INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Relationships between Functional Load and Auditory Confusability under Different Speech Environments

Shinae Kang[1], *Clara Cohen*[2]

[1]Georgetown University, USA
[2]The Pennsylvania State University, USA
sk1718@georgetown.edu, cpccohen@psu.edu

Abstract

Functional load (FL) is an information-theoretic measure that captures a phoneme's contribution to successful word identification. Experimental findings have shown that it can help explain patterns in perceptual accuracy. Here, we ask whether the relationship between FL and perception has larger consequences for the structure of a language's lexicon. Since reducing FL minimizes the risk of misidentifying a word in the case where a listener inaccurately perceives the initial phoneme, we predicted that in spoken language, where perceptual accuracy is important for successful communication, the lexicon will be structured to reduce FL in auditorily confusable initial phonemes more than in written language. To test this prediction, we compared FL of all initial phonemes in spoken and academic written genres of the COCA corpus. We found that FL in phoneme pairs in the spoken corpus is overall higher and more variable than in the academic corpus, a natural consequence of the smaller lexical inventory characteristic of spoken language. In auditorily confusable pairs, however, this difference is relatively reduced, such that spoken FL decreases relative to academic FL. We argue that this reflects a pressure in spoken language to use words for which inaccurate perception does minimal damage to word identification.

Index Terms: functional load, genre-specific lexical choice, auditory confusability

1. Introduction

As anyone who has attempted to hold a conversation with a person in the next room can attest, not all communication is successful. A distracted listener, an inarticulate speaker, or an intervening door can all contribute to failed transmission of the message. For communicative success, utterances must be structured in such a way as to make misperception of crucial information unlikely. Theories of audience design hold that speakers do this explicitly, by intentionally adjusting their production based on their evaluation of the needs of the listener [1, 2]. Theories of efficient communication, on the other hand, hold that speakers do not consider the needs of the listener directly, but rather modulate their utterances to hold constant the rate at which information is transferred [3, 4, 5, 6]. In both theories, however, decisions that speakers make drive only the forms of the individual utterance. In this project, we investigate whether these pressures that influence the form of individual utterances have larger consequences for the structure of the English lexicon. Specifically, we explore the differences between written English and spoken English. These two genres are distinguished in large part by the fact that spoken English is subject to communicative failure due to misperception of acoustic cues, while written English is insensitive to this source of error. We ask whether the structures of the written and spoken English lexicons reflect the different pressures inherent in preventing communication failure in spoken speech compared to written language.

1.1. Functional load

As a way of capturing the structure of a lexicon, we used a metric called functional load (FL) of a phoneme across the lexicon. FL is one of the measures adopted to represent information in information-theoretic approaches. Although other measures, such as the information content in a word, have been used to explain behaviors such as stop deletion during production or phonetic reduction [7, 8] (see also Section 1.2 in [9] for review), we chose functional load because it can capture the relationship between the structure of the lexicon and individual phonemes, rather than words, in a principled way.

To understand FL, consider first a language with a lexicon L, which can be conceptualized as the finite set of all speech elements σ of a given size, such as the syllable or word. The base entropy of that lexicon, $H(L)$, represents the amount of inherent uncertainty associated with selecting an element σ from L. As defined in Equation 1, entropy is the weighted sum of the log probabilities of selecting each element σ_i from L. Crucially, entropy increases as the probabilities of each element in the lexicon become more evenly distributed, and also as the size of the lexicon increases.

