Contents

1 Introduction **4**

1.1 High information 8

1.2 Robust communication 10

1.3 Efficient linguistic communication 12

2 Data and Treatment **14**

2.0.1 Word probability 19

2.0.2 Lexical baselines and novel lexicon generation 20

2.1 Language data preparation 29

2.1.1 Conclusion 59

3 An Efficient Lexicon for Incremental Processing **61**

3.1 Balanced contrasts 66

3.1.1 Relation between type and token frequency 68

3.1.2 Results - Linear Relationship between Token and Type Frequency 80

3.1.3 Word-initial entropy 94

3.1.4 Results - Word-initial vs. Non-initial Segment Entropy 99
3.1.5 Discussion . . . 105
3.2 Entropy across word forms . . . 108
3.2.1 Preparation . . . 115
3.2.2 Results - Significance Tests . . . 124
3.2.3 Results - Novel Lexicons . . . 130
3.2.4 Discussion . . . 138
3.3 General Discussion . . . 144

4 Synergy between Redundancy and Efficiency 149
4.1 Background . . . 149
4.1.1 Noisy channel and redundancy . . . 152
4.1.2 Channel capacity and a smooth signal . . . 157
4.1.3 Mutual information and an efficient lexicon . . . 161
4.2 Methods and Results . . . 166
4.2.1 Preparation . . . 166
4.2.2 Results . . . 169
4.3 Discussion . . . 177
4.3.1 Post-hoc analysis . . . 178
4.3.2 Trade-offs in the Lexicon and Language . . . 181

5 Conclusion 185
5.1 Novel lexicons . . . 186
5.2 Different shapes of information in words . . . 190

5.3 Uniformity beyond word forms 192
5.4 Why ‘optimized’ but not ‘optimal’? 194

Chapter 1

Introduction

The famed linguist Ferdinand de Saussure claimed that phonological words and their meanings were arbitrarily linked (de Saussure, 1916). That is, there is no connection between the sounds that make up a word and the concepts it represents. Superficially, at least, this claim seems to hold true. For instance, there appears no a priori reason for /ˈdag/ to be related to the meaning of *dog* or /tɹa͡jˈsɛɹəˌtaps/ to be related to the meaning of *triceratops*. Yet, words are not all of language. Words are systematically combined to form larger hierarchically-structured complex meanings (e.g., Chomsky 1957), which then are often communicated from one person to another. Because language is at least partially used as a communication system, it stands to reason that biases for efficient communication at large may affect the individual structuring of human languages and their grammars, in particular the mapping between meaning of words and the sounds used to represent them.

Consider again the words *dog* and *triceratops*. Though the individual sounds of either word do not have an obvious connection with their meanings, *dog* is by far a much more frequently used word. It is also a much shorter word. If the pronunciations of the two words were switched, the longer *triceratops* would become more frequent, making the average English sentence would be longer, though not by much. Be that as it may, these slightly longer sentences would communicate the same information, though they would take more time and more effort to do so. Is this an accident? That is, is the relative shortness of the more commonly used word a fortunate result of a random process of arbitrary assignment between form and meaning or is it evidence of a relationship between the way a word is shaped and how it is used for communication?

In fact, this pattern extends beyond these two words and beyond English altogether and forms one of the most robust trends across languages, Zipf's 1935 *law of abbreviation* (see Fig. 1). The cross-linguistic law states that the most frequently used words have the shortest forms and vice versa, making the world's languages more efficient communication systems than if the pattern were absent. If the opposite were true and the most frequently used words were in fact the longest in a language, it would take significantly more time to communicate and the language as a whole would be less efficient. Here, I will use a definition of *efficiency* that is very similar to that of Gibson et al. (2019), which itself borrows heavily from the principles of *Information theory* (Shannon, 1948). A communication system is efficient

when it is able to successfully communicate as much information as possible, with as little time or effort as needed.

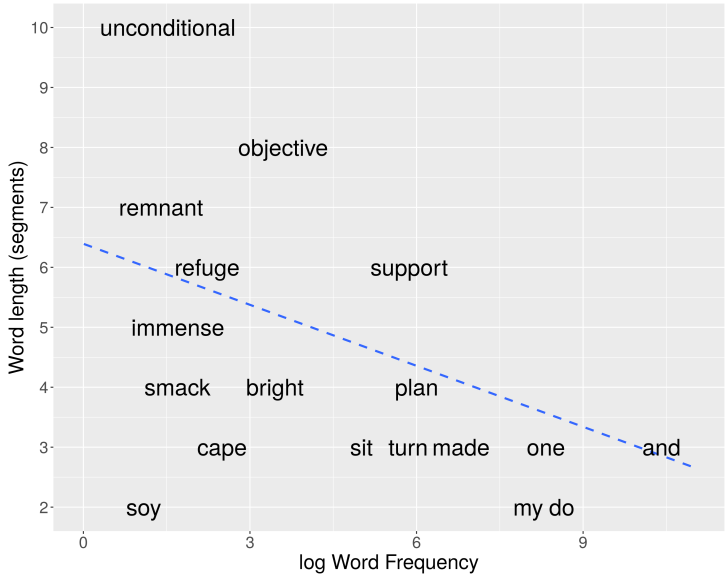

Figure 1.1: Example of Zipf's 1935 *law of abbreviation* for English. The x-axis show log-transformed word frequency and the y-axis shows word length. The visible words represent a random subset of English words to demonstrate specific examples while not cluttering the graph. Overall, there is a strong trend for more frequent words to be shorter.

The length of words is not the only aspect of language that appears structured for efficient communication, rather it is one of many robust aspects of language that make it an efficient communication system (for review see Gibson et al. 2019). For example, Piantadosi et al. (2009) found that stressed syllables often contain more information than unstressed syllables, and the greater articulatory effort of stress (e.g., Morton and Jassem 1965) helps ensure that the information will be understood accurately. Graff (2012) looked

at words across a language and found that there were fewer minimal pairs for contrasts that were perceptually similar (e.g., /p/ vs. /b/) than would be expected otherwise, suggesting that words are partially shaped to avoid confusion between words. Piantadosi et al. (2012) investigated homonyms, i.e., multiple words which share a pronunciation, and found that, despite local chances of confusion, the inclusion of homonyms actually makes language a more efficient system. Dautriche (2015); Dautriche et al. (2017); Meylan and Griffiths (2017); Mahowald et al. (2018) looked at the segments that make up words and found that words were preferentially constructed of easier-to-produce segments (for more, see Vitevitch et al. 2004; Tomaschek et al. 2018), making language less effortfull and thereby more efficient.

In this dissertation, I will seek to ask whether the contrastive segments, i.e., *phonemes*, themselves and their order within words may cause language to be a more efficient communication system for spoken language. All things being equal, a communication system that communicates *more* information than another over the same period of time is more efficient. Therefore, if language is structured for efficiency, the contrastive phonological segments of words should individually possess more information than would be expected otherwise in a system which was not structured for efficiency. Furthermore, words should be structured to ensure that the information within a word's segments is likely to be communicated successfully. If so, this will provide additional insight into the extent that efficiency plays a role in the structuring of human languages.

To talk about this in detail, it is important to flesh out a definition of information and to discuss properties that affect the success of communication.

1.1 High information

Consider a game of guessing the identity of a randomly drawn card from a standard deck of 52 playing cards. The aim of this game is to identify the mystery card with as few questions as possible, while being restricted to binary - 'yes' or 'no' - questions. This game is unusual for a game of chance, as there is an ideal strategy for choosing questions: always split the remaining cards into equal halves, e.g., as a first question ask whether the card is red. This strategy is slightly nonintuitive in that excludes any chance of guessing the correct card with the first question. For example, one could begin with a guess for specific card and potentially identify the card with a single question. However, for $\frac{51}{52}$ of the cards, the answer will be 'no', meaning that there will still be 51 possible candidates for the mystery card. On the other hand, a question which splits the deck in half is structured to reduce the set of possible cards as much as possible regardless of the actual answer.

Another way to think about this strategy is that it structures questions such that each provides maximal *information*. On average, each new question reveals more about the identity of the mystery card than any other of question. For sake of this work, I will formalize the definition of information as it

is done in *Information Theory* (Shannon, 1948), i.e., as the $-\log_2$ probability of an event. With this definition, less probable features of a card convey higher information and vice versa. For example, one out of every four cards is a diamond ($p(\diamondsuit) = \frac{1}{4}$) while one of every two cards is red ($p(\heartsuit \cup \diamondsuit) = \frac{1}{2}$), meaning that learning either property of a card will equal $-\log_2 \frac{1}{4} = 2$ or $-\log_2 \frac{1}{2} = 1$ bit of information respectively (assuming nothing else about the card is already known).

However, questions oftentimes have multiple possible answers and so the maximally informative question is one which results with the greatest information, on average. For example, asking if the card is a specific one, such as the two of clubs, has the potential to provide $-\log_2 \frac{1}{52} \approx 5.7$ bits of information, identifying the card with a single question. Yet, there is a $\frac{51}{52}$ chance that the card will not be the two of clubs and this answer will have only yielded $-\log_2 \frac{51}{52} \approx .028$ bits of information, having only excluded a single possible card. Across all possible answers to this question, the *entropy* of this question, or weighted average of all possible answers, is $(\frac{51}{52} * -\log_2 \frac{51}{52}) + (\frac{1}{52} * -\log_2 \frac{1}{52}) = .137$ bits. On the other hand, asking a question that that splits the remaining space in half will always equal $-\log_2 \frac{26}{52} = 1$ bit of information, meaning that the expected information of that question will be 1 bit. As this shows, entropy is greatest when all values are equiprobable (Shannon, 1948). By continuing to ask questions with this strategy, the game will always be won in roughly 6 questions questions, which is provably the fewest questions needed for the game to be reliably won, making the

strategy of equal splits highly efficient.

1.2 Robust communication

Now consider playing the same game, but in a noisy environment, an environment where a response of a simple 'yes' or 'no' may find its single syllable unsuccessful in transmitting the information contained in the answer. In this case, it is better to respond with longer forms of either answer, such as 'affirmative' or 'negative'. Though these words are longer, their greater length makes them less likely to be distorted due to noise. Because neither of these longer words reduce the set of potential cards more than the shorter 'yes' and 'no', the system as a whole is less efficient and thereby more redundant.

I use the term *redundancy* here to mean additional material or information beyond the minimum needed for successful communication without a chance of confusion; redundancy allows for information to be successfully communicated when there is a chance of confusion, though it may cause communication to take longer and more effort. As such, redundancy is key aspect of an efficient communication when there is a chance that part of the message may be distorted (Shannon, 1949).

To discuss the benefits of redundancy in detail, imagine a more abstract communication system where messages are sequences of only two symbols: 0's and 1's. This system communicates over a *noisy channel* where there is a .9 probability that a 0 or 1 is accurately transmitted and .1 probability

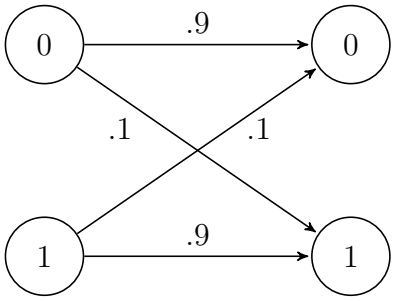

Figure 1.2: A noisy channel with two symbols where there is a probability of .1 that one symbol is received incorrectly. This is commonly known as a *binary symmetric channel* (for more, see MacKay 2003).

that one is confused with the other (see Fig. 1.2). This is a relatively inaccurate system as one of every ten symbols will be inaccurately understood, on average, potentially interfering with successfully recovering of the message as a whole. However, if each symbol in a message were redundantly encoded three times, e.g., 0 1 0 becoming 000 111 000, then at least two of the three copies would need to be affected by noise for the original symbol to be confused with the other, meaning that the probability of confusions would drop to .01[1]. Therefore, though it does make messages longer, an efficient communication system should include some redundancy if there is a chance of part of the message being effected by noise.

[1]In this scheme, a receiver would infer the original symbol as the majority class for the received three symbols. If only one of the three was affected by noise, the original symbol could be inferred because of the two remaining unaffected symbols. However, if two or three symbols changed ($p = .1 * .1 = .01$), then confusion would occur.

1.3 Efficient linguistic communication

How are these abstract examples relevant? Put simply, both have many parallels to how spoken language is used to communicate. Abstractly, spoken language begins with a *speaker* encoding a semantic *message* into into a linearized acoustic signal, which a *listener* must receive, process and decode to be able to correctly infer the original meaning of the message.

In the game, a guesser must identify a card by asking a series of questions, incrementally reducing the set of possible remaining cards until a single card remain that matches the sequences of answers and information gained. A crucial and early step in understand a message is for a listener to identify the *words*[2], processing the message incrementally with each subsequent segment within a word reducing the set of possible others words of the language (Dahan and Magnuson 2006; Weber and Scharenborg 2012 for review, further discussion will follow in later sections). For example, by hearing a word-initial [v], a listener gains sufficient information to rule out words that do not begin with /v-/. As more phonological material from the word is processed, the listener will continue to incorporate new information, until sufficient information is accumulated to allow for the correct identification of the intended word. Crucially, the information contained within a word

[2]There is evidence that the processing of morphologically complex words is affected by properties of the enitre word itself as well as properties of the individual morphemes (for more, see Schriefers et al. 1991; Balling and Baayen 2008, 2012). For the purposes of simplicity, I will assume that morphologically complex words are identified all together as a single unit.

is incrementally incorporated with other sources of information during processing, meaning that words that are less probable on average require more from the segments themselves to be accurately identified. The maximally efficient language then, should structure its *lexicon*, or set of meanings and their associated *word forms*, so that the segments of word forms concisely encode the sufficient information needed to identify words while remaining robust.

This dissertation will investigate if and how this lexical structuring occurs in the lexicons of a natural languages. Structurally, it will take the following shape.

