Language-users choose short words in predictive contexts in an artificial language task

Jasmeen Kanwal (jasmeen.kanwal@ed.ac.uk), Kenny Smith (kenny.smith@ed.ac.uk), Jennifer Culbertson (jennifer.culbertson@ed.ac.uk), and Simon Kirby (simon@ling.ed.ac.uk)
Centre for Language Evolution, University of Edinburgh
3 Charles St., Edinburgh, Scotland EH8 9AD

Abstract

Zipf (1935) observed that word length is inversely proportional to word frequency in the lexicon. He hypothesised that this cross-linguistically universal feature was due to the *Principle of Least Effort*: language-users align form-meaning mappings in such a way that the lexicon is optimally coded for efficient information transfer. However, word frequency is not the only reliable predictor of word length: Piantadosi, Tily, and Gibson (2011) show that a word's predictability in context is in fact more strongly correlated with word length than word frequency. Here, we present an artificial language learning study aimed at investigating the mechanisms that could give rise to such a distribution at the level of the lexicon. We find that participants are more likely to use an ambiguous short form in predictive contexts, and distinct long forms in surprising contexts, only when they are subject to the competing pressures to communicate accurately and efficiently. These results support the hypothesis that language-users are driven by a least-effort principle to restructure their input in order to align word length with information content, and this mechanism could therefore explain the global pattern observed at the level of the lexicon.

Keywords: Information theory; Efficient communication; Artificial language learning; Uniform Information Density

Introduction

Zipf (1935) observed that word length tends to be inversely proportional to word frequency in the lexicon. He hypothesised that this widespread cross-linguistic pattern was due to the *Principle of Least Effort*: language-users align form-meaning mappings in such a way that effort is minimised while expressivity is still maintained. However, word frequency is not the only reliable predictor of word length. Using corpora from 11 different languages, Piantadosi et al. (2011) show that a word's predictability in context (where they define context as the two words preceding the target word) is even more strongly correlated with word length than frequency is: words that are, on average, more predictable in context tend to be shorter.

Measuring how predictable or unpredictable a word is in a particular context gives us a way of defining the *information content* of a word. For example, consider the two sentences:

(1) The early bird catches the worm.
(2) Our early bird special today is a baked-apple worm.

In sentence (1), a well-known proverb, the word *worm* is entirely predicted by the preceding words. The word itself thus gives us practically no new information, and so it has *low information content*. In sentence (2), the same word is highly unlikely given the preceding words, and thus we find it surprising. This element of surprise is associated with *high information content*.

Using these concepts, we can apply Zipf's Principle of Least Effort to hypothesise that a speaker's drive to reduce effort will be directed towards words that are already highly predictable given the context, i.e. have low information content. Words that are more surprising in a particular context will be less likely to be reduced, or more likely to be lengthened. The resulting state in which low-information words are shorter than high-information words, and thus the length of a word is roughly proportional to the amount of information associated with the word, is consistent with the Uniform Information Density (UID) principle (Jaeger, 2010) or the Smooth Signal Redundancy (SSR) hypothesis (Aylett & Turk, 2004), which state that information is distributed roughly evenly across words in an utterance.

There are many ways to operationalise the information content of a word. One way is to use the *N-gram probability* of a word, i.e. its probability conditioned on a window of N preceding or following words. This is the method used by Piantadosi et al. (2011). Zipf's word frequency measure is in fact just a limiting case of this N-gram probability, where N=0. Other measures include *syntactic probability*, a word's probability of appearing in a particular syntactic structure (Jaeger, 2010, e.g.), and *givenness*, a word's predictability given the semantic context (Aylett & Turk, 2004).

