & van Santen, J. P. (2005). Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication*, *47*(1–2), 100–123.

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*, 387–401.

Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, *128*, 179–186.

Wedel, A., Nelson, N., & Sharp, R. (2018). The phonetic specificity of contrastive hyperarticulation in natural speech. *Journal of Memory and Language*, *100*, 61–88.

Zipf, G. K. (1935). *The psycho-biology of language*. Boston: Houghton Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Reading, MA: Addison-Wesley.