Towards the Use of Pretrained Language Model GPT-2 for Testing the Hypothesis of Communicative Efficiency in the Lexicon

1st Yuqing Zhang
School of Information Science
Beijing Language and Culture University
Beijing, China
yuqingelsa@gmail.com

2nd Zhu Li
School of Information Science
Beijing Language and Culture University
Beijing, China
lzblcu19@gmail.com

3rd Jinsong Zhang
Research Institute of International Chinese Language Education
Advanced Innovation Center for Language Resources and Intelligence
School of Information Science, BLCU
Beijing, China
jingsong.zhang@blcu.edu.cn

***Abstract*—Functional load (FL) is a quantitative measure that computes a phonological contrast's information contribution to successful word identification. Work on communicative efficiency in the lexicon has demonstrated that words rely more on phonological contrasts in phonological contexts where those contrasts are readily perceived by the listener. However, these studies have only examined the preference for perceptual distinctiveness based on FL as the number of minimal pairs. This study demonstrates that recent progress in language modeling pre-training can be effectively utilized to quantify the information contribution of phonemic contrasts under contexts and bring new evidence to the hypothesis of communicative efficiency in the lexicon. Building on top of the GPT-2 model and the text-phoneme-text transmission model, we calculated the information contributions of English consonants and vowels. Results indicated that they were strongly correlated FL values based on other computational algorithms. In addition, we computed information contributions and perceptual distinctiveness of consonants in different phonological contexts. Results indicated that phoneme contrasts had higher information contributions in the word-initial context, i.e., the context where those contrasts were readily perceived. Altogether, these results show the feasibility of using the pre-trained language model GPT-2 to test the hypothesis of communicative efficiency in the lexicon.**

***Index Terms*—pre-trained language model, communicative efficiency, functional load, perceptual distinctiveness, phonological context**

I. Introduction

Languages use contrasts of features to transfer information or convey messages in social environments. As stated in [1], "the function of a phonemic system is to keep the utterances of a language apart". Analyzing phoneme inventories and the organization of phonemic systems is crucial for our understanding of information transmission in speech communication. Many linguistic phenomena show that different phonological oppositions (or the same phonological opposition in different languages) exhibit different levels of importance. For example, in the evolutionary process of language, the presence of some contrasts has little impact on information transmission, and phonemes in these contrasts have merged into a unique symbol.

In order to be an efficient communicative system, languages may be structurally shaped by communicative efficiency so as to achieve easy, rapid, and robust communication. Guided by this theoretical framework, a series of experiments conducted by [2] have shown that the phonological lexica of natural languages are globally optimized for the recoverability of words so as to facilitate efficient message transmission. For example, gathering a sample of 49 languages from 25 different language families, [2] demonstrated that word distinctions preferentially rely on perceptible contrasts for distinctness. [3] investigated the relationship between auditory confusability and FL in written and spoken English, and suggested that due to pressures inherent in preventing communication failure in spoken speech, the distinction exists in the structure of written and spoken lexicons.

As a quantification of the amount of work a contrast does in conveying information/distinguishing meanings in a language, the notion of the functional load of a phonological contrast arose and developed several decades ago and various modeling frameworks have been proposed to find the functional load of phonemic oppositions. In its simplest expression, functional load can be computed as the number of lexical minimal pairs for a given contrast [4,5]. Taking word and syllable structure into consideration, Surendran and Niyogi proposed an information-theoretical measure (i.e., the overall loss of information in a language induced by the merger of a contrast) to quantify the importance of phonemic oppositions, distinctive features, and suprasegmentals [4-6]. It has been shown that FL derived from these definitions may play an important role in language change [5,9], phoneme acquisition by children [10,11], and foreign language acquisition [12]. In consideration of the fact that different types of contexts (lexical, syntactic, semantic, and interpretative) influence the process of recognizing a spoken word, [13] utilized n-gram language models capable of modeling contextual information of individual words to model information contributions of phonemic contrasts and computed FL as the change in

mutual information of spoken texts and phoneme sequences (F) induced by the merger of a phoneme pair. However, conventional n-gram models used in [13] are still weak in modeling contextual information.

Recent progress in training high-capacity auto-regressive language model GPT-2 [14] on large datasets opens the question of whether such models can be useful in estimating information contributions of phonemic contrasts. The GPT-2 was trained on a massive 40GB WebText dataset that the researchers crawled from the internet. It has been successfully used to estimate log probabilities for a sentence via the chain rule [15].