Both corpus studies and controlled behavioural experiments have linked low information content, operationalised in these different ways, to various types of linguistic reduction. Lieberman (1963); Aylett and Turk (2004); Gahl and Garnsey (2004); Tily et al. (2009); Kuperman and Bresnan (2012), and Seyfarth (2014) show that words with low information content are more likely to undergo various types of phonetic reduction. Bell, Brenier, Gregory, Girand, and Jurafsky (2009) show that each of the four different measures of information content mentioned above may in fact contribute separately to the phonetic duration of a word. Fedzechkina, Jaeger, and Newport (2012) show that case markers are more likely to be omitted on nouns in more probable syntactic roles. Jaeger (2010) shows that *that*-complementisers are more often dropped when the following word is less surprising in context.

If predictability in context can lead to phonetic reduction, as well as deletion of morphemes and entire words, then these effects might make their way to the overall distribution of form-meaning mappings in the lexicon. However, there is relatively little work directed at understanding how predictability affects this widely observed pattern at the level of the lexicon.

One way of investigating the issue is by tracking language-users' online choices when producing words that are part of a 'clipped pair', i.e. when both a long form and an abbreviated or 'clipped' form exist that have the same or very similar meanings (Mahowald, Fedorenko, Piantadosi, & Gibson, 2013). E.g. in English, *info/information* is a clipped pair. Mahowald et al. presented participants with sentences containing a blank and asked them to complete the sentence with either the long or the clipped form corresponding to the relevant meaning. They found that participants were more likely to choose the short form in predictive contexts, which is consistent with the hypothesis that the lexicon-level patterns observed by Piantadosi et al. (2011) may be due in part to a *least-effort* mechanism, in which speakers balance communicative efficacy with efficiency.

However, because this study uses English sentence frames and target words, we cannot rule out potentially confounding contributions from register, prosody, and participants' learned preferences to their word choice in particular instances. Moreover, we cannot assess whether the effect is really driven by the competing pressures for communicative accuracy and efficiency without manipulating the presence or absence of these different communicative pressures. For instance, in Mahowald et al.'s task, it seems participants clicked on a word rather than typing it in, and thus there was no difference in effort between choosing the long or short form. In addition, participants were told to choose a word based on "which sounded more natural", rather than being directly engaged in a task requiring successful communication.

Here, we present a new artificial language learning study investigating the question of whether language-users align word length with information content when communicating. Our results are consistent with previous findings that language-users tend to use shorter forms in more predictive contexts. Furthermore, the behaviour we observe across different experimental conditions supports the hypothesis that this effect is driven at least in part by a least-effort principle, in which language-users balance the competing pressures for communicative accuracy and efficiency to reshape the lexicon into one where word length is roughly proportional to average information content.

Method

Artificial language learning studies have previously been used to shed light on the cognitive mechanisms and environmental pressures that shape large-scale linguistic structure. In this paradigm, participants learn an artificial language, and then we observe how they reshape their input as they use the language, in this case to communicate with a partner (e.g., Winters, Kirby, & Smith, 2015; Kirby, Tamariz, Cornish, & Smith, 2015; Fehér, Wonnacott, & Smith, 2016).

Participants

120 participants (53 females, 66 males; one did not report their gender) were recruited and remunerated via Amazon Mechanical Turk. 108 of these reported themselves as native English speakers, of which 96 were monolingual. A range of other languages were represented across the remaining participants. Ages ranged from 18 to 70 (mean=32.9, SD=9.5).

The Training Language

The study was run online. Participants were trained on two names for each of two plant-like alien objects, by repeatedly being shown pictures of the objects labeled with a simple sentence. The sentence consisted of a framing word followed by the object's name. There were two possible frames, *bix* and *gat*. Overall there were 64 training trials, with each object appearing 32 times and each frame appearing 32 times. Crucially, one object appeared seven times more frequently with the frame *bix* than *gat* (28 and 4 times, respectively), while the other object appeared seven times more frequently with the frame *gat* than *bix* (again, 28 and 4 times, respectively). This meant that each object appeared in both a predictive context and a surprising context; which frame signified which of these contexts was flipped between the two objects.