In Chapter 2, I will describe the data sources, preparation and some general methodology for the corpus experiments found later.

In Chapter 3, I will demonstrate evidence that the individual segments that make up word forms posses more information than would be expected otherwise, i.e., inter-word contrasts are between probabilistically balanced sets.

In Chapter 4, I will show how the lexicon is structured to both contain high information, while still being sensitive to possible sources of interference such as noise and channel capacity, i.e., the overall entropy and amount of beneficial redundancy in the lexicon is greater than would be expected otherwise.

In Chapter 5, I will finish by discussing the results and possible avenues for future work.

Chapter 2

Data and Treatment

The thesis of this work is that the lexicon is structured to be an efficient code, tailored for the specifics of human language communication. Specifically, I predict that the lexicon should arrange contrastive phonological segments, i.e., phonemes, in word forms in an efficient matter. That is, segments should convey as much information as possible, given incremental word processing, while not adversely affecting the likelihood of successful communication.

To test this, I will investigate the data from of a variety of languages, arguing that each significantly demonstrates properties of an efficient communicative code as I have described. Together, I argue that the data suggest a statistical cross-linguistic trend (Dryer, 1998) for the lexicon to be partially structured to facilitate communication, given the specific pressures of human language communication.

As language is a constantly evolving system (e.g., Köhler 1987), the ideal

data source to test my predictions would cover a language's history over several generations, showing how word forms changed over time. However, finding a data source that accurately captures this for a wide set of languages would prove difficult, to say the least. If language does evolve under pressure for communicative efficiency, the lexicon at any given point of time should show some effect of these pressures. Because of this, a synchronic corpus is a useful proxy for the language's evolutionary history, and importantly, synchronic corpora are available for a large set of the world's languages. For these reasons, I will use word frequnecy information from synchronic corpora as a proxy for the relevant evolutionary history.

In this section, I will describe the language corpora that I will be using, how they were constructed and further processing that I did. In order to ensure that my results are not be skewed by the effects of a single language family or macro-area, I have endeavored to include a typologically and geographically diverse range of languages in my dataset (see Fig. 2), with as many languages outside of Europe and the Indo-European language family as possible, given available resources. I have also restricted the set of word forms, when possible, to word uninflected stems, i.e., *lemmas*, in order to minimize the effect of inflectional morphology[1].

[1]In calculating frequency, all inflected forms contributed to the overall frequency for the lemma.

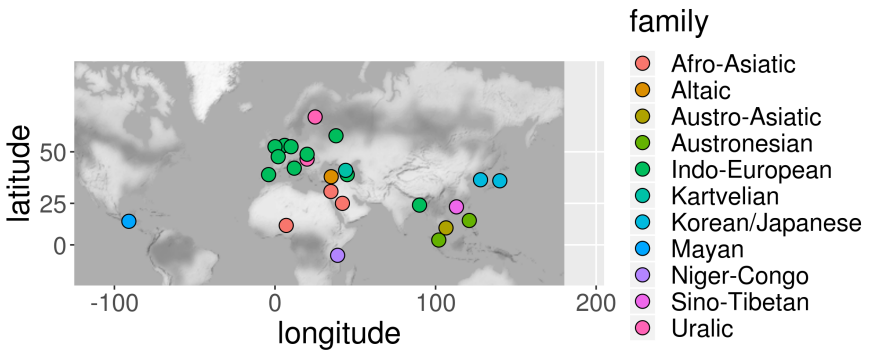

Figure 2.1: (Original) geographic location and family for languages in the dataset. For a list of languages and sources, see Tabs. 2.1 - 2.2.

Language	Source	Approx. Corpus size
Arabic	Canavan et al. (1997)	30,000 tokens
Armenian	Khurshudian and Daniel (2009)	100 million tokens
Bengali	Kilgarriff et al. (2014)	5 million tokens
Cantonese	Leung and Law (2001)	160,000 tokens
Dutch	Baayen et al. (1995)	40 million tokens
English	Davies (2008)	44 million tokens
Finnish	*"Finnish Text Collection"* (2005)	45 million tokens
French	New et al. (2001)	31 million tokens
Georgian	Gippert and Tandashvili (2012)	136 million tokens

Table 2.1: Table of sources for language corpora.

Language	Source	Approx. Corpus size
German	Baayen et al. (1995)	5 million tokens
Hausa	Kilgarriff et al. (2014)	2.4 million tokens
Hebrew	Linzen (2009)	80 million tokens
Hungarian	Váradi (2002)	112 million tokens
Italian	Lyding et al. (2014)	250 million tokens
Japanese	Canavan et al. (1997)	30,000 tokens
Kaqchikel	Tang and Bennett (2018)	700,000 tokens
Korean	Lee (2006)	1.3 million tokens
Malay	Lison and Tiedemann (2016)	7 million tokens
Russian	Kilgarriff et al. (2004)	50 million tokens
Slovak	Krajčovič (1988)	450,000 tokens
Spanish	Mendonca et al. (2009)	150 million tokens
Swahili	Hurskainen (2004)	25 million tokens
Tagalog	Goldhahn et al. (2012)	7 million tokens
Turkish	Sak et al. (2008)	130 million tokens
Vietnamese	Kilgarriff et al. (2014)	90 million tokens

Table 2.2: Table of sources for language corpora.

For all languages, data was collected from a variety of different genres, e.g., news transcripts, movie subtitles, to better represent the language as a whole[2]. If I was not able to find a curated corpus for a language, I aggregated

[2]As an amusing anecdote, the curator of my Russian corpus mentions that he was

texts for that language to create a new corpus and then used other resources to phonemically transcribe the words of the new corpus. I will detail each language's corpus and further preparations I performed in detail in Section 2.1.

Whenever possible, I chose a corpus of spoken dialogs as resource for a language. Though reading is also incremental (e.g., Ehrlich and Rayner 1981; Rayner and Well 1996; Levy et al. 2009; Smith and Levy 2013), I am particularly focused on how phonological material is arranged to create an efficient communicative code during speaking. For this reason, resources that more accurately reflected what listeners are likely to hear when speaking the language were preferable.

For lower-resourced languages where spoken corpora were not available, e.g., Kaqchikel, I relied solely on written sources. Though the words and frequencies in written corpora are not perfectly representative of what a listener is likely to hear, I felt it was better to include a wider range of languages, rather than a small set of high resource languages which are more likely to belong to the Indo-European language family.

I followed the suggestions of Jurafsky and Martin (2014) and limited my data to words above a strict frequency threshold (1 per million) and to a maximum of set of word forms (10,000), which together minimize the likelihood

inspired to create a Russian corpus as the largest Russian corpus at that time was composed solely from Soviet newspapers that lent some political ideology to the word counts. For example, *comrade* was as frequent as a function word, showing an obvious skewing of the corpus at large.

of misspellings, out-of-language borrowed words or names, etc., skewing the analysis in undesirable and unpredictable ways. All corpus preparation was done with Python (Van Rossum and Drake, 2011), using the NLTK (Loper and Bird, 2002) and EPITRAN (Mortensen et al., 2018) libraries. Analyses and were performed in R (R Development Core Team, 2008) and visualizations were generated using the GGPLOT2 library (Wickham, 2009). Computation was carried out via servers at the University of Arizona and all code is available at: `http://github.com/AdamKing11`.

2.0.1 Word probability

A word *type* is a distinct word form in a particular language's corpus. A word *token* is an instance of a word type and the number of tokens for that word type determines the word's *frequency*. Because the corpora I used were of very different sizes (e.g., Egyptian Arabic had roughly 30,000 tokens, English had roughly 40 million), I represented frequency as tokens per million, making frequency values more comparable between languages.

$$p(w) = \frac{freq(w)}{\sum_{w'} freq(w')} \tag{2.1}$$

I chose to use a context-free measure of word probability, which is a relative measure of the word's frequency. Explicitly, this was equal to the frequency of the target word in the corpus, w, divided by the sum of all words in the corpus including the target word, w' (Eq. 2.1). If a particular

lemma appeared with multiple forms in the corpus, e.g., *sing*, *singing*, *sang*, I determined the frequency for the lemma form as the sum of all inflected forms.

Though other means of operationalizing word probability have been shown to be of interest, e.g., average n-gram probability (Piantadosi et al., 2011), I decided upon this relatively simple method as it allowed for the incorporation of data from each language in my dataset, many of which do not have a corpus of adequate size or level of pre-processing for more sophisticated estimations of word probability. For example, the corpora of well resourced languages such as English and Russian are sufficiently large to create trigram language models, though this would be less feasible for less resourced languages such as Kaqchikel or Hausa. Fortunately, context-free probability is highly correlated with more sophisticated measures of word probability (Cohen Priva and Jaeger, 2018), meaning that the chance that any result would be an artifact of the representation of word probability is low. Be that as it may, further work may be merited to piece apart different results for different methods of determining word probability.

2.0.2 Lexical baselines and novel lexicon generation

Within this work, I am arguing that the lexicon is structured to facilitate communication across a noisy channel, given incremental word processing. As support for this claim, I will present data to show that lexicons of natural languages *significantly* demonstrate various aspects of an efficient commu-

nicative code. At this point, I would like to take a moment to define what exactly I mean by 'significantly'.

In standard statistics, the term 'significant' indicates that one variable accounts for a sufficient amount of observed variance in another, such that there is an acceptably low probability (standard practice is 5%) that the data would be such as they are under the 'null-hypothesis' (Baayen, 2008). In many cases, the null-hypothesis is assumed to mean that two variables have no correlation at all. Yet, this assumption is not always satisfactory for an investigation into a system as complex as human language.

For example, Miller et al. (1958) argued that Zipf's law of abbreviation could arise from so-called "*monkeys on typewriters*" or from any string of a random of symbols. Assuming that each Latin letter and the whitespace character are all equally probable - $p = \frac{1}{27}$ - the probability of any particular string of letters, i.e., word form, that is of length two is $\frac{1}{27^2}$, the probability of any string of length three is $\frac{1}{27^3}$, length four is $\frac{1}{27^4}$, and so on. This yields a clear negative relationship between word length and probability since longer words will be less probable by necessity. If the null-hypothesis was assumed to be no correlation between word probability and length, this random string of letters would also 'significantly' demonstrate Zipf's law, making the correlation in natural language less interesting. However, the fact that a non-zero correlation might exist a priori does not make testing relationships impossible. Ferrer-i Cancho and Elvevåg (2010) used the *typewriting monkey* as a baseline and showed that natural language held a tighter correlation between

word probability and length than would be expected otherwise and better fit Zipf's original prediction.

Following from this, the 'null-hypothesis' should often be adjusted in order to be able to claim that the lexicon demonstrates 'significant' correlations between the variables in question (for more discussion, see Moscoso del Prado Martin 2013; Ferrer-i Cancho 2016). As such, I will employ two main families of tests in this work. Firstly, I will rely on standard statistical tests (primarily, Pearson's correlations and mixed-effects models) to demonstrate whether a correlation exists between a word's probability and its form or not. For many tests, I will then move on to a more conservative set of tests where I compare the lexicon against a more realistic baseline for the effect. This baseline will represent an expected value for a code that shares many similarities with a lexicon, while not having been subject to the same evolutionary pressures for communication, i.e., a more applicable null-hypothesis. The most straightforward means to create a baseline is to generate a large set of randomly generated variations on a real-world lexicon, as was done by Graff (2012) and Dautriche (2015) to show other optimized properties of several languages.

For example, Dautriche (2015) tested the number of minimal pairs in the lexicon by comparing against sets of nonce word forms that were equally-sized to the original lexicon. These nonce word forms represented *possible* words of the language that may not have been found in the target language, e.g., *blick* /blɪk/ in English, which, as a whole, represented possible alternative forms

of the lexicon for that language. By using a baseline of alternative forms of the lexicon, she was able to estimate the expected number of minimal pairs for each language and then show that the real-world lexicons possessed more pairs on average, which she argued to be evidence of lexical structuring for lower effort productions.

With this methodology, deciding how to generate variations of the lexicon is an important choice. They must differ from the original, real-world lexicon while remaining a possible lexicon of a natural language. Furthermore, they must not differ in a way that could affect the tested hypothesis. For instance, if the task were to evaluate the entropy of phonemic contrasts, it would be necessary to compare the original lexicon to alternatives with the same number of phonemic contrast. All things being equal, contrasts between more possible values will be higher information on average than contrasts between fewer, meaning that the size of a lexicon's phonemic inventory plays a large role in determining entropy. Therefore, to show that a language has higher entropy than would be expected otherwise, the baseline should be constructed from alternative lexicons with identically sized phonemic inventories.

The question now turns to the exact method of generating word forms for the alternative lexicons. All word forms in a viable alternative should share certain similarities with the forms of the original, while differing in the exact association between word probability and form. In addition, it should be possible to generate a suitably large set of variations for each language; it could be the case that any difference between the original lexicon and a

single generated alternative is due to chance. From a suitably large set of generated alternative lexicons, it is possible to show whether the identical value for the real-world lexicon falls beyond that distribution, which itself is a convincing secondary definition for 'significant'.

Within this dissertation, I will utilize two methods for creating alternative forms of the lexicon. The first is to generate novel word forms from a phonotactic model extracted from the target language. The second is to redistribute the probability values of word forms among words of the same length. Both represent different means to vary the lexicon and as such are useful for testing different aspects. A lexicon generated by phonotactic models varies both the relationship between word probability and form, as well the size and make up of the various cohorts within the lexicon. On the other hand, a lexicon with an identical set of forms but where the word probabilities are shuffled creates a novel mapping between form and probability while maintaining the overall cohort structure. I will discuss both in detail below.

Phonotactic models

One means to generate a novel lexicon is by generating word forms via a phonotactic model. This paradigm has been used in similar investigations into the lexicon (e.g., Dautriche et al. 2017; Mahowald et al. 2018) and is a useful but still feasible means to create a comparable baseline for the lexicon. A phonotactic model is a probability distribution that models the probability of a particular segment given the context of the previous segments in a

word form, e.g., the probability that one segment follows another (for more discussion, see Dautriche 2015).