$$H(L) = -\sum_{i=1}^{N_L} Pr(\sigma_i) log_2 Pr(\sigma_i). \quad (1)$$

The FL of a phoneme x with respect to another phoneme y, expressed $FL(x, y)$, is defined as the relative difference between the base entropy $H(L)$ and the changed entropy of the lexicon after the contrast between x and y is neutralized $H(L^*_{xy})$ [9, 10, 11]

$$FL(x, y) = \frac{H(L) - H(L^*_{xy})}{H(L)} \quad (2)$$

Then, the overall FL of a phoneme x, expressed $FL'(x)$, is the sum of all such pair-wise FL $FL(X, y)$, multiplied by 1/2 for normalization purposes [9];

$$FL'(x) = \frac{1}{2}\sum_{y} FL(x, y) \quad (3)$$

Intuitively, FL captures the importance of a particular phoneme in identifying distinct words in a language's lexicon. Consider the phoneme [b]. This sound forms many minimal pairs with the phoneme [p] (e.g., *bat/pat*; *batter/patter*; *black/plaque*; *beer/peer*; *beak/peak*, and so on). If the contrast

http://dx.doi.org/10.21437/Interspeech.2016-906

between [p] and [b] is neutralized in English, then all of those words will now be homophones, and the size of the lexicon will shrink by the number of minimal pairs that are no longer distinct. This reduction in the size of the lexicon has the effect of reducing the entropy of the lexicon of English. By summing up all these reductions in entropy for all contrasts between [b] and the other phonemes of English, we arrive at FL, or the work that [b] performs in rendering words distinct from other words in a lexicon. Thus, the higher the FL of a phoneme, the more minimal pairs it is responsible for distinguishing, and hence the more likely it is that misidentifying that phoneme will result in misidentifying the target word.

FL has proven a useful tool for explaining sound-related linguistic phenomena. Sound mergers, for example, are more likely between phonemes that have lower pairwise FL [12], and people are more likely to correctly distinguish a pair of phonemes when the FL of that pair is higher [9]. The current project asks whether FL can also help explain the structure of the English lexicon.

1.2. Confusability

Not all phonemes are equally likely to be confused with each other (cf. confusion matrices in [13, 14]). It is extremely unlikely that a voiceless postalveolar affricate [ʧ] will be misheard as a vowel like [a]; yet it is quite likely that it might be mistaken for the voiceless postalveolar fricative [ʃ]. Acoustic information is crucial in perceptual identification, and can interact in important ways with FL. For example, there might be little danger of communication failure in using a phoneme with high FL if it is never misperceived, while there could be a great danger of communicative failure if a phoneme with low FL is so confusable with many other phonemes that it is frequently misperceived.

Confusability is closely associated with perceptual similarity between two sounds. Among several methods that could induce a similarity metric from confusion matrices [15, 16], we use the Phi-value, as described in [17, 18]. The Phi-value, or Phi-square statistic, is the distribution describing the perceptual similarity of two phonemes x and y, derived from the comparison the two phonemes' experimental response distributions. It as defined in (4) below:

$$P(ID) = 1 - \sqrt{\frac{\sum \frac{(x_i - E(x_i))^2}{E(x_i)} + \sum \frac{(y_i - E(y_i))^2}{E(y_i)}}{N}} \quad (4)$$

where N is the total number of responses, x_i is the frequency with which x was identified as category i, and y_i is the frequency with which y is identified as category i. $E(x_i)$ and $E(y_i)$ represent the expected rates at which x and y, respectively, are identified as category i, and are defined as the sum of all frequencies with which x and y, respectively, are identified as i, divided by 2. When x_i and y_i are perceptually identical, $E(x_i) = E(y_i)$.

The Phi-value is a useful measure of perceptual similarity, because the more similar x and y, the closer the term under the square root approaches to 0, and hence the closer $P(ID)$ is to 1. The phi-value thus ranges between 0 (completely distinct, minimally confusable) and 1 (perceptually identical, maximally confusable) It has an advantage over more traditional probabilistic measure of confusability, because response biases and asymmetric data are normalized.

The hypothesis driving this investigation is that issues of confusability drive the decisions people make when producing spoken language more than in written language, and that this has distinct effects on the emergent lexicon used in spoken compared to written English. Crucially, we posit that, in spoken English, FL in more confusable pairs of phonemes should be reduced relative to less confusable pairs. In other words, since misperceiving phonemes with higher FL raises the risk of communication failure, people producing spoken language will select lexical items that minimize FL in the case where those items are especially confusable and hence vulnerable to misperception. In written language, however, confusability is not an issue, and so choices of lexical items will not be constrained by such a pressure, and the relationship between FL and confusability will be less pronounced in written English. We tested this prediction by examining the relationship between FL and confusability of initial phonemes in the lexicons extracted from a corpus of written and spoken English. We restricted our analysis to the role of initial phonemes because words are perceived incrementally over time [19], so the maximum danger of misidentifying a word results from misperceiving the initial phoneme.