In this paper, we demonstrate that the language model GPT-2 pre-trained on large general-domain corpora can be effectively utilized to quantify the information contributions of phonemic contrasts under contexts, and FL calculated using GPT-2 can be effectively used to test the hypothesis of communicative efficiency in the lexicon, e.g., phoneme contrasts had higher information contributions in the more perceptible word-initial context.

II. Methods

A. FL as change in mutual information

In the Text-Phoneme-Text transmission modeling framework, FL of a phoneme pair can be calculated as the change in mutual information of spoken texts (W) and phoneme sequences (F) induced by the merger of a phoneme pair α [13], as shown in equation 1. Since upon merger of a phonemic contrast, the number of word sequences sharing the same phoneme transcription will increase, hence the mutual information between the spoken text and the phoneme transcription will decrease as compared to before. The mutual information loss reflects the reduction of the amount of shared information due to the merger of a phoneme pair, and thus it can be utilized to quantify information contributions of phonemic contrasts.

$$\mathrm{FL}(\alpha) = \frac{\mathrm{I(W;F)} - \mathrm{I\,(W;F_{\alpha})}}{\mathrm{I(W;F)}} \tag{1}$$

According to the Shannon-McMillan-Breiman theorem [17], we can mathematically derive the formula 2, in which $W'_1, W'_2, \ldots, W'_m$ are all text sequences sharing the same transcription F. The probability of the text sequence $P\,(W'_i)$ can be efficiently computed by language models. The high-capacity auto-regressive language model GPT-2, with high contextualized information, naturally becomes a candidate to complete this task.

$$\mathrm{I(W;F)} = \lim_{\mathrm{n}\to\infty} -\frac{1}{\mathrm{n}} \log \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{P\,(W'_i)} \tag{2}$$

B. FL as the number of lexical minimal pairs

Functional load of a phonemic pair can be calculated by the number of occurrences of the target pair in the static lexicon or the number of occurrences of the minimal pair in the dynamic corpus (#MP). The higher the number of lexical minimal pairs based on a phonemic pair, the more this phonemic pair contributes to meaning distinctions of a language's lexicon and the more information this phonemic pair carries than other pairs. In this study, the open software Phonological Corpus Tools [18] was used to calculate the number of minimal pairs of a phonemic pair, and the parameter was set as the raw number of minimal pairs in the corpus.

C. Perceptual distinctiveness

A confusion matrix captures the frequencies of phoneme identification errors, and thus it can be used to compute perceptual similarity/distinctiveness estimates. Among several available measures, we applied the phi-square statistic to confusion data from a phoneme identification experiment [19]. The phi-square statistic characterizes the degree of perceptual distinctiveness of a phonemic pair x and y, derived from quantifying the similarity of the response distributions of two phonemes [16]. It is expressed mathematically as:

$$\Phi^2 = \sqrt{\frac{\sum \frac{(x_i - E(x_i))^2}{E(x_i)} + \sum \frac{(y_i - E(y_i))^2}{E(y_i)}}{N}} \tag{3}$$

D. Experimental setup

For the present study, we randomly sampled 425,500 sentences (around 2,000,000 word tokens) from the "TV and Movies subtitles" (TVM90-19) sub-corpus of the Corpus of Contemporary American English [20] and converted them into ARPAbet sequences using the CMU dictionary. Then we decoded ARPAbet sequences into all possible word sequences. To calculate $\mathrm{FL_{MI\_GPT2}}$, the pre-trained GPT-2 model was used to calculate sentence probabilities. To compute $\mathrm{FL_{MI\_trigram}}$, a trigram LM trained with the TVM90-19 corpus using KenLM [21] was used to score the lattice of all word sequences sharing the same phonological transcription.