Furthermore, the object name appeared half the time in its full form, a 7-letter word, and half the time in shortened form, a 3-letter word derived by clipping the last two syllables off the long name. These short and long forms were evenly distributed across both predictive and surprising contexts, ensuring that the input language contained no inbuilt bias towards using one form in any particular context.[1] A schematic diagram of the object frequencies and labels is provided in Fig. 1A.

In natural languages, shorter words are subject to greater confusability for a number of reasons: shorter forms have less space for signal redundancy and thus are more likely to be completely lost in noisy signal transmission; and because languages have a finite phoneme inventory, there are more unique possible long strings than short strings, and thus word shortening often results in ambiguity. Indeed, shorter words are more likely to be polysemous and homophonous (Piantadosi, Tily, & Gibson, 2012). To model this fact in our miniature lexicon, we designed the names such that the short name for both objects was identical (*zop*), while the long names were unique (*zopekil* and *zopudon*).

Procedure

Participants were assigned to one of four conditions, where we manipulated the presence of pressures to communicate accurately and quickly in a between-subjects 2x2 design (Kanwal, Smith, Culbertson, & Kirby, 2017). Each experiment consisted of two phases: training and testing. The training phase was uniform across conditions, while the testing phase varied by condition.

[1] Which object (the blue fruit or the red stalk) appeared more frequently with which frame, as well as which object was paired with which long name, were both counterbalanced between participants, giving a total of 4 possible object-frame-name pairings which a participant might be trained on. This ensured that potential factors such as sound symbolism, or higher saliency or learnability of any specific object-word pairing, could not systematically bias our results.

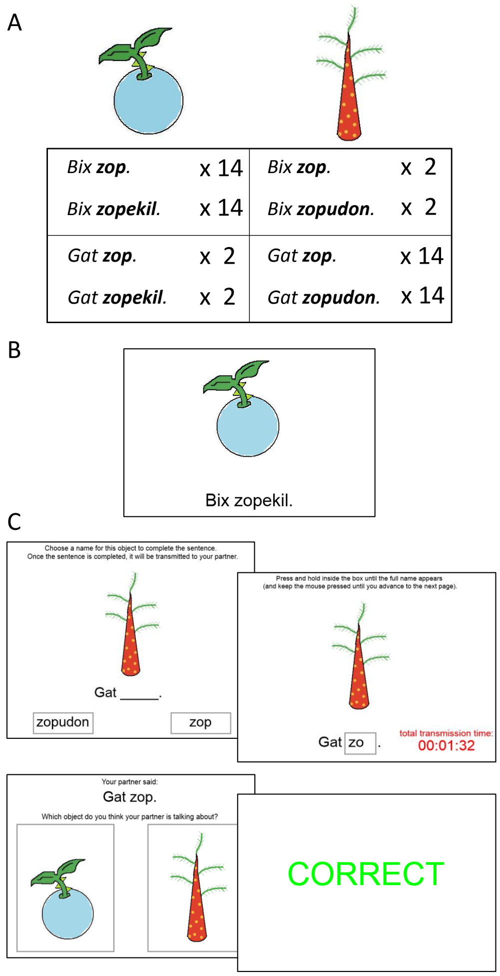

Figure 1: (A) The input frequencies of the objects and framing sentences presented during training trials in all four experimental conditions. (B) A sample training trial. (C) A sample director trial in the Combined condition (top) and a matcher trial followed by feedback (bottom).

Training phase On each training trial, an object was presented on screen alone for 700ms. The appropriate sentence then appeared beneath the object for a further 3000ms, yielding a total trial duration of 3700ms. A blank screen showed for 500ms between trials. The 64 training trials were presented in a different randomised order for each participant.

Testing phase After the training phase, testing procedures varied depending on the experimental condition. In the Combined condition, participants were under a pressure to communicate accurately *and* efficiently, as according to the Principle of Least Effort, it is balancing these competing pressures that leads language-users to distribute word length inversely to word predictability. The remaining three conditions removed one or both of these accuracy and time pressures.