I will limit the context to the two previous segments. In other words, the probability of a segment at position i of a word, s_i, is only affected by the number of words that also share the two previous segments, s_{i-2}, s_{i-1} (Eq. 2.2). I opted for using a context of two segments because it maintains similar syllable structure, i.e., allowed consonant clusters and sonority sequences in the language, while not being too large a context to make novel word forms too similar to those of the original lexicon.

To construct the model, I counted segment sequences of length three, s_{i-2}, s_{i-1}, s_i and two, s_{i-2}, s_{i-1}, and then determined the probability of a segment as its count given its context and divided by count of context altogether. I included distinct word boundary symbols at the beginning and end of each word form.

$$p(s_i|s_{i-2}, s_{i-1}) = \frac{count(s_{i-2}, s_{i-1}, s_i)}{count(s_{i-2}, s_{i-1})} \tag{2.2}$$

I will be using a type-based model, ignoring word frequency when determining the probability of segment sequences. That is, probability values are only determined by the number of word types that a particular sequence is found in, irrespective of the frequency of those words. I opted for a type-based model in order to create a more diverse set of word forms, while still limiting the forms to those that are phonotactically licit. Put simply, a token-

based phonotactic model is likely to generate a less diverse set of word forms. In detail, this is because the Zipfian nature of word frequencies causes that probability of the majority of segment sequences in the original lexicon to fall far below $\frac{1}{N}$, where N is the number of word forms in the lexicon. This in turn means that, when generating N novel forms, a vast majority of *possible* segment sequences will not be created and the resulting novel word forms will be very similar.

To generate a word form, I began by generating a word boundary segment and then sampling from the phonotactic model (Eq. 2.2) until another a word boundary segment was generated. To create a variant form of a lexicon, I began with the word forms and their associated probabilities from the real-world lexicon and replaced each form with a generated form of identical length, which I did to maintain the relationship between word probability and length. For example, *desk* /dɛsk/ could be replaced with /blɪk/ but not by /blɪkə/. By maintaining this relationship, I was able to ensure that the average 'sentence' for any novel lexicon would be constructed of an identical number of segments as the original lexicon. Therefore, if information was efficiently encoded into the word forms of the real-world lexicon, it would not be due to the possibility that a novel lexicon included longer word forms.

Across languages, more than 50% of word forms in novel lexicons that were five or fewer segments long were also found in the original lexicon, on average. As might be expected, for longer words, there was much less overlap, causing the overall number of shared word forms in the original and novel

lexicons to be roughly 10%, on average. Nevertheless, though this method did generate word forms that were not found in the original lexicon, word generation was relatively conservative.

Probability-shuffling

The second method that I used for creating a novel lexicon maintained the same set of word forms, but shuffled the probability values for those forms of the same length. For example, the 3-segment words *and* /ænd/ and *lute* /lut/ could switch probability values but *and* and *story* /stori/ could not. This creates a novel assignment of probability to form, while keeping the relationship between word probability and length constant. As an example, consider Fig. 2.0.2. This graph shows the total length of the original English corpus in segments ($\sum wordlength(word) * freq(word)$), compared to a distribution of probability-shuffled variants where length has not been held constant, i.e., *the* and *divestment* could switch frequencies. As it is clearly made obvious in Fig. 2.0.2, shuffling probability values without holding length constant would create a novel lexicon that is less efficient a priori; the average sentence in one of these shuffles will be much longer, though it would convey the same information altogether.

Furthermore, probability-shuffling is a much more conservative strategy to create a novel lexicon compared to phonotactic models, as each word form is one that was found in the original language. Though phonotactic models generate *possible* words of a language, they do so ignoring some higher level

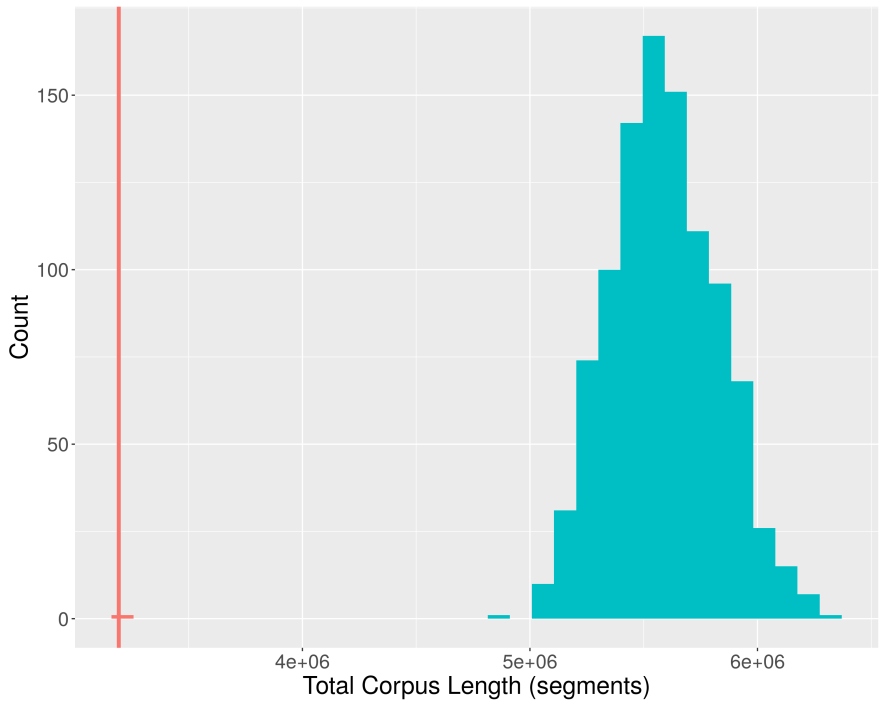

Figure 2.2: Comparison of the total length of the real-world English corpus (red) against a distribution of 1,000 variants of English (blue) where word probability values have been shuffled but length has not been held constant. Because the real-world lexicon displays a relationship between word probability and length, i.e., Zipf's law of abbreviation, the real-world lexicon requires far fewer segments to encode the information of the corpus.

constraints on word formation. For example, words in Semitic languages such as Hebrew and Arabic are based on a root-and-pattern system, meaning that a vast majority of words have exactly three consonants. A trigram phonotactic model may approximate this, though it is possible that it will generate words with more or less than three consonants. On the other hand, a probability-shuffled lexicon is guaranteed to always consist solely of word forms that satisfy all constraints on word form, simply because all word forms

are found in the original lexicon.

On a more pragmatic note, because the set of word forms is the same, certain properties of the lexicon remain the same automatically, minimizing the time and computational resources it takes to compare against the real-world lexicon. For example, the tests in Chapt. 4 require comparison lexicons that not only maintain the relationship between word probability and length, but also maintain the relationship between word probability and neighborhood size (e.g., Luce and Pisoni 1998, more discussion will follow in the relevant chapter). Creating a novel lexicon via phonotactic models that maintains both would be very computationally intensive; creating a novel lexicon via probability-shuffling does by default.

2.1 Language data preparation

For each language, I aggregated the relevant resource to create lists of word forms and their frequency. For all languages, I ensured that there was a one-to-one correspondence between characters used in the word form's transcription and contrastive segments, i.e., phonemes. That is, if multiple characters in the languages orthography were used to represent a single segment, I replaced those characters with a single symbol to represent the sound. This made sure that word length calculations and segment positions within word forms were easily compared across languages. The scheme that I used to represent phonological segments was not identical to the IPA or the language's

orthography (if it uses the Latin alphabet), nor was it always uniform across languages. For example, I use 'c' to represent [t͡ʃ] in English, while I use a 'T' to represent an ejective consonant in Kaqchikel but an aspirated consonant in Armenian.

Though a totally uniform transcription system would perhaps be ideal, I found it more important to use a system where the chosen character was decipherable by one who is familiar with the language and for all characters to be limited to standard ASCII characters[3]. Regardless, what is most important in this case is that each contrastive phonological segment is represented as a single character, so that tests on word length and segment position are comparable across all languages in the dataset.

It is important to note that I treated all different segments as discrete contrasts, without any appeal to similarities between segments due to articulatory or perceptual features. For example, the *distance* between 't' and 'd' was identical to that of 't' and 'e'. Though there is a large body of literature to strongly suggest that this is a gross oversimplification (e.g., Smits et al. 2003; Flemming 2004; Cutler et al. 2004; Mielke 2012), it is a simplification that reduces the assumptions I must make as to the exact *distance* between segments and makes testing multiple language feasible[4]. As with my

[3]I restricted the set of possible characters to ASCII rather than Unicode-8 characters because it was more memory efficient and more efficient to run.

[4]For example, the feature [+/− STRIDENT] of Chomsky and Halle (1968) seems to exist solely for the purpose of explaining certain English phonological processes. Whether 'strident' should be included as a feature for all languages is a separate branch of research itself.

selection of operationalizing method for word probability, I expect that more sophisticated representations for contrasts will yield interesting findings and should be explored in the future.

For certain phonological segments, there were some choices in transcription that needed to be made. I chose to represent geminate consonants, long vowels and diphthongs as sequences of segments, rather than with their own unique segments. For example, the word /kːa/ would be considered three segments long with segments 1 and 2 both being a token of 'k'. I did this for two primary reasons. Firstly, doing so models the similarity between long and short variants of vowels and consonants that representing either with a distinct symbol would not. Secondly, representing longer variants of contrasts as multiple symbols captures the fact that these sounds are in fact longer in duration. In the example of /kːa/ above, I felt it better to count this as a length 3 compared to 2, which would be the case if I represented /kː/ and /k/ with two different single symbols.

For languages with contrastive tone, I opted to note tones as digits that directly followed the vowel, i.e., tone 1, tone 2, etc., much like diphthongs. Though this may be imperfect considering that tone is not perceived only after the syllabic nucleus, I chose to do this to limit the number of contrasts in tone languages. For example, Cantonese possess 11 contrastive monophthongs, 11 diphthongs and 6 distinct tone patterns, meaning that for each of these to be represented as a distinct symbol, it would require $(11 + 11) * 6 = 132$ symbols. Logistical problems aside, this posed problems

for the analysis itself, in that the potential entropy (given how I determine segment information) of tone language nuclei was significantly higher than that of non-tone languages. This meant that by the simple property of having tones, tone languages would be biased towards having higher information segments on average than non-tone languages. Though the differences in tone and non-tone languages may in fact be an interesting avenue for future work, I felt that it would be more conservative to avoid this potential confound.

When determining word length, I removed tone segments, only counting the consonants and vowels.

Egyptian Arabic (Semitic)

I collected data for Arabic from the Callhome corpus of Egyptian Arabic (Canavan et al., 1997). I chose this corpus because it contained phonetic transcriptions for word forms (including vowels) and frequency information and was collected from spoken sources (phone calls). A crucial thing to note is that due to Arabic's root-and-pattern morphology, many word forms, especially verbs, do include some amount of inflection. Fortunately, many prefixes and suffixes were explicitly marked in the corpus (separated from stems with a '-') and I was able to remove the affixes from the analysis. That is, words with affixes would be counted for determining the frequency of the word's stem, but not for the word itself.

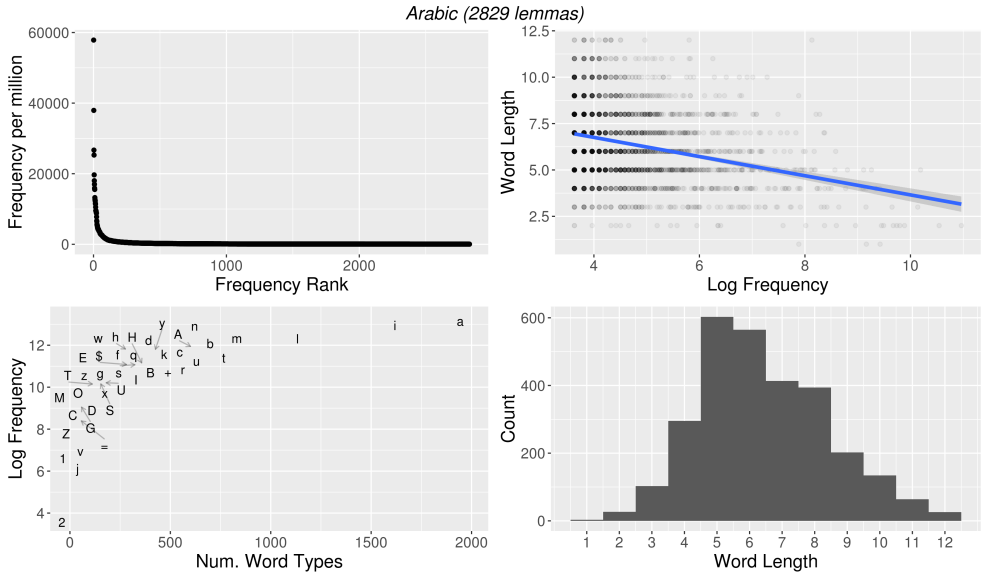

Figure 2.3: Graphs showing details for Egyptian Arabic word forms. The top-left graph shows a word's frequency rank (x-axis) and its actual frequency (y-axis). The top-right shows the relationship between log word frequency (x-axis) and length (y-axis). The bottom-left shows the number of word types that contain a particular segment, in the transcription scheme used for the language (x-axis), plotted against the total frequency of those words (y-axis). The bottom-right shows a histogram of word lengths.

Armenian (Indo-European, Armenian)

I collected data for Armenian from the Eastern Armenian National Corpus (Khurshudian and Daniel, 2009). This corpus is the largest collection of written works of Eastern Armenian, comprising over 100 million tokens, with each token in the corpus being morphologically parsed[5]. Though the corpus does include historical works, I only collected word frequency information from the "Modern" subsection of the corpus.

[5]The morphological information provided by the corpus was done programmatically (more information available at `http://www.eanc.net`).