2. Method

2.1. Corpus

2.1.1. Corpus of Contemporary American English (COCA)

Information about lexicon size was drawn from the Corpus of Contemporary American English [20], a 440 million word corpus that is split into five evenly-sized subcorpora of different genres of English: spoken, fiction, magazine, newspaper, and academic. For this project, the subcorpora containing unscripted spoken English (spoken) and written academic prose (academic) were selected, as we considered that the choices which people make in forming spoken utterances and the pressures affecting comprehension would be maximally distinct from the choices and pressures relevant in written academic prose.

2.1.2. Preprocessing

In each sub-corpus, all words were extracted by Python script, and their usage frequency was recorded. Additionally, their first phones were determined by matching the unique lexical items with the entries in the Carnegie Mellon University (CMU) pronouncing dictionary [21]. All lexical items that did not have an entry in the CMU dictionary were discarded. This extraction provided the frequency with which each phoneme in English served as the initial phoneme of a word in each corpus, as well as the frequency of every word in each corpus. Overall, there were 24 distinct initial phonemes represented in 76,163 distinct word types in the spoken sub-corpus, and 95840 types in the academic sub-corpus. Note that since the corpora are balanced at roughly 88 million *tokens* each, the different sized lexicons reflect the fact that academic prose draws on a larger set of words than spoken English. In other words, the different lexicons do not represent unbalanced corpora, but genuine differences in the shape of the two lexicons.

2.2. Statistical analysis

Pairwise and general FL of initial phonemes extracted from both the academic written and and spoken sub-corpora was calculated following the formulas in Equations 2-3. The Phi-values for each phoneme were extracted from [11]. In theory, there were 576 (24 x 24) phi-values for each pair of the initial consonant phonemes. However, because Phi-square defines the con-

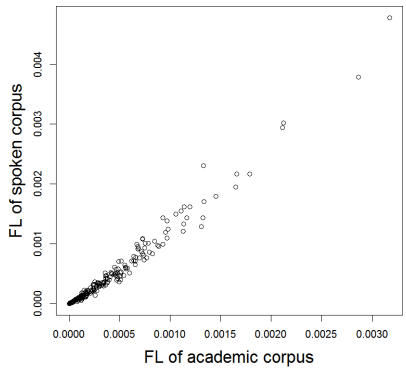

Figure 1: *Correlation between FL of academic corpus and of spoken corpus.*

Table 1: Regression results

	Dependent variable:
	Spoken FL
Academic FL	1.324^{***} (0.013)
Intercept	-0.00005^{***} (0.00001)
Observations	276
R^2	0.974
Adjusted R^2	0.973
Residual Std. Error	0.0001 (df = 274)
F Statistic	$10{,}092.510^{***}$ (df = 1; 274)
Note:	*p<0.1; **p<0.05; ***p<0.01

fusability of x to y and y to x as identical, only the upper diagonals of the matrix (576-24)/2=276 were taken and aligned with FL.

The underlying idea under investigation here is that the degree to which more confusable phoneme pairs (as indexed by Phi-value) play a role in word meaning differentiation (as indexed by FL) will be reduced in spoken language compared to written language. Therefore, we would expect reduced FL for confusable pairs in the spoken corpus relative to the academic corpus.