Condition 1: Combined In the testing phase of this condition, participants were paired with a partner to play a communication game, using the method developed for running two-player online experiments in Kanwal et al. (2017). On each trial, the 'director' was shown an object on the screen with a framing word followed by a blank. The director was instructed to choose a name for the object to complete the sentence, and once the name was entered, the sentence would be transmitted to the 'matcher'. The director could choose one of two options to complete the sentence: the unique long name for the object or the (ambiguous) short name. Once the chosen name was selected by clicking on the appropriately labeled button, it had to be entered into the blank space by pressing and holding the mouse as each letter appeared one after the other at 1200 ms intervals. Only after all the letters in the name had appeared in the box was the completed sentence transmitted to the matcher. This belaboured method of production, in which the long name was significantly slower to produce than the short name, was introduced to model the difference in effort and speed associated with producing long versus short utterances.

Once the director completed their description, it was transmitted to the matcher, who was asked to choose which of the two objects they thought the director was referring to. Both players were then given feedback as to whether the matcher's choice was correct.

The players alternated roles after every trial, with the matcher becoming the director and the director becoming the matcher, until both completed 32 director trials and 32 matcher trials. The proportion of times each object appeared with each frame in each player's director trials matched those of the training proportions: one object appeared seven times more frequently with the frame *gat* than *bix*, and the other appeared seven times more frequently with *bix* than *gat*. The order of each participant's 32 director trials was randomly shuffled.

To model the pressures in spoken communication to be both efficient and accurate, pairs were told at the beginning that they would be rewarded a bonus payment of $1 if they were the pair to complete the game in the quickest time with the highest number of correct match trials. Time was only counted when the director was entering a name into the blank, and the total time count was displayed next to the blank during this process, to emphasise the time pressure. Screenshots of sample director and matcher trials are shown in Fig. 1C.

In this condition, with pressures to be speedy yet accurate, we expected participants to converge on an optimal strategy in which the short name is used for an object when it appears in its predictive context, and the long name otherwise. In predictive contexts, the framing word already provides a lot of information to the matcher about which object is likely under discussion, and thus participants can minimise effort by using the short form. Conversely, in surprising contexts, the full object name is required to ensure disambiguation.

In order to establish a causal link between these purported mechanisms and the behaviour we observe, we included three further experimental conditions, described below, for a full 2x2 manipulation of the pressures for accuracy and efficiency.

Condition 2: Accuracy In this condition, participants were paired to play a communication game as described above, but in the director trials, there was no intermediate step between the director choosing a name to complete the sentence and the matcher receiving the sentence; the names were entered instantaneously, thus removing any difference in effort between producing long or short names. Pairs were told that the goal of the game was simply to have their partner make as many correct guesses as possible. No bonus prize was offered in this condition, as we expected many pairs to hit ceiling as they did in Kanwal et al. (2017)—however, fewer than expected actually did so here.

In this condition, we predicted that participants would be more likely to use the long names for both objects across all contexts, as the long names are less confusable, and without a pressure to be efficient, there is little reason to shorten.

Condition 3: Time In this condition, communication was taken out of the game entirely; participants played a one-player game consisting of 64 director trials. In each trial, participants completed the sentence with either the long or short name for the object shown, but there was no subsequent communicative task. The name was simply entered as in the Combined condition, by pressing and holding the mouse in the blank space, with each letter appearing at 1200 ms intervals, while a timer displayed the total time count. The next trial began once all the letters had appeared in the box. Participants were told at the beginning of the game that they would be rewarded a bonus payment of $1 if they were the player with the shortest total time count.

Here, we expected participants to use the short name for both objects across all contexts: with no communicative purpose attached to the transmissions, and an incentive to be as quick as possible, using the short name in every trial is the best strategy.