To collect word frequency information, I scraped the online corpus using a combination of UNIX bash scripts. From this, I counted the lemma form of each token in the corpus. Being a morphologically complex language, for some word forms, there were multiple possible parses, though for the majority, there was only a single possible parse. For example, [avɛli] could be 'more' or the genitive form of the noun 'broom'. When there were multiple suggested lemmas, I counted the first proposed lemma only, which was the lemma with the greater frequency altogether.

I then transliterated the Armenian script into ASCII characters using regular expressions. In addition, I included additional regular expressions to account for spelling differences between the Eastern and Western dialect of Armenian and to account for common Armenian phonological processes, e.g., final devoicing, insertion of glides before word initial mid-vowels, etc. (Vaux, 1998).

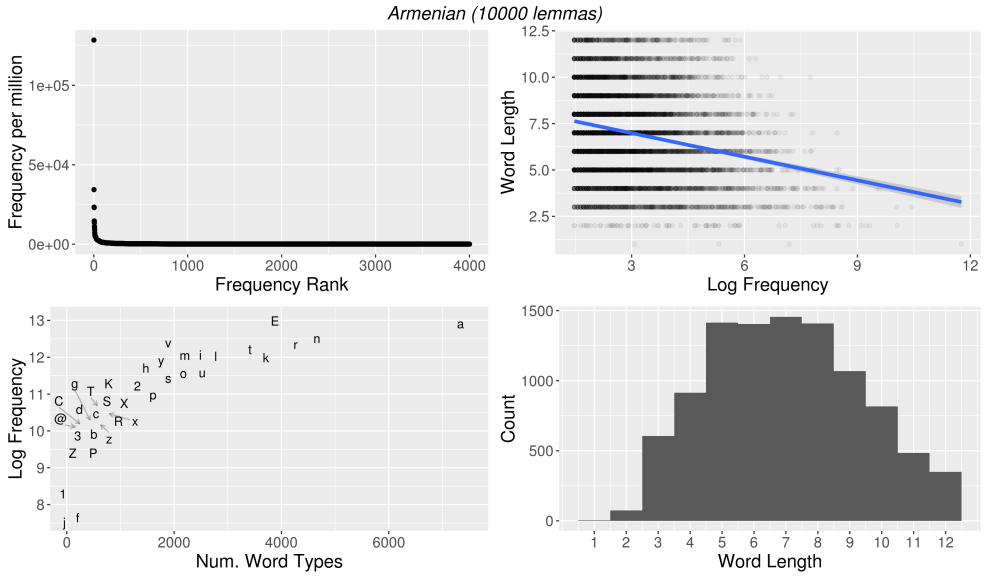

Figure 2.4: Details for Armenian word forms.

Bengali (Indo-European, Indic)

I collected Bengali data from a Bengali-script collection of web text, Wikipedia articles and news articles (Kilgarriff et al., 2014). Word forms were presented with morphological annotation, which allowed counting lemmas instead of raw tokens. Because the data was presented in its natural Bengali script, I utilized the epiTran package to transliterate the Bengali script into IPA transcriptions, which I further refined by the condensing IPA transcriptions that did not have a 1:1 correspondence between symbol and contrastive phoneme.

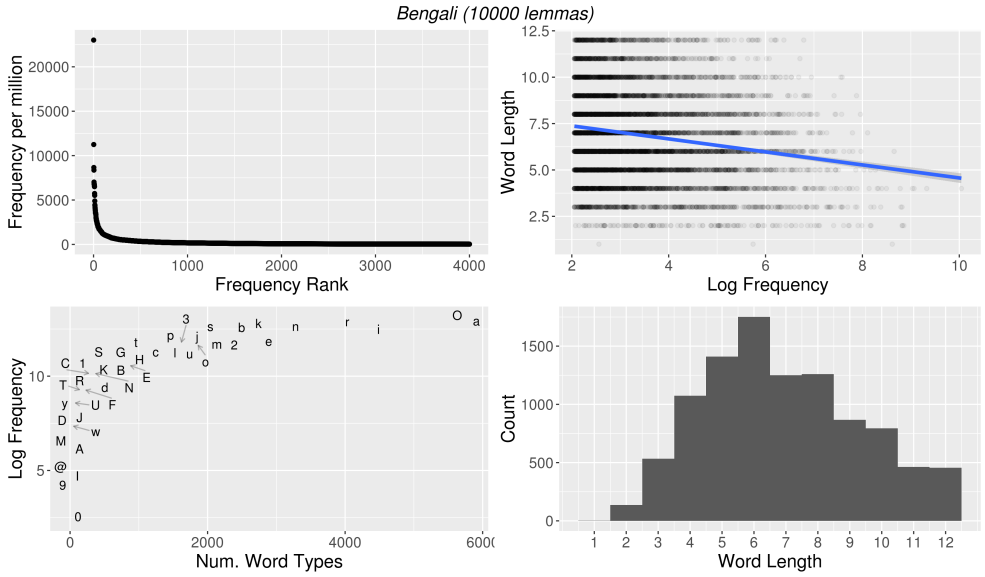

Figure 2.5: Details for Bengali word forms.

Cantonese (Sino-Tibetan)

For Cantonese, I used the Hong Kong Cantonese Corpus, which was collected from transcribed conversations of spontaneous speech and radio programs (Leung and Law, 2001). Each token was annotated with the Linguistic Society of Hong Kong's romanization scheme. For phonemes that were represented with multiple characters, e.g., 'ch', I used regular expressions to reduce these to a single character. To ensure that word lengths were more comparable with non-tone languages, I removed the tone characters when determining word length.

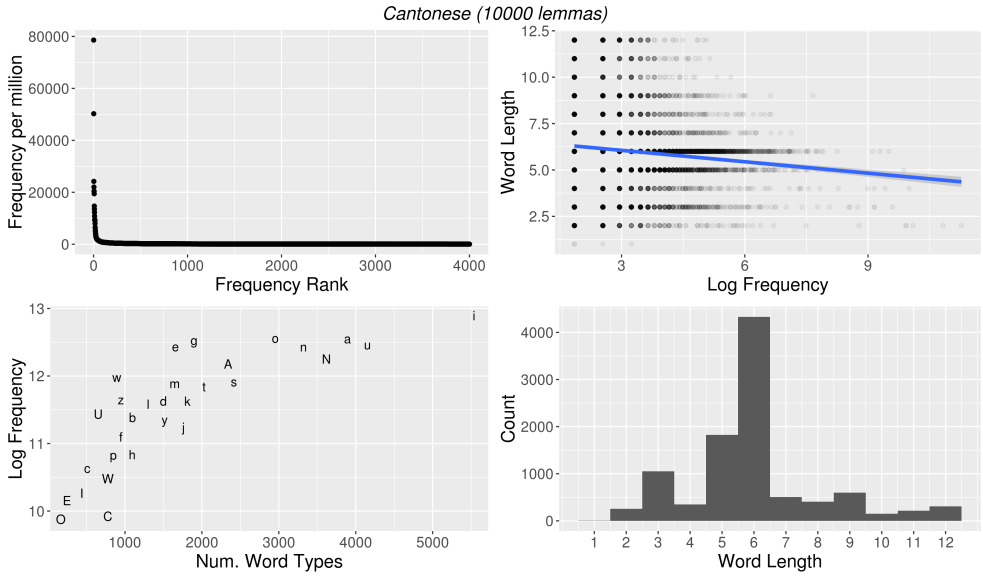

Figure 2.6: Details for Cantonese word forms.

Dutch (Indo-European, Germanic)

I collected Dutch data from the CELEX corpus (Baayen et al., 1995), which included phonetic transcriptions, morphological parsing and frequency information. Using the morphological information, I only collected word forms that were marked as lemma forms, i.e., no word forms that explicitly included inflections. I did not alter transliterations given in CELEX, except for when multiple characters were used to represent a single contrastive sound, e.g., 'dZ' [d͡ʒ] to 'J'.

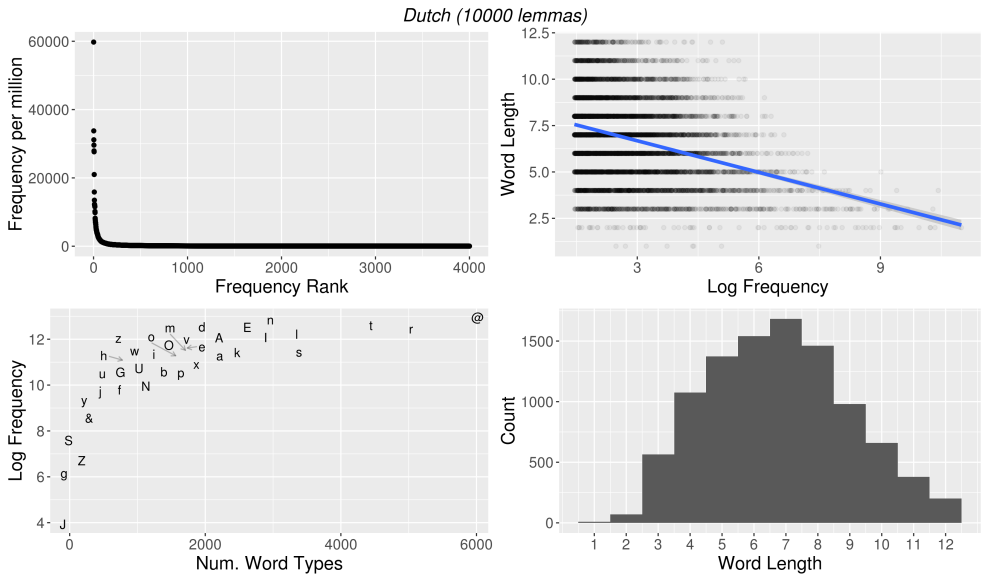

Figure 2.7: Details for Dutch word forms.

English (Indo-European, Germanic)

My English source was comprised of multiple corpora. I used the *Carnegie Mellon Pronouncing Dictionary* (Weide, 2005) for phonetic transcriptions of American English words and I used the *Corpus of Contemporary American-English* (Davies, 2008), a collection of subtitles and television transcripts from 1990-2012, for frequency information. I then used CELEX (Baayen et al., 1995) to exclude inflected forms.

I chose to not use the English subcorpus of CELEX as it is based off of British English and I was better able to verify the correctness of transcriptions for American English.

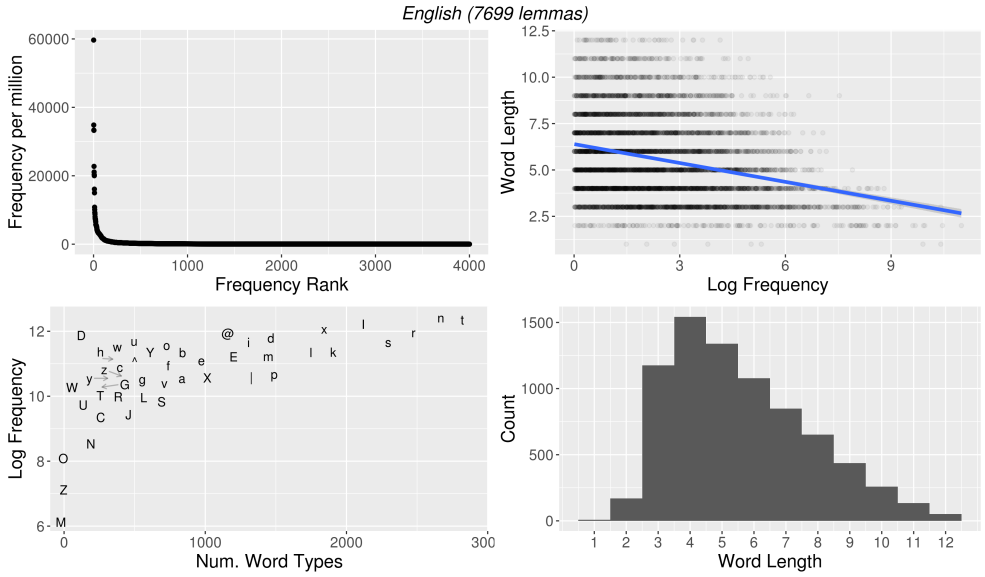

Figure 2.8: Details for English word forms.

Finnish (Uralic)

My source for Finnish was collected from *"Finnish Text Collection"* (2005), with frequency information collected from Finnish newspapers over a several year period. This source was constructed by its original authors to only include word stems. Due to the transparency of the Finnish orthographic system, I left the word forms in their original forms from the corpus.

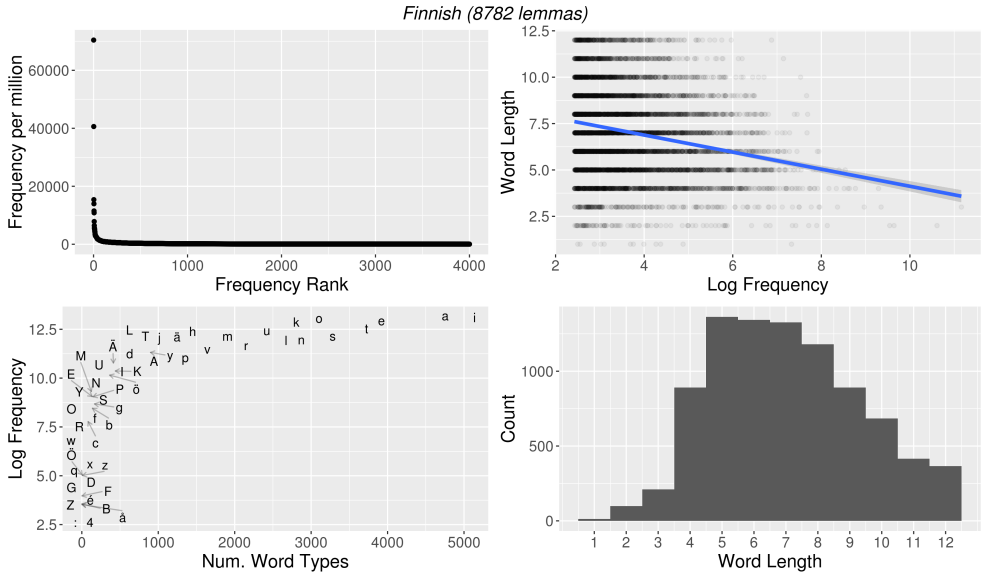

Figure 2.9: Details for Finnish word forms.