The first step of our analysis was designed to evaluate the baseline similarity between the FL in the two corpora. To this end, we performed a paired t-test and correlation test on the respective FL in each corpus for each of the 276 pairs of initial phonemes. The second step of our analysis was to see if the more confusable sound pairs have lower FL in the spoken corpus than in the academic corpus. To this end, we focused on the *difference* between academic FL and spoken FL. A greater difference means that spoken FL is smaller relative to academic FL. We analyzed the relationship between FL and confusability by regressing the FL differences in the two corpora for each phoneme pair against that pair's confusability. The prediction is that more confusable pairs in the spoken corpus have lower FL compared to those same pairs in the academic corpus. In the regression we therefore expect a positive effect of confusability: As phoneme pairs become more confusable, the *difference* between academic FL and spoken FL should become larger, indicating lower spoken FL compared to academic FL.

3. Results

3.1. Comparison of academic and spoken corpora

As shown in Figure 1, the correlation between FL of the spoken corpus (spoken FL) and that of the academic corpus (academic FL) is very high ($r = 0.99, t = 100.46, df = 274, p < 0.01$*), meaning that the distribution of FL across phoneme pairs is extremely similar across the two corpora. This is to be expected, given that they are both corpora of English, drawing on substantially overlapping lexicons. Yet the distributions are not identical. A paired t-test reveals that in general, spoken FL is quite significantly higher than academic FL ($t = 5.64, df = 275, p < 0.01$*). Table 1 presents the results of a simple regression model predicting spoken FL from academic FL. In this model, the coefficient of academic FL is greater than 1, indicating that the range of FL is not identical. For every one-unit increase in academic FL, there is a corresponding increase of more than one unit in spoken FL, which means that there is a greater range of spoken FL values than academic FL values.

3.2. Relationship between FL and confusability

3.2.1. Simple correlation

FL from both the academic corpus and spoken corpus showed a significant negative correlation with the confusability (for spoken FL: $r = -0.12, t = -2.07, df = 274, p < 0.05$* ; and for academic FL: $r = -0.13, t = -2.10, df = 274, p < 0.05$*). Given 0.99 of correlation between the two corpora found in Section 3.1, such a similar pattern in the correlation with confusability is not surprising. The negative correlation coefficient indicates that more confusable phoneme pairs tend to have lower FL. In other words speakers are less likely to use the words beginning with sounds that are more confusable if those sounds are important in differentiating word meanings, in both the academic and spoken corpus.

3.2.2. Relating FL differences to auditory confusability

The negative correlation between confusability and both spoken and academic FL indicates a dispreference for auditorily confusable initial phonemes with high FL. To see whether this dispreference is stronger in the spoken corpus, where auditory confusability has greater implications for communicative success than in the academic corpus, we subtracted the spoken FL from the academic FL for each phoneme pair, and used the resulting difference to represent the FL of the spoken corpus relative to the academic corpus. Since spoken FL is overall much higher than academic FL, this difference is almost always negative. It increases, however, becoming less negative, when spoken FL is lower relative to academic FL.

We used this difference as a dependent variable in a correlation test and simple regression model with confusability, or Phi-value, as an independent variable. As shown in Figure 2, most of the spoken/academic FL differences are close to 0, with departures from 0 mostly negative, corresponding to relatively higher values of spoken FL compared to academic FL. Crucially, these departures are primarily at the lower values of confusability. This pattern yields a trend towards a positive effect of Phi-value on the difference between academic and spoken FL. To probe this relationship more deeply, we ran a linear regression model predicting the difference between academic and spoken FL as a function of Phi-value. Table 2 summarizes this

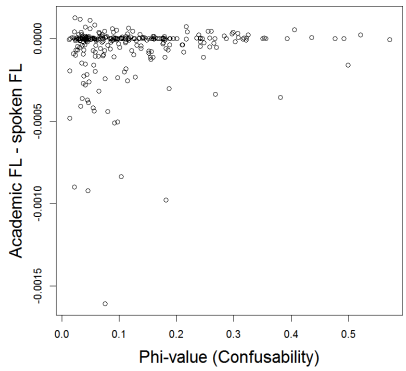

Figure 2: *Correlation between the difference of spoken and academic FL (y-axis) and confusability (x-axis).*