Condition 4: Neither The fourth and last condition contained neither a pressure for efficiency nor a pressure for accuracy. As in the Time condition, participants played a one-player game with no explicit communicative element. Additionally, there was no time difference associated with transmission; once a label was chosen to complete a sentence, it was instantaneously recorded and the player advanced to the next trial. We included this condition to provide a baseline for participants' behaviour from which to assess the effects of the accuracy and time pressures in the other three conditions.

In this condition we expected that participants would either probability-match—i.e. use the long and short forms for both objects with equal frequency, as in the training trials (Hudson Kam & Newport, 2005)—or their behaviour would reveal prior biases language users bring to the task, such as a preference against using ambiguous forms, as observed in Kanwal et al. (2017).

Results

Fig. 2 shows the proportion of trials in which the short name was produced by each participant or pair of participants in

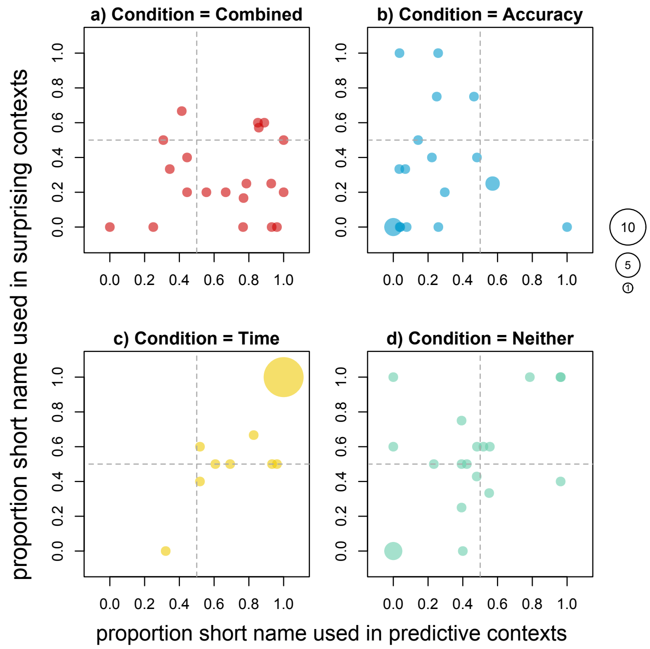

Figure 2: The proportion of trials in which the short name was used in predictive contexts versus the proportion of trials in which it was used in surprising contexts. For the Combined and Accuracy condition, each data point combines a pair of communicating players, representing the sum of their director trial productions. For the Time and Neither condition, each data point corresponds to an individual player's productions. The size of the circles is perceptually scaled (Tanimura et al., 2006) to reflect the number of data points coinciding at each value. Data from only the second half of testing trials is shown here, as participants were more likely to have converged on a stable mapping by this time. These results demonstrate that behaviour consistent with the principles of UID or SSR—using short forms in predictive contexts and long forms in surprising contexts, generating systems that fall in the bottom right corner of each graph—only reliably arises in the Combined condition.

predictive versus surprising contexts. Our predictions were borne out by the results in all four conditions. In the critical Combined condition, in which participants were subject to the combined pressures for accuracy and efficiency, pairs of communicating participants produced systems in which the short name was used in predictive contexts and the long name in surprising contexts. Crucially, only when both pressures were present did participants reliably produce systems where word length was conditioned on context in this way. In the Accuracy condition, participants tended to use the long name for both objects regardless of context, and in the Time condition, they used the short name for both objects regardless of context. In the Neither condition, some participants stuck with the long name or the short name throughout the trials regardless of context, as in the Accuracy or Time conditions; however, most participants probability-matched.