III. Results and Discussion

A. Similarity in distribution patterns

Figure 1 shows the distribution of $\mathrm{FL_{MI\_trigram}}$, #MP and $\mathrm{FL_{MI\_GPT2}}$ for consonant pairs in English. It can be observed that FL values computed by the three methods share a common trend towards power law distributions. The general distribution curve can be roughly divided into two parts: the first part is for some phonemic oppositions ranked at the top, and the curve shows a steep descending trend; the second part is for most of the phonemic oppositions ranked at the bottom, and the curve is characterized by a slow decrease. This indicates that English may rely unevenly on the phonemic pairs in the phonological subsystem to carry information, i.e., the majority of phonemic oppositions in the phonological subsystem have relatively little information contribution. The shape of $\mathrm{FL_{MI\_trigram}}$ and $\mathrm{FL_{MI\_GPT2}}$ was more regular and smooth compared to #MP. One possible explanation is that with proper contextual modeling, information contributions of phonemic contrasts could be estimated more concretely. Thus the distribution becomes more similar to regular distributions found in other scientific disciplines.

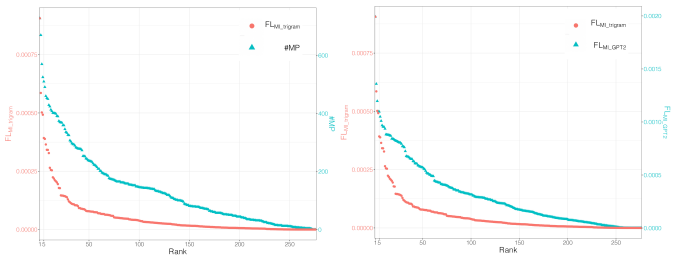

Fig. 1. Distribution of consonant pairs: (left)- $\mathrm{FL}_{\mathrm{MI\_trigram}}$ on the left y-axis (in red) and #MP on the right y-axis (in blue). (right)- $\mathrm{FL}_{\mathrm{MI\_trigram}}$ on the left y-axis (in red) and $\mathrm{FL}_{\mathrm{MI\_GPT2}}$ on the right y-axis (in blue). Pairs are listed by their decreasing order of FL values.

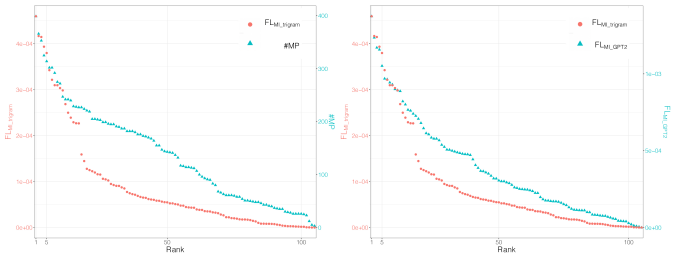

Fig. 2. Distribution of vowel pairs: (left)- $\mathrm{FL}_{\mathrm{MI\_trigram}}$ on the left y-axis (in red) and #MP on the right y-axis (in blue). (right)- $\mathrm{FL}_{\mathrm{MI\_trigram}}$ on the left y-axis (in red) and $\mathrm{FL}_{\mathrm{MI\_GPT2}}$ on the right y-axis (in blue). Pairs are listed by their decreasing order of FL values.

B. Correlations between different computational methods

To explore the feasibility of using the pre-trained language model GPT-2 to quantify the information contributions of phonemic contrasts and test the communicative efficiency hypothesis concerned in this study, correlation coefficients between different computational methods were calculated and shown in Tables I and II. It can be observed that $\mathrm{FL}_{\mathrm{MI\_GPT2}}$ has relatively strong correlations with both $\mathrm{FL}_{\mathrm{MI\_trigram}}$ and #MP. In comparison, correlation between $\mathrm{FL}_{\mathrm{MI\_trigram}}$ and #MP is relatively low. This result indicates that $\mathrm{FL}_{\mathrm{MI\_GPT2}}$ may be intermediate between the other two methods. With no contextual modeling, #MP lies at one extreme, while with strong contextual modeling using the test corpus, $\mathrm{FL}_{\mathrm{MI\_trigram}}$ lies at the other extreme. $\mathrm{FL}_{\mathrm{MI\_GPT2}}$ remedies their deficits since it is trained on general-domain corpora and closely approximates real-life communication.