French (Indo-European, Romance)

My source for French was collected from *Lexique* (New et al., 2001), with possessed phonetic transcriptions of lemma forms and frequency information from a collection of written French sources. Due to the quality of this source, I did not make any further modifications to the transcriptions.

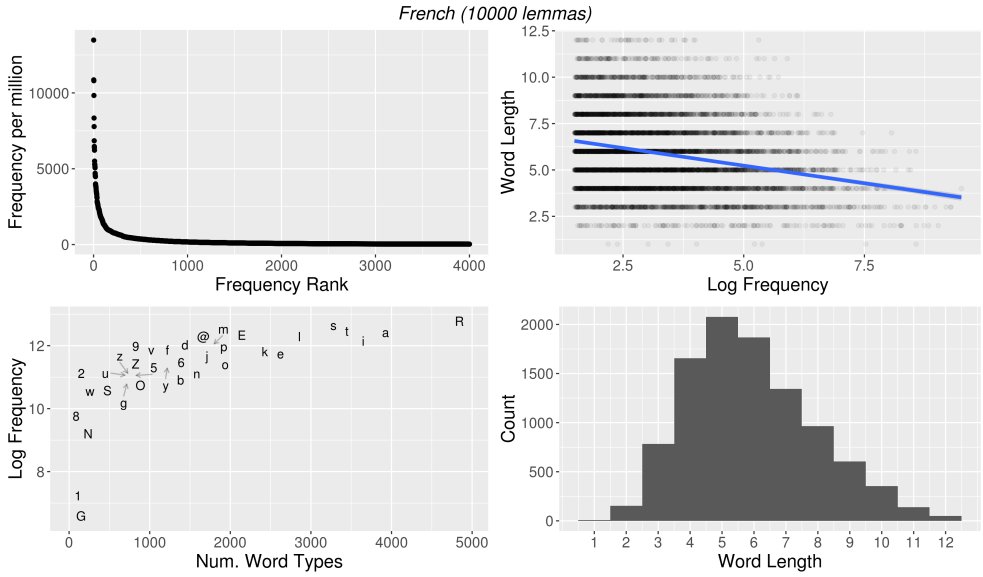

Figure 2.10: Details for French word forms.

Georgian (Kartvelian)

To construct the Georgian data source, I used the Georgian National Corpus (Gippert and Tandashvili, 2012), which is a large collection of Georgian texts. To create the source, I collected frequency information for the 20,000 most frequent words listed as "simplified lemmas" from the "Modern" subcorpus (not including most derivational affixes). Though I restricted my collection of word forms to modern Georgian, I removed all forms that possessed characters that were made obsolete by the 19th century Georgian spelling reform (Aronson, 1990), as these forms likely came from a markedly non-modern document.

Following this, I was able to use regular expressions to transliterate the Georgian script to ASCII since the Georgian orthography is relatively trans-

parent.

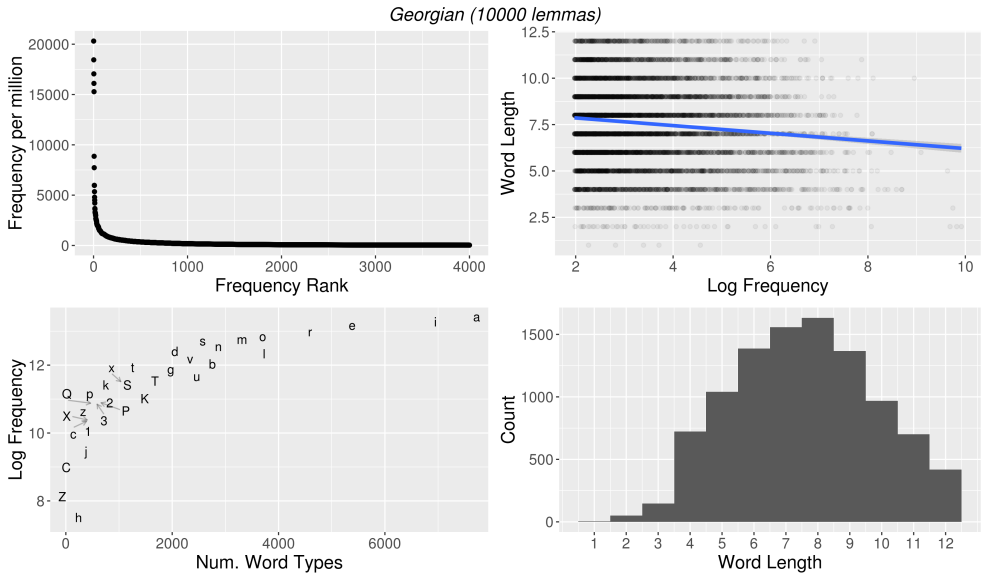

Figure 2.11: Details for Georgian word forms.

German (Indo-European, Germanic)

I collected German data from CELEX (Baayen et al., 1995), only collecting form without inflectional morphology. As with Dutch, I did not alter the transliteration, except when two characters were used to represent a single contrastive sound.

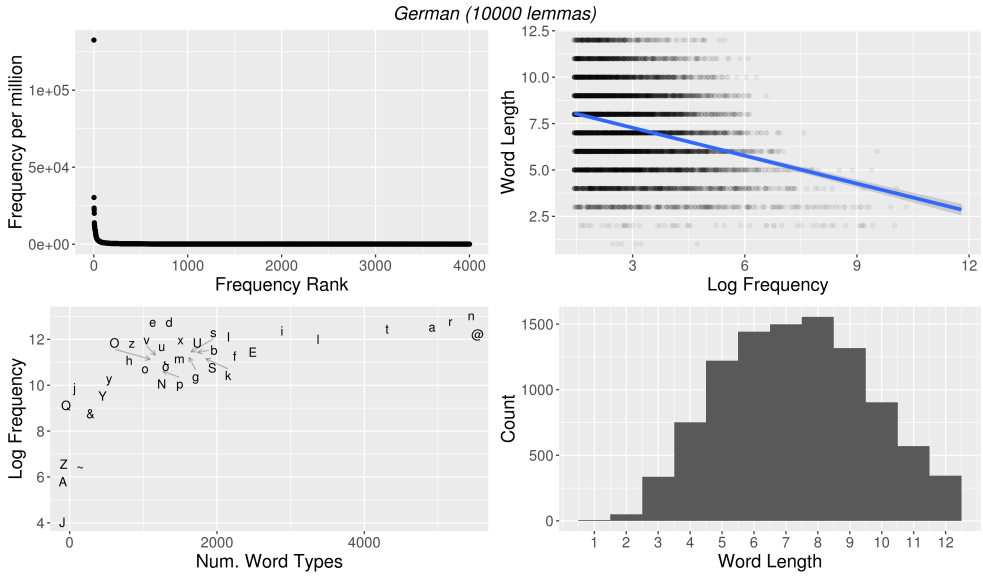

Figure 2.12: Details for German word forms.

Hausa (Chadic)

I collected Hausa data from the HAWAC corpus of web texts and Wikipedia articles (Kilgarriff et al., 2014). The data was presented in the Boko script, which is a transparent orthography based on the Latin alphabet. However, Hausa possesses contrastive tones (high vs. low), but this distinction is not always represented in the Boko script. To this end, some documents within the corpus explicitly denoted tone, while others did not. To deal with this, I opted to condense all vowels of the same quality to a single character, e.g., {'o', 'ó'} → 'o'. Though this does lose tone information, I felt it was more necessary to have a uniform representation throughout the data.

Because of the otherwise transparency of the Hausa orthography, I kept word forms as is, except to condense multi-character sequences for a single

phonemic contrast.

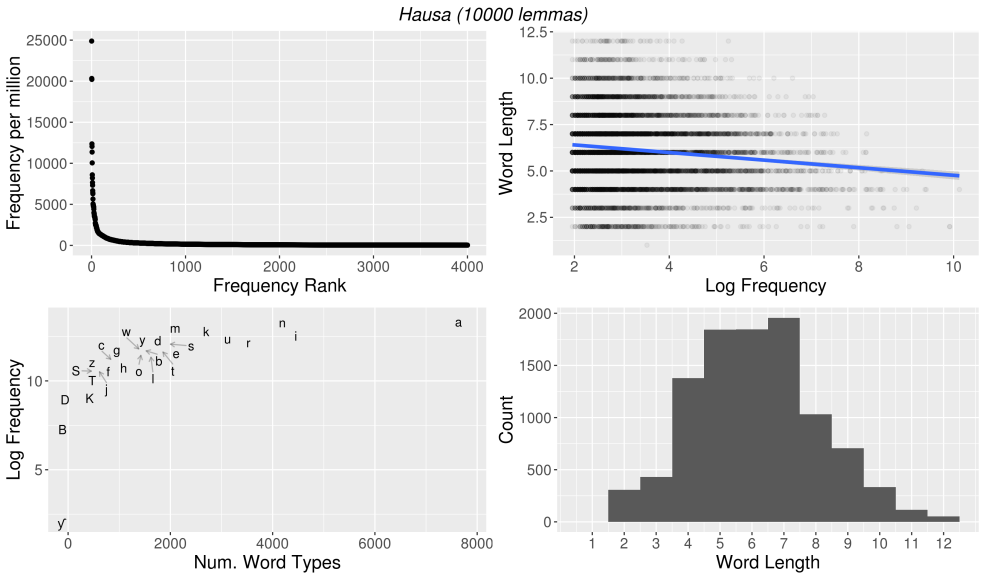

Figure 2.13: Details for Hausa word forms.

Hebrew (Semitic)

For Hebrew, I employed two sources. The first was a written corpus of the language that listed frequency with orthographic forms in the Hebrew alphabet (Linzen, 2009). However, because of the orthographic system of Hebrew, these forms all lacked vowels. To account for this, I used a second source, a list of the 10,000 most frequent Hebrew word forms with their transliteration into a sequence of ASCII characters, which included vowels[6]. I found the intersection of the Hebrew words in either source, finding the precise frequency count for the written forms.

[6]Available at `https://www.teachmehebrew.com/hebrew-frequency-list.html`.

Due to the fact that the Hebrew script does not explicitly mark vowels, some transliterated forms would share a written form. When this was the case, I chose to divide word frequency evenly among the transliterated forms. Though, of course, this may result in certain forms being represented at different frequencies than they occur in actual Hebrew, it allowed the analysis to utilize a large number of forms without arbitrarily excluding certain forms.

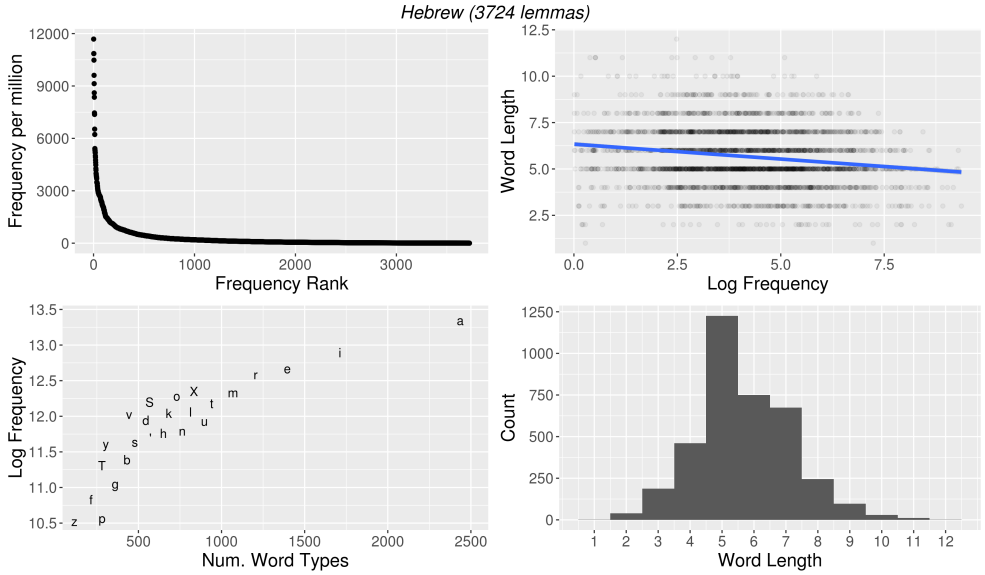

Figure 2.14: Details for Hebrew word forms.

Hungarian (Uralic)

I used the Hungarian National Corpus (Váradi, 2002) to collect data for the Hungarian language. This corpus is a collection of written texts for the language over several years and for several genres or written work. In addition, each word form is morphologically parsed, showing a morpheme-

by-morpheme breakdown for individual words and separately noting where compounding occurs. Because Hungarian is a very morphologically rich language (and the source I used was well annotated), I only included words that as a whole met a frequency threshold (10 occurrences in the corpus) and whose composing morphemes all met another frequency threshold (10 occurrences in the corpus). I did so to exclude word forms that were overly infrequent as well as words that included morphemes that themselves were overly infrequent. This was in addition to the frequency requirement I had for all languages (1 per million). Though some word forms in the dataset did include derivational morphology, I removed all inflectional morphology.

With this reduced list of Hungarian word forms, I utilized regular expressions to convert the relatively transparent Hungarian alphabet to ASCII characters with a 1:1 relationship between contrasts and characters.