A logistic regression model was fit to the full dataset in R using the lme4 package, with short name use (as contrasted with long name use) as the binary dependent variable; context (predictive or not), experimental condition, and their interaction as fixed effects; and by-participant random slopes

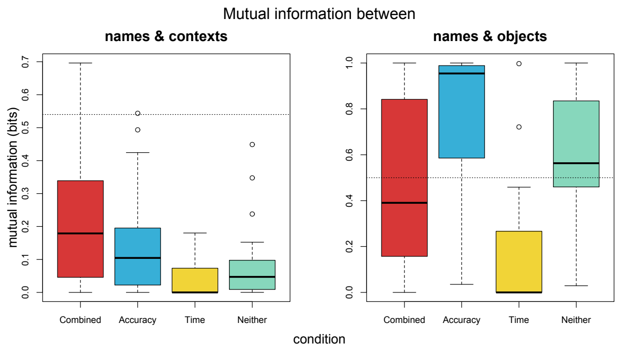

Figure 3: This figure shows the extent to which participants' name choices are conditioned on context (lefthand graph) and object (righthand graph). The dotted line in the lefthand graph represents the mutual information (MI_C) associated with the 'optimal' language in least-effort terms—the language in which the short form is used only in predictive contexts, and the long form only in surprising contexts. MI_C=0 for the input language. In the righthand graph, mutual information (MI_O) can range from 0 (same name fixed for both objects) to 1 (distinct names fixed for each object). MI_O=0.5 for the input language, marked by the dotted line. Data from only the second half of testing trials is shown in this figure, as participants were more likely to have converged on a stable mapping by this time.

and intercepts for context. The model was sum coded, setting the grand mean as the intercept, to which each level was then compared. The results yielded a significant positive interaction of context in the critical Combined condition ($\beta = 0.619, SE = 0.158, p < 0.001$), indicating that in this condition, participants were significantly more likely to use the short name in predictive contexts. The only other significant effects found were as follows: a positive overall effect in the Time condition ($\beta = 2.187, SE = 0.292, p < 0.001$), indicating that participants were more likely to use the short form in this condition regardless of context; a negative overall effect in the Accuracy condition ($\beta = -1.470, SE = 0.233, p < 0.001$), indicating that participants were *less* likely to use the short form in this condition regardless of context; and finally a negative interaction effect of context in the Accuracy condition ($\beta = -0.490, SE = 0.161, p = 0.002$), indicating that in fact participants were *even* less likely to use the short form in the predictive context in this condition.

An analysis of how participants conditioned the variation in their name usage sheds further light on the differing patterns of behaviour seen across conditions. We calculated the average mutual information between name produced and context (predictive or not) in each participant's output language (MI_C). The more reliably participants are conditioning their use of the long and short names on context, the higher we would expect the value of MI_C to be. The distributions for all four conditions are plotted on the lefthand graph of Fig. 3.

We also calculated the average mutual information between name produced and *object* (the blue fruit or the red stalk) in each participant's output language (MI_O). This measure allows us to determine whether some participants are using fixed names for each object, regardless of context. The results are plotted by condition in the righthand graph of Fig. 3. If participants are using a distinct name for each object, MI_O will be close to 1; if they are using the same name for both objects, MI_O will be close to 0. The former pattern is what we see in the Accuracy condition: most participants use the unique long name for each object, regardless of context. The latter pattern is what we see in the Time condition: most participants use the ambiguous short form for both objects, regardless of context.

In the Combined and Neither conditions, MI_O hovers around that of the input language. Based on this graph alone, participants may be probability matching in both these conditions, or perhaps reliably conditioning their output on other factors. Looking back at MI_C disambiguates: it is significantly higher in the Combined condition than in any other condition. A linear regression on MI_C with condition as predictor variable (fit to the second half of testing trials, as in Fig. 3) yielded a significant negative effect of the Accuracy ($\beta = -0.081, SE = 0.033, p = 0.016$), Time ($\beta = -0.184, SE = 0.041, p < 0.001$), and Neither ($\beta = -0.128, SE = 0.041, p = 0.002$) conditions, with the Combined condition set as the intercept. This result is consistent with what we saw in Fig. 2: in the Combined condition, many participants are optimally conditioning their responses on context, generating systems that fall in the bottom right corner of the graph; in the other conditions, almost no data points fall in this region.