TABLE I
CORRELATION COEFFICIENTS OF DIFFERENT MEASURES OF FL CARRIED BY THE CONSONANTAL CONTRASTS

Pearson's r	#MP	$\mathrm{FL}_{\mathrm{MI\_trigram}}$	$\mathrm{FL}_{\mathrm{MI\_GPT2}}$
#MP	1.00	0.31	0.72
$\mathrm{FL}_{\mathrm{MI\_trigram}}$	0.31	1.00	0.72
$\mathrm{FL}_{\mathrm{MI\_GPT2}}$	0.72	0.72	1.00

C. Effect of perceptual distinctiveness on functional load in different phonological contexts

1) Perceptual distinctiveness in different phonological contexts: Perceptual distinctiveness values calculated for the 21 phonemic oppositions of the English unvoiced stops and fricatives /p, t, k, f, θ, s, ʃ/ at word-initial and word-final positions are shown in Figure 3. One-tailed paired sample t-test indicates that, overall, phonemic pairs have higher perceptual distinctiveness at the word-initial position (M = 0.49, SD = 0.06) than at the word-final position (M = 0.42, SD = 0.09) (t(20) = 4.70, p <0.001). This result is in line with our expectation, as there is a perceptual advantage of consonant-initial syllables compared to consonant-final syllables, and the acoustic features of consonant-initial syllables have been found to be more intelligible [19].

TABLE II
CORRELATION COEFFICIENTS OF DIFFERENT MEASURES OF FL CARRIED BY THE VOWEL CONTRASTS

Pearson's r	#MP	$\mathrm{FL}_{\mathrm{MI\_trigram}}$	$\mathrm{FL}_{\mathrm{MI\_GPT2}}$
#MP	1.00	0.39	0.73
$\mathrm{FL}_{\mathrm{MI\_trigram}}$	0.39	1.00	0.79
$\mathrm{FL}_{\mathrm{MI\_GPT2}}$	0.73	0.79	1.00

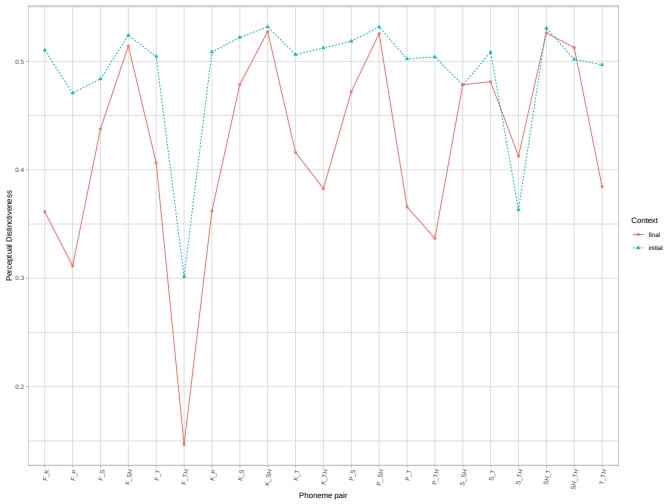

Fig. 3. Perceptual Distinctiveness of phonemic contrasts in different phonological contexts.

2) Functional load in different phonological contexts: $\mathrm{FL}_{\mathrm{MI\_trigram}}$ calculated for the 21 phonemic oppositions of the English unvoiced stops and fricatives /p, t, k, f, θ, s, ʃ/ at word-initial and word-final positions are shown in Figure 4. One-tailed paired sample t-test suggests that phonemic pairs have higher information contributions at the word-initial position (M = 2.68e-05, SD = 4.59e-05) than at the word-final position (M = 1.38e-05, SD = 1.76e-05) (t(20) = 2.72, p = 0.007).

The number of minimal pairs calculated for the 21 phonemic oppositions of the English unvoiced stops and fricatives /p, t, k, f, θ, s, ʃ/ at word-initial and word-final positions are shown in Figure 5. One-tailed paired sample t-test indicates that, overall, phonemic pairs have higher information contributions at the word-initial position (M = 97.14, SD = 54.06) than at the word-final position (M = 44.67, SD = 31.27) (t(20) = 4.70, p <0.001).

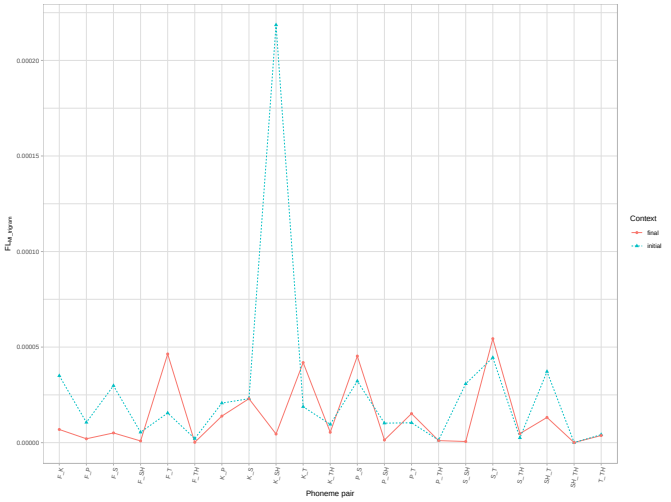

Fig. 4. $\mathrm{FL}_{\mathrm{MI\_trigram}}$ of phonemic contrasts in different phonological contexts.