Table 2: Regression results

	Dependent variable:
	Academic FL–spoken FL
Confusability	0.0002^{*} (0.0001)
Intercept	-0.0001^{***} (0.00002)
Observations	276
R^2	0.011
Adjusted R^2	0.007
Residual Std. Error	0.0002 (df = 274)
F Statistic	2.944^{*} (df = 1; 274)
Note:	*p<0.1; **p<0.05; ***p<0.01

model, showing a negative intercept and positive effect of Phi-value. The negative intercept reflects the fact that spoken FL is overall greater than academic FL for the least confusable pairs (i.e., those with a Phi-value of 0), while the positive effect of Phi-value indicates that as confusability increases, spoken FL decreases relative to academic FL.

4. Discussion

4.1. The spoken vs. academic corpus

The high correlation between the functional loads in the two corpora is to be expected, since speakers of the same language will exhibit largely similar linguistic behavior irrespective of genre or contexts. Yet the pairwise comparison of FL between the corpora showed a clear difference between the spoken and academic language, with FL of a spoken corpus being significantly higher than FL in the academic corpus, and more variable. This difference could reflect the diversity of the lexicon comprising each corpus. Since FL increases as a function of the *relative* reduction of lexicon size caused by neutralizing a distinction between two phonemes, then the same absolute reduction in lexicon size will yield a higher FL in a smaller lexicon. As we observed here, the size of the spoken lexicon was substantially smaller than the size of the written lexicon, so it is only natural that there would be higher and more variable FL across phoneme pairs in the spoken corpus.

4.2. Relationship between FL and confusability

The first finding of the study is that there is a negative correlation between FL and confusability in both spoken and academic corpus. This pattern in the lexicon emerges naturally if speakers select lexical items so that the danger of misperceiving a word balances out the damage such misperception might cause to comprehension of the message. Highly confusable phonemes tend to have low FL, so that if they are misperceived, as is likely, the danger to word identification is minimized, while high FL phonemes tend to be less confusable, so that the misperception, and its ensuing damage to communicative success, is less likely. Over time, these word choices could cause the lexicon of English to shift so that lexical items balancing FL and confusability in their initial phonemes will be favored. Even though misperception is not an issue in written language, it draws on substantially the same lexicon as spoken English, and so this emergent relationship between FL and confusability can still be observed.

What was most intriguing about these findings was the trend towards a higher FL in the academic corpus relative to the spoken corpus as confusability increased. This result implies that more word differentiation in the spoken corpus is done by less confusable phonemes than by more confusable phonemes, compared to the academic corpus. If this reflects a real effect, then the preference for balanced FL and confusability is still present and active in our language, and more active in spoken English than written English.

4.3. Possible limitations and future directions

The relationships we observed here were not robust, but rather suggestive trends. It is possible that there is so very little difference between the spoken and academic FL that any effect of auditory confusability and functional load is necessarily minimal. However, to the extent that there are differences, it is necessary to explain their source, and the work presented here proposes one such account: namely, that auditory confusability affects FL distributions differently in contexts where auditory accuracy is crucial for communicative success.

Although initial phonemes are the earliest possible source of miscommunication in word recognition, they are not the only source. In principle, misperceiving any phoneme before a word's uniqueness point [22] could lead to communicative failure. A more thorough analysis, therefore, would examine the relationship between auditory confusability and FL of all pre-uniqueness-point phonemes, with the prediction that any such relationship would dissipate for phonemes the follow a word's uniqueness point. Other directions include comparing the distributions of all five COCA subcorpora, to see whether patterns in other genres can shed further light on how people balance FL and auditory confusability; performing similar analyses in other languages, to see whether the distinction between written and spoken English reflects a general desire to minimize communicative failure across all language users.

5. Conclusion

In this work we tested the hypothesis that the communicative pressures to avoid misperception affect spoken and written language differently, and as a result the structures of the lexicons employed in the two modalities have evolved to be slightly different. It would be interesting to return in 100 years, perform the same analysis, and see if structures of the spoken and written English lexicon have diverged further.