Discussion

There is mounting evidence that utterance length is linked to information content (Lieberman, 1963; Aylett & Turk, 2004; Gahl & Garnsey, 2004; Tily et al., 2009; Bell et al., 2009; Jaeger, 2010; Piantadosi et al., 2011; Kuperman & Bresnan, 2012; Fedzechkina et al., 2012; Seyfarth, 2014). The explanation put forth in much of this previous work is that speakers are driven by pressures much like those outlined in Zipf's Principle of Least Effort: the competing demands for accurate and efficient communication lead speakers to converge on an optimal system in which information content is spread roughly uniformly across the utterance, resulting in low-information units being shorter than high-information units. This resultant effect appears to have made its way into the structure of the lexicon as a whole: shorter words appear on average in more predictive contexts than longer words (Piantadosi et al., 2011). But is this effect really due to the proposed mechanism? Can speaker choice lead to the reshaping of a lexicon to align it with the principles of Uniform Information Density and Smooth Signal Redundancy?

Here, we presented the first study that concretely addresses these questions. Previous studies either lacked a manipulation of the communicative pressures operating in the task, or lacked a communicative element entirely. In our study, by observing participants' online behaviour in a task in which the pressures to communicate accurately and efficiently were manipulated across four experimental conditions, we have shown that participants use shorter words in more predictive contexts *only* when both competing pressures were acting on them. When these pressures were isolated or removed entirely, participants failed to reliably condition their word choices on context.

Furthermore, because our study employed an artificial language learning paradigm, our findings avoid potential confounds from factors such as register, prosody, and participants' learned preferences in their native or second languages. Our results are nevertheless consistent with previous findings that language-users tend to use shorter forms in more predictive contexts when using their native language.

Our results serve as a proof of concept that the lexicon-level effect observed by Piantadosi et al. (2011) could be driven at least in part by a least-effort principle in which language-users balance the competing pressures for communicative accuracy and efficiency to reshape the lexicon into one where word length is roughly proportional to information content. However, there is a crucial step between what we have observed here—language-users alternating between long and short variants for a single meaning depending on context—and what Piantadosi et al. (2011) observed in the lexicon of different languages, where most meanings don't correspond to both a long and a clipped variant, but rather map to a single fixed form. For these cases, which make up the majority of the lexicon, the length of the form is strongly correlated with the *average* predictability-in-context of the meaning, across all its different occurrences. We can hypothesise a link between these two phenomena: as a word appears in increasingly more predictive contexts, a reduced variant may come into use. If speakers use the reduced variant in predictive contexts, then this reduced form will consequently become much more frequent than the long form, leading to the long form eventually dying out altogether. This would end in a scenario where a short word, with no alternative variants currently in use, appears on average in a high number of predictive contexts, and thus has a low average information content. Though this story sounds reasonable, a precise mechanistic explanation of how this preference for short forms in more predictive contexts leads to permanent shifts in form-meaning mappings has yet to be thoroughly investigated. We hope this topic is given more attention in future work.

References

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.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, *60*(1), 92–111.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012, October). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*(44), 17897–17902.

Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language*, *91*, 158–180.

Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, 748775.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language learning and development*, *1*(2), 151–195.

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

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*. (In press.)

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Kuperman, V., & Bresnan, J. (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language*, *66*(4), 588–611.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and speech*, *6*(3), 172–187.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013, February). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155.

Tanimura, S., Kuroiwa, C., Mizota, T., et al. (2006). Proportional symbol mapping in R. *Journal of Statistical Software*, *15*(i05).

Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., & Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition*, *1*(2), 147–165.

Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, *7*(3), 415-449.

Zipf, G. K. (1935). *The psycho-biology of language* (Vol. ix). Oxford, England: Houghton Mifflin.