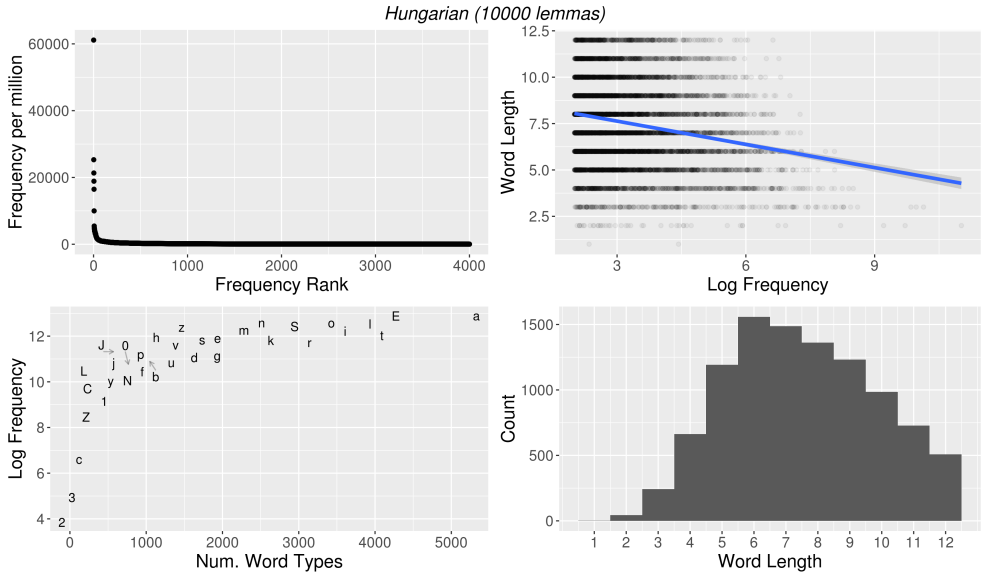

Figure 2.15: Details for Hungarian word forms.

Italian (Indo-European, Romance)

I collected Italian word frequency data from the Paisa corpus (Lyding et al., 2014), a collection of web texts collected via the same method as the WaCky Corpus (Baroni and S. Zanchetta, 2009). The corpus had already undergone cleaning and morphological parsing, making it trivial to restrict the word forms that I used to lemmas. I then used the epiTran (Mortensen et al., 2018) package to convert Italian orthography to IPA representations. I further processed the transliteration forms via regular expression to condense multi-character sequences for single contrasts to single symbols.

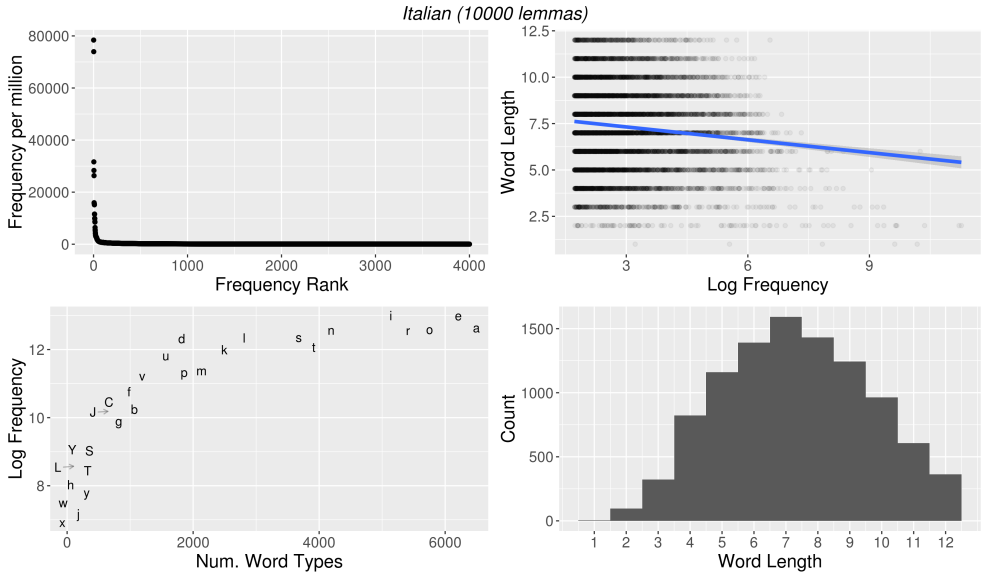

Figure 2.16: Details for Italian word forms.

Japanese

I collected Japanese data from the Callhome corpus for Japanese (Canavan et al., 1997), which contained a collection of transcribed phone calls. All word forms were transcribed in the Romaji transliteration scheme and included morphological information. Becuase Romanji does not always have a one-to-one relationship with contrastive phonemes, I used additional regular expressions to condense all multi-character sequences for a single sound to a single character.

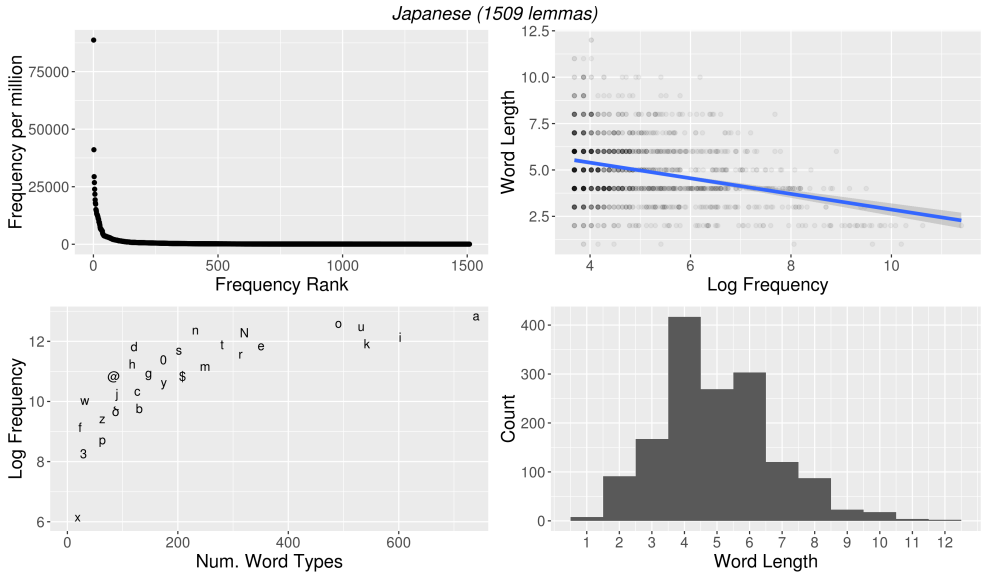

Figure 2.17: Details for Japanese word forms.

Kaqchikel (Mayan)

My data for Kaqchikel was provided to me by the authors of Tang and Bennett (2018). As Kaqchikel is a relatively lower resourced language, their source was compiled from a variety of legal, religious and literary texts. The the raw data (which I was graciously given) still possessed some obvious Spanish borrowings, i.e., words or names that did not conform to Kaqchikel spelling conventions, and other non-word entries, e.g., the page number 'xvii'. Because of this, I further cleaned the source to remove forms that did not conform to Mayan spelling conventions, attempting to exclude incorrectly spelled words or borrowings.

I did so not because I believe that borrowed works are psycholinguistically processed differently than native lexical items, but because words

forms spelled with a different orthographic system may cause problems with my analyses. For instance, a 'c' (outside of the digraph 'ch') is not used in Mayan orthography and so a word that begins with that letter might be judged to be phonotactically very different from other words, when it is in fact just spelled differently from other words.

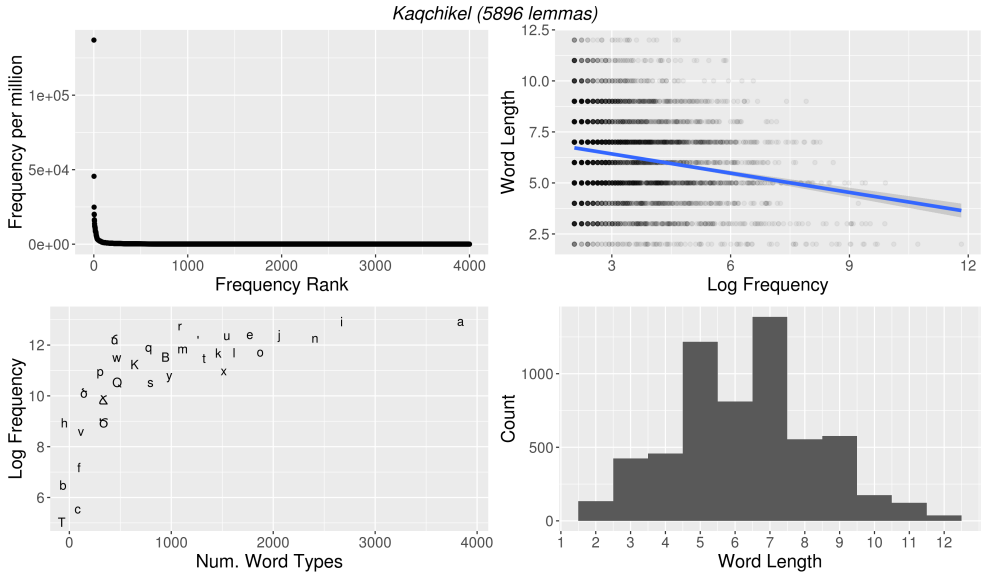

Figure 2.18: Details for Kaqchikel word forms.

Korean

My source for Korean was Lee (2006), which compiled and transcribed part of the Korean Academy Database from a large collection of written work of various genres. Due to the quality of the source, I left all phonetic transcriptions as is.

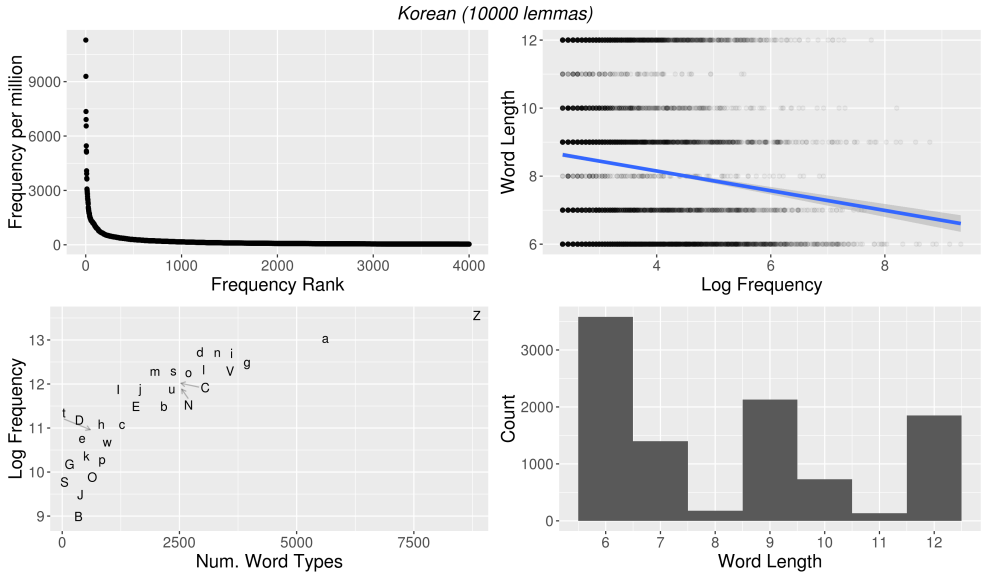

Figure 2.19: Details for Korean word forms.

Malay (Austronesian)

I collected Malay word frequency information from a corpus of Malay subtitles Lison and Tiedemann (2016). Though this source lacked any morphological information, it was collected from subtitles of movies and TV shows meaning that word frequencies it contained were more in line with spoken language. Fortunately, Malay orthography is relatively transparent and I was able to use simple regular expressions to convert Malay spelling to sequences of contrastive phonological segments.

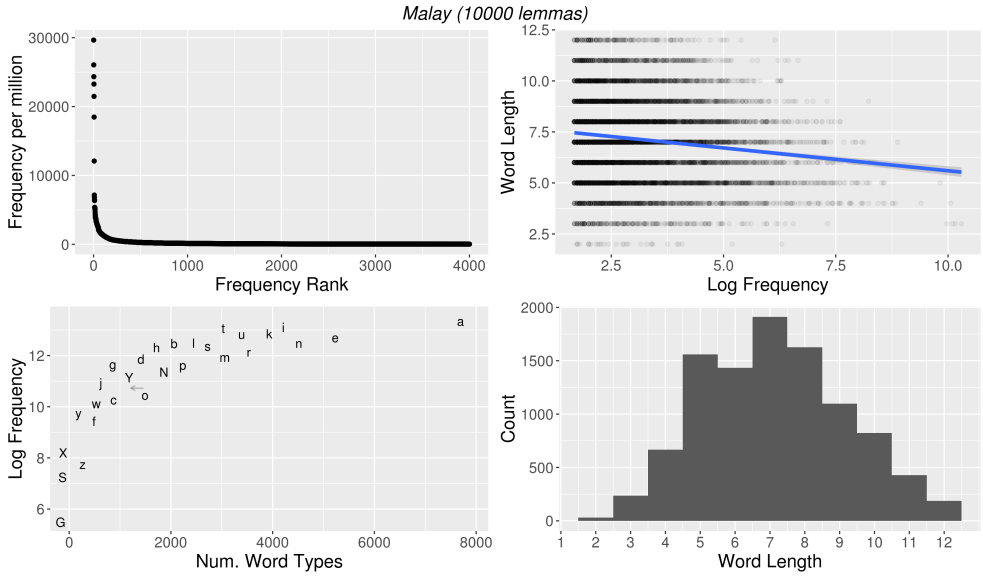

Figure 2.20: Details for Malay word forms.

Russian (Indo-European, Slavic)

The Russian source for this project was created from a collection of texts for modern fiction, politics, newspapers and popular science from 1970 - 2002 using the method described by Kilgarriff et al. (2004)[7]. I converted the Cyrillic orthography to IPA using the Python package epiTran (Mortensen et al., 2018). I then reduced multi-character sequences into into single characters for each phoneme using regular expressions.