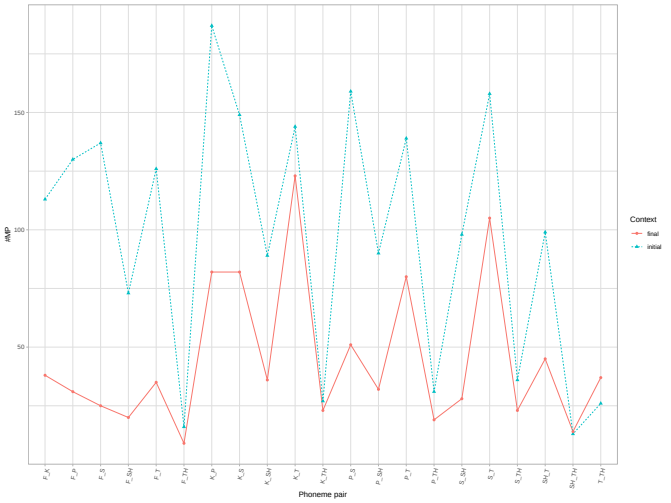

Fig. 5. #MP of phonemic contrasts in different phonological contexts.

$\mathrm{FL}_{\mathrm{MI\_GPT2}}$ calculated for the 21 phonemic oppositions of the English unvoiced stops and fricatives /p, t, k, f, θ, s, ʃ/ at word-initial and word-final positions are shown in Figure 6. One-tailed paired sample t-test indicates that, overall, phonemic pairs have higher information contributions at the word-initial position (M = 1.59e-04, SD = 1.01e-04) than at the word-final position (M = 1.02e-04, SD = 1.04e-04) (t(20) = 2.89, p = 0.005).

Based on these results, it can be inferred that there are more lexical words in contexts relying on consonantal contrasts at the word-initial position to distinguish meanings. In other words, in speech communication, English makes more use of a contrast at the word-initial position than at the word-final position. These findings are in line with our expectation: for a language to be an efficient communicative system, phoneme contrasts in this language will have higher information contributions in the word-initial context, i.e., the context where those contrasts are easily perceived by listeners.

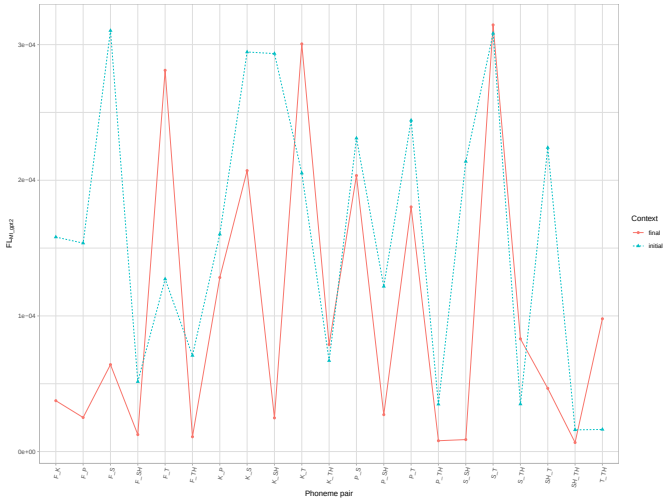

Fig. 6. $\mathrm{FL}_{\mathrm{MI\_GPT2}}$ of phonemic contrasts in different phonological contexts.