6. References

[1] J. Liljencrants and B. Lindblom, "Numerical simulation of vowel quality systems: the role of perceptual contrast," *Language*, vol. 48, pp. 839-862, 1972.

[2] B., Lindblom, "Phonetic universals in vowel systems," *Experimental Phonology*. Orlando: Academic Press, pp. 13–44, 1986.

[3] M. Aylett and A. Turk, "The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," *Language and Speech*, vol. 47, No. 1, pp. 31–56, 2004.

[4] T. F. Jaeger, "Redundancy and reduction: Speakers manage syntactic information density," *Cognitive Psychology*, vol. 61, no. 1, pp. 23–62, 2010.

[5] R. Levy and T. F. Jaeger, "Speakers optimize information density through syntactic reduction," *Proceedings of the 20th Conference on Neural Information Processing Systems (NIPS)*, 2007.

[6] C. Kuramada and T. F. Jaeger, "Communicatively efficient language production and case-marker omission in Japanese," *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 2013.

[7] U. Cohen Priva, "Using information content to predict phone deletion," *Proceedings of the 27th West Coast Conference on Formal Linguistics*, pp. 90–98, 2008.

[8] S. Seyfarth, "Word informativity influences acoustic duration: effects of contextual predictability on lexical representation," *Cognition*, vol. 133, pp. 140–55, 2014.

[9] S. Kang, "Relationship between perceptual accuracy and information measures: A cross-linguistic study," PhD Dissertation, UC Berkeley, 2015.

[10] C. F. Hockett, "The quantification of functional load: A linguistic problem," *Word*, vol. 23, pp. 320–339, 1967.

[11] D. Surendran and P. Niyogi, "Quantifying the Functional Load of Phonemic Oppositions, Distinctive Features, and Suprasegmentals," *Competing Models of Language Change: Evolution and Beyond*, Amsterdam; Philadelphia, PA: John Benjamins, pp. 43–58, 2006.

[12] A. Wedel, A. Kaplan, and S. Jackson, "High functional load inhibits phonological contrast loss: a corpus study," *Cognition*, vol. 128, pp. 179–186, 2013.

[13] G. A. Miller and P. E. Nicely "An Analysis of Perceptual Confusions among some English Consonants," *Journal of the Acoustical Society of America*. vol. 27, pp. 338–352, 1955.

[14] J. M. Pickett, "Perception of vowels heard in noises of various spectra," *Journal of the Acoustical Society of America*. vol. 29, pp. 613–620, 1957.

[15] R. N. Shepard, "Psychological representation of speech sounds," In E. E. David, Jr.& P. B. Denes (Eds.), *Human Communication: A Unified View*, New York: McGraw-Hill, pp. 67–113, 1972.

[16] P. A. Luce and D. B. Pisoni, "Recognizing spoken words: The neighborhood activation model," *Ear & Hearing*, vol. 19(1), pp. 1–36, 1998.

[17] P. Iverson, L. Bernstein, and E. Auer, Jr., "Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition," *Speech Communication*, vol. 26, no. 1-2, pp. 45–63, 1998.

[18] J. F. Strand and M. S. Sommers, "Sizing up the competition: Quantifying the influence of the mental lexicon on auditory and visual spoken word recognition," *Journal of the Acoustical Society of America*. vol. 130, pp. 1663–1672, 2011.

[19] J. Magnuson, J. Dixon, M. Tanenhaus, and R. Aslin, "The dynamic of lexical competition during spoken word recognition," *Cognitive Science* vol. 31, pp. 1–24.

[20] M. Davies, "The corpus of contemporary American English as the first reliable monitor corpus of English," *Literary and Linguistic Computing*, vol. 24, No. 4, pp. 447–464, 2010.

[21] R. L. Weide, "CMU Pronunciation Dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 1994.

[22] L. H. Wurm, M. T. C. Ernestus, R. Schreuder, and R. H. Baayen, "Dynamics of the auditory comprehension of prefixed words: Cohort entropies and Conditional Root Uniqueness Points," *The Mental Lexicon*, vol. 1, No. 1, pp. 125-146, 2006.