[7]The corpus is available at `http://bokrcorpora.narod.ru/frqlist/frqlist-en.html`.

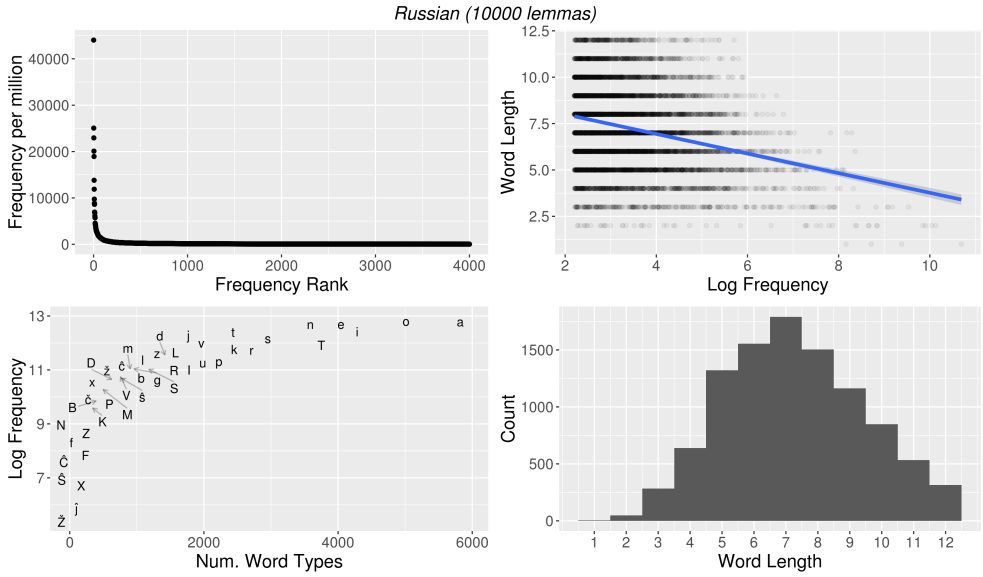

Figure 2.21: Details for Russian word forms.

Slovak (Indo-European, Slavic)

My source for Slovak was prepared section of the Slovak National Corpus (Krajčovič, 1988), which included some preprocessing, i.e. indicating lemma forms, and phonetic transcriptions. Together, the corpus was compiled from newspapers, books and other written texts over several years. Due to the quality of the source, I left transcriptions as is.

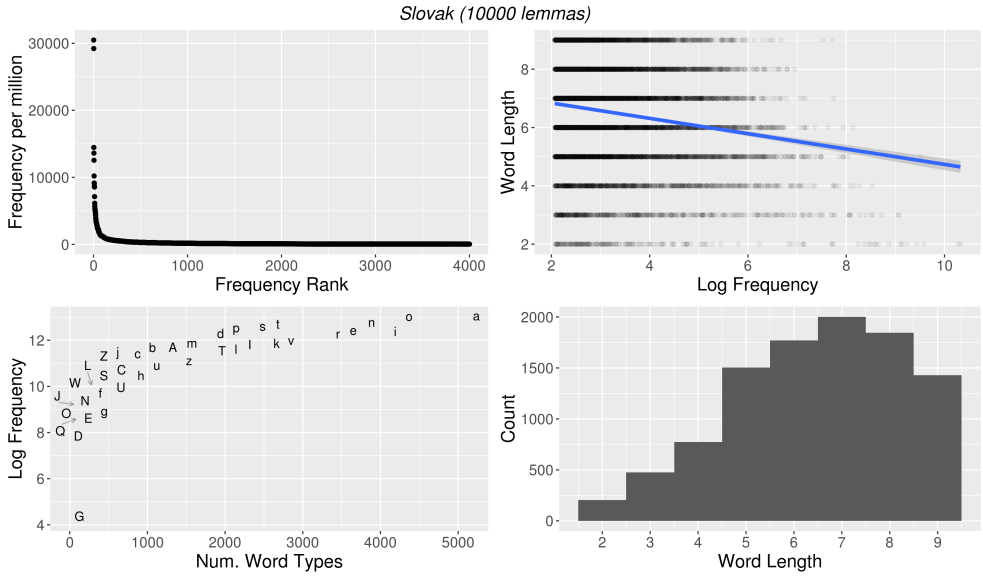

Figure 2.22: Details for Slovak word forms.

Spanish (Indo-European, Romance)

I used the GIGAword corpus (Mendonca et al., 2009) for Spanish data, which was collected from newswires over several years. All words were marked for their lemma form and provided with transcriptions for Castilian Spanish. I left all transcriptions as is, though I did condense all accented variants of vowels to unaccented forms.

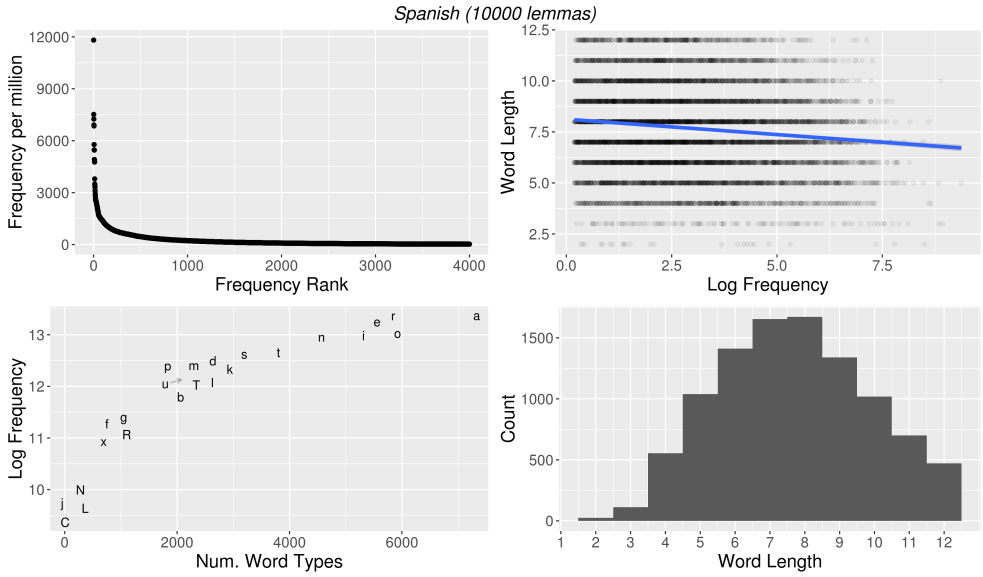

Figure 2.23: Details for Spanish word forms.

Swahili (Bantu)

I collected Swahili data from the Helsinki Corpus of Swahili (Hurskainen, 2004). This source was a compilation of newspapers and other written sources of the language, and included a morphological parse for each token. To create this resource, I downloaded and parsed the entire corpus, collecting lemma frequencies for each token. Because the Swahili orthography is relatively transparent, I kept word forms as is, save for a few regular expressions to account for some multi-character sequences.

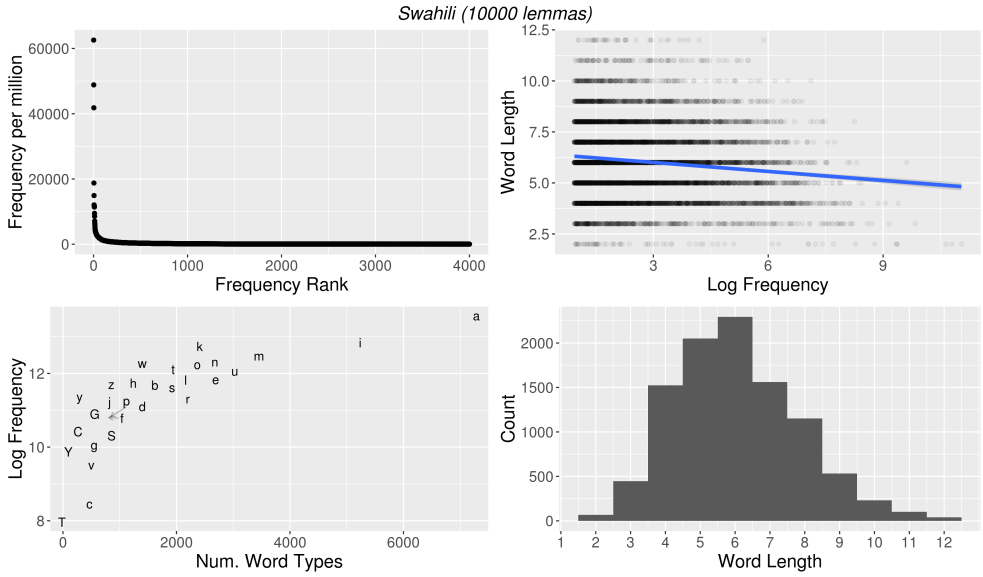

Figure 2.24: Details for Swahili word forms.

Tagalog (Austronesian)

My source for Tagalog was a collection of 1,000,000 Wikipedia pages, compiled by the Leipzig Source for Under-resourced languages (Goldhahn et al., 2012)[8]. Because all documents came from Wikipedia, there were some English words, e.g., 'subscribe' or 'donate', that had relatively high frequency. To mitigate this, I used a list of English words in NLTK (Loper and Bird, 2002) to remove any English words from the data. In addition, I removed all word forms that did not conform to Tagalog spelling conventions, similar to what I did for Kaqchikel.

[8]I did attempt to use a collection of subtitles, as I did with Malay. However, the amount of data available (roughly 10 thousand Tagalog tokens compared to 7 million for Malay) was much smaller than was available for written sources. Because of the degree of this difference, I opted to use written sources instead.

Due to the transparency of the Tagalog orthography, I left word forms as is, except for condensing multi-character sequences with regular expressions.

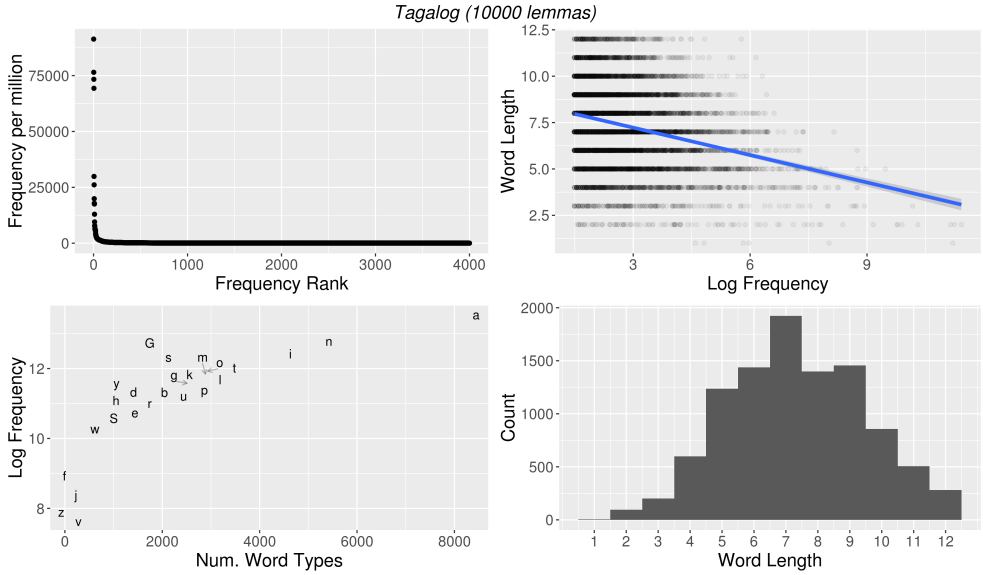

Figure 2.25: Details for Tagalog word forms.

Turkish (Altaic)

The Turkish data that I used was made available by Sak et al. (2008). This corpus was collected from web data and processed via a proprietary parsing algorithm to decompose the (often very complex) agglutinative forms of Turkish words. For this work, I only collected the forms that were marked as lemmas. Fortunately, the corpus included high quality phonetic transcriptions, which I left as is.

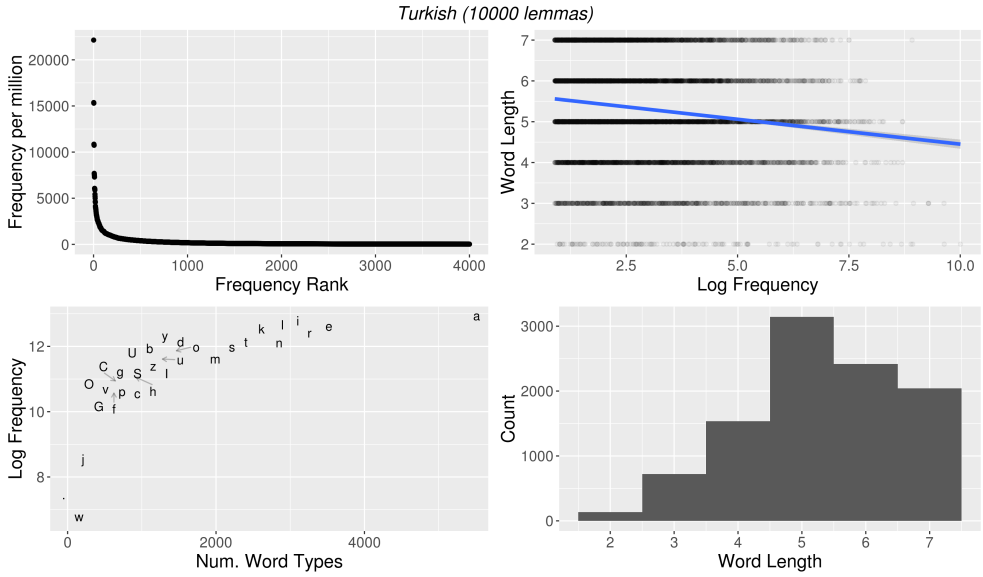

Figure 2.26: Details for Turkish word forms.

Vietnamese (Austroasiatic)

My source for Vietnamese data came from viWac (Kilgarriff et al., 2014), a collection of web pages, Wikipedia articles and news articles in Vietnamese. A crucial benefit of this source was that it included morphological parsing, providing counts for word lemmas. Because the Vietnamese orthography regularly separates morphemes with white space, without this parsing it would be impossible to determine the difference between compounding/affixation and separate words.

I represented contrastive tones as a character after a monophthong or diphthong, which I removed when determining word length. As I discussed for Cantonese, this may not be a perfect means to represent contrastive tone, though it avoids many other problems. To convert the Vietnamese orthog-

raphy to ASCII characters, I used system of ordered regular expressions.

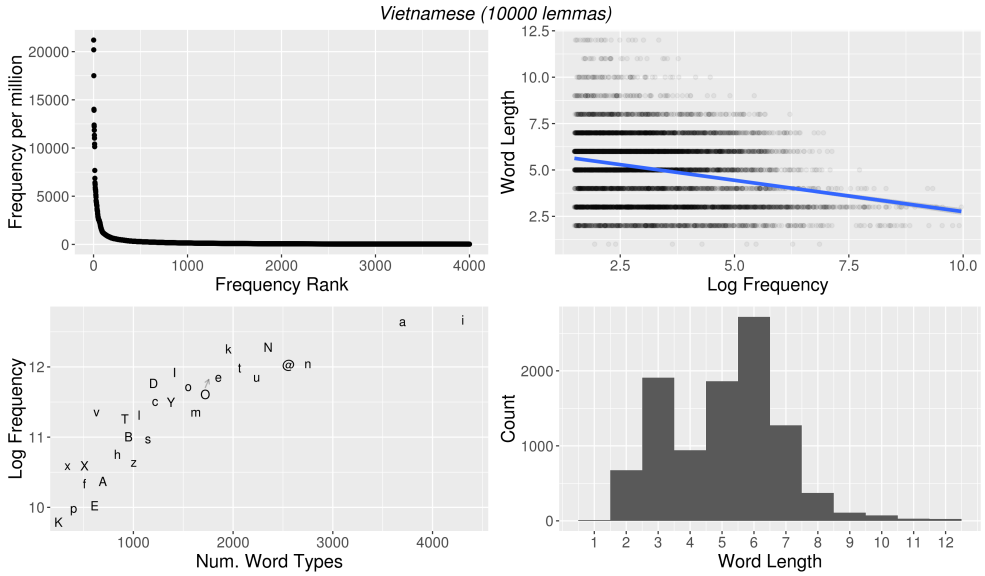

Figure 2.27: Details for Vietnamese word forms.

2.1.1 Conclusion

In the previous chapter, I have discussed the data that I will use and how it was prepared. I have also discussed some of the general methodology that I will use in the following chapters. The purpose of this chapter was to give an overall idea of the type of data and decisions that I made while collecting and preprocessing it. As best as I could, I endeavored to err on the side of caution when choosing how to represent the data, in order to limit the effects that my choices while curating the data may have on the results.

Nevertheless, it is my hope that number of represented languages and the general quality of the prepared data will add strength to any results found

in later sections. As much as possible, I attempted to collect a diverse set of languages in order to minimize the effect of a single language family or macro-area. All language data is currently available on GitHub[9] for those who wish to replicate the results here or use the data in their own experiments.

[9] http://github.com/AdamKing11

Chapter 3

An Efficient Lexicon for Incremental Processing

Background

The standard Information Theoretic definition of an efficient communication system is one that transmits as much information as possible with as little time and effort as possible (Shannon, 1948)[1]. For example, consider a variant of English where each word was twice as long. This new English would be able to transmit equal information to the original English, though it would

[1]The original definition of efficient communication only referred to time, not effort. The focus on time or number of symbols as the main focus for efficiency arose as a product of Information Theory's birth in the era of telegraph communication where the transmission of each symbol was equally costly, meaning that the raw number of transmitted symbols was what defined the overall cost of transmitting an entire message.

take twice as long and be less efficient[2]. On the other hand, consider a language where every word is [ba]. This language would have maximally short word forms, though it would fail as a communication system because word forms would not contain sufficient information to be distinguished from others in the lexicon. Neither English nor any other language is at either of these extremes, posing the question whether language is shaped to transmit as much information with as little time and effort as possible, while still maximizing accuracy. In other words, lanugage exists as a balance between different pressures and is structured to benefit both speakers and listeners (e.g., Zipf 1949; Köhler 1987)

To discuss this in detail, it is necessary to piece apart how exactly information is transmitted in spoken language, at least abstractly. Messages in spoken language are realized as a sequences of meaningful units, i.e., words, and the words are realized as a semi-unique sequence of contrastive units, i.e., *word forms*, which a listener uses to identify the original words and decode the original message[3]. Identifying words is a balance between information in the acoustic signal and the information of the larger linguistic context (for

[2]That is, it would be less efficient assuming that the extra length does not make messages more robust to noise.

[3]Note that I am using the terms 'word' and 'word form' to simply mean the mapping between a meaning and a sequence of contrastive phonological symbols. That is, I am not predicting a salient difference in how information is transmitted in isolating languages such as English or Korean or more agglutinative language like Finnish or Turkish. In both cases, linguistic messages are composed of sequences of meaningful units which must be identified. Whether or not these units lie in the same morphological domain, i.e., as different 'words' or affixes in the same 'word', is not relevant, though an interesting question for future work.

more, see Lindblom 1990; Bell et al. 2003; Aylett and Turk 2004, 2006; Jaeger 2010; Seyfarth 2014; Hall et al. 2016). As such, words that are less probable on average often require more information from the acoustic signal to be accurately identified. Generally, longer words should contain more information as they contain more material to distinguish themselves from others in the lexicon. However, the information that each sub-part of the word contains is not necessarily equal. Rather, it is a function of how uniquely it disambiguates the current word from from the other word forms of the lexicon. As a result, some parts of word forms can potentially contain more information than others. The earlier question can now be made more explicit. Are the word forms of languages constructed to communicate the information needed to identify words efficiently, beyond Zipf's *law of abbreviation* and the effect of short lengths?

Interestingly, there is strong, cross-linguistic evidence that the forms of the least probable words are longer than those of probable words (Zipf, 1935; Piantadosi et al., 2011; Bentz and Ferrer-i Cancho, 2016), allocating more information to word forms that are less able rely on contextual information. This relationship also causes the lexicon to be more efficient, because the words that are used the most are shorter. Of course, an efficient lexicon should go beyond short lengths. As the literature stands at this point, there is in-depth work on word forms as easy to produce, beyond short length (Zipf 1949; Dautriche 2015; Dautriche et al. 2017; Meylan and Griffiths 2017; Mahowald et al. 2018, for review see Gibson et al. 2019). If the lexicon is

indeed structured to communicate as much information as possible with as little time and effort as possible, word forms should be structured to be maximally informative. To discuss this, it is necessary to talk about how words are processed in speech.

Word recognition is a competitive process (e.g., Luce and Pisoni 1998; Allopenna et al. 1998); each sub-part of the word provides information as the word unfolds over time and listeners incrementally update their beliefs for the current word, keeping in mind the larger linguistic context (for more, see Dahan and Magnuson 2006; Magnuson et al. 2007; Weber and Scharenborg 2012). Because words are processed incrementally, the information that part of a word provides is dependent on what has already been processed. For example, the [v] in *vacuum* contains enough information for a listener to exclude all word forms that do not begin with [v]. The [m], on the other hand, contains significantly less information because the previous segments have already radically reduced the set of possible word forms. Compare this with the word *motive* where the [m] is relatively high information and the [v] relatively low. In these cases, the difference in information between [v] and [m] is not dependent on the segments themselves, but rather on their intra-word contexts.

Considering this, a measure of segment information can be operationalized as a function of the word forms excluded by new material. In this case, I will represent the information of a segment at position i of a word, $h(s_i|s_1...s_{i-1})$, as the $-\log_2$ contextual probability of that segment, given the

previous segments of the word form, $s_1...s_{i-1}$ (Eq. 3.1, c.f. van Son and Pols 2003). For example, the information of the [æ] in *vacuum* is equal to 4.2 bits because the probability of encountering a token with an [æ] following a word-initial [v] is $\frac{392}{7666} \approx .05$[4].

$$h(s_i|s_1...s_{i-1}) = -\log_2 p(s_i|s_1...s_{i-1}) = -\log_2 \frac{count(s_1...s_i)}{count(s_1...s_{i-1})} \quad (3.1)$$

Note that this measure focuses solely on intra-word contexts, ignoring the larger context of preceding words and discourse. In other words, under this equation, the determination of the information of the [æ] in *vacuum* ignores whether or not the word *vacuum* was relatively likely given the larger linguistic context or not. Though there is certain to be an effect of the larger context on how information of a contrast is processed by actual listeners, the use of solely intra-word contexts allowed me to include a larger number of language corpora. In order to accurately estimate the likelihood of words (which would be necessary in building a context-dependent measure of segment information), it is necessary to have a large corpus. This requirement would exclude some languages, e.g., Arabic, Hausa, Vietnamese, from the analysis.

Over the next sections, I will show that word forms are structured so that the lexicon as a whole is a higher information code than might be expected

[4]This is assuming a token-based measure of segment information.

otherwise. Specifically, I will show that inter-word phonological contrasts are organized to be higher information on average then would be expected otherwise. I will then show that the association between word probability and form causes the most probable points of contrast in the lexicon to be the most informative on average. Both of these are important features of an efficient code, focused on the information content of sub-parts of the word.

3.1 Balanced contrasts

Given this definition for the information of part of a word form, it is now possible to discuss how to structure the lexicon for high information contrasts. Because segment information is a function of competing word forms that are excluded by it, each segment should aim to exclude as many words as possible. Consider Fig. 3.1 below. Here, the lexicon is modeled as a probabilistic tree where each segment represents a contrast point between word forms. With each successive branch, part of the lexicon is excluded until a single word remains. In this case, the relative probability of each branch at a contrast is perfectly balanced. That is, the summed probability of the set of word forms of each branch is equivalent, e.g., the probability of [ð]-initial words is equal to that of [s]-initial words. Because each contrasting point is organized to have equiprobable branches, the *entropy*, or average information, at each contrast is maximal. Therefore, given the size of the lexicon and the relative probabilities of each of the word forms, this is an maximally efficient way to

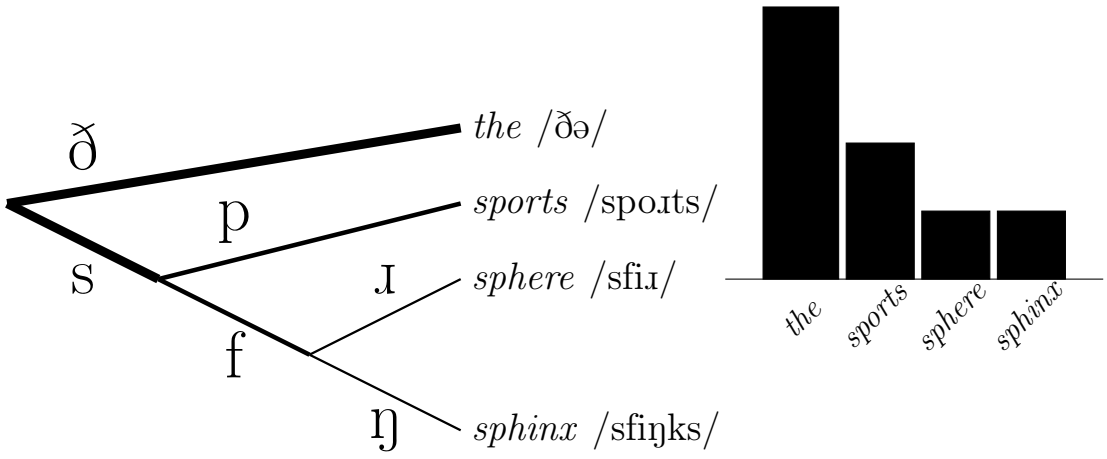

Figure 3.1: Contrast structure of a toy version of English. On the right, the bar chart represents the probability of each word. On the left is a branching diagram of contrasts in all word forms, with line thickness representing that probability of each branch, i.e., the summed probability for all words that lie along a branch. The ideal lexicon is one where the average information at each branching point is maximal. Given the specific set of branches and probabilities of words, each branch is probabilistically balanced, meaning that entropy at each point is maximal.

structure the lexicon.

Granted, this is an overly simple representation of word forms and word processing broadly, but it allows for explicit measurement and testing of the lexicon as a communicative code. Furthermore, regardless of how the lexicon is represented, the requirement for an efficient code remains the same. Explicitly, the entropy of each of the various contrasts throughout a message should be high. This in turn means that the probability for each of the possible values for a contrast, e.g., segments, should be close to balanced; the entropy for any variable is maximal when all possible outcomes are equally likely (Shannon, 1948).

Therefore, if the lexicon is structured for high information, contrasts should show a significant effect of probabilistic balancing. That being said, testing whether *all* contrasts in the lexicon are balanced is likely to be a difficult and non-straightforward task. Firstly, a pressure for high information contrasts is but one of many pressures that likely affect the shapes of words, meaning that other pressures may cause certain parts of words to be imbalanced. Secondly, there is no a priori baseline for the expected effect of probabilistic balance. For these reasons, I will focus on a single part of word forms, word-initial position, and I will employ two means to test the relative balance of probabilities indirectly. Put together, this represents a testable way to investigate whether the lexicon shows greater evidence of balance than might be expected otherwise.

3.1.1 Relation between type and token frequency

As mentioned before, determining whether or not contrasts are balanced in the lexicon is not straightforward. Contrasts will very likely not be *perfectly* balanced, the question here is whether they are closer to perfect than would be expected otherwise. To visualize a "