IV. Conclusions and Future Work

With the development of data and computational power, large-scale corpus-based studies in recent years have enabled the scientific test of traditional theoretical hypotheses. Using a large-scale spoken corpus and the high-capacity language model GPT-2, this study demonstrates that recent progress in pre-trained language models can be effectively utilized to quantify the information contribution of phonemic contrasts under contexts and bring new evidence to the hypothesis of communicative efficiency in the lexicon. We calculated the information contributions of English consonants and vowels as $\mathrm{FL}_{\mathrm{MI\_GPT2}}$ and results indicated that they were strongly correlated FL values based on other computational algorithms. In addition, we computed information contributions and perceptual distinctiveness of stops and fricatives in different phonological contexts. Results indicated that phoneme contrasts had higher information contributions in the word-initial context, i.e., the context where those contrasts were readily perceived. In conclusion, this study demonstrates the feasibility of using the pre-trained language model GPT-2 to test the hypothesis of communicative efficiency in the lexicon. For future work, we plan to perform similar analyses in other languages, to see whether this is a universal pattern across all languages. Besides, we'd like to use these computational models to test other related hypotheses of communicative efficiency in the human lexicon.

Acknowledgment

This study was supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Beijing Language and Culture University (21YCX180), Advanced Innovation Center for Language Resource and Intelligence (KYR17005), National Social Science Foundation of China (18BYY124), the Science Foundation and Special

Program for Key Basic Research fund of Beijing Language and Culture University (the Fundamental Research Funds for the Central Universities) (21YJ040004), Wutong Innovation Platform of Beijing Language and Culture University (19PT04). Jinsong Zhang is the corresponding author.

REFERENCES

[1] C. F. Hockett, "The quantification of functional load–a linguistic problem." 1966.

[2] P. N. H. M. Graff, "Communicative efficiency in the lexicon," Ph.D. dissertation, Massachusetts Institute of Technology, 2012.

[3] S. Kang and C. Cohen, "Relationships between functional load and auditory confusability under different speech environments." in *INTERSPEECH*, 2016, pp. 2821–2825.

[4] D. Ingram and I. David, *First language acquisition: Method, description and explanation*. Cambridge university press, 1989.

[5] A. Wedel, S. Jackson, and A. Kaplan, "Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change," *Language and speech*, vol. 56, no. 3, pp. 395–417, 2013.

[6] D. Surendran and P. Niyogi, "Measuring the functional load of phonological contrasts," *arXiv preprint cs/0311036*, 2003.

[7] ——, "Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals," *Amsterdam studies in the theory and history of linguistic science series 4*, vol. 279, p. 43, 2006.

[8] D. Surendran and G.-A. Levow, "The functional load of tone in mandarin is as high as that of vowels," in *Speech Prosody 2004, International Conference*, 2004.

[9] A. Wedel, A. Kaplan, and S. Jackson, "High functional load inhibits phonological contrast loss: A corpus study," *Cognition*, vol. 128, no. 2, pp. 179–186, 2013.

[10] S. F. Stokes and D. Surendran, "Articulatory complexity, ambient frequency, and functional load as predictors of consonant development in children," *Journal of Speech, Language, and Hearing Research*, pp. 577–591, 2005.

[11] L. Van Severen, J. J. Gillis, I. Molemans, R. Van Den Berg, S. De Maeyer, and S. Gillis, "The relation between order of acquisition, segmental frequency and function: the case of word-initial consonants in dutch," *Journal of child language*, vol. 40, no. 4, pp. 703–740, 2013.

[12] M. J. Munro and T. M. Derwing, "The functional load principle in esl pronunciation instruction: An exploratory study," *System*, vol. 34, no. 4, pp. 520–531, 2006.

[13] J. Zhang, W. Li, Y. Hou, W. Cao, and Z. Xiong, "A study on functional loads of phonetic contrasts under context based on mutual information of Chinese text and phonemes," in *2010 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 2010, pp. 194–198.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[15] A. Warstadt, A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang, and S. R. Bowman, "Blimp: The benchmark of linguistic minimal pairs for english," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 377–392, 2020.

[16] P. Iverson, L. E. Bernstein, and E. T. 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.

[17] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[18] K. C. Hall, B. Allen, M. Fry, S. Mackie, and M. McAuliffe, "Phonological corpustools," in *14th Conference for Laboratory Phonology*, 2016.

[19] M. A. Redford and R. L. Diehl, "The relative perceptual distinctiveness of initial and final consonants in cvc syllables," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1555–1565, 1999.

[20] M. Davies, "The Corpus of Contemporary American English as the first reliable monitor corpus of English," *Literary and linguistic computing*, vol. 25, no. 4, pp. 447–464, 2010.

[21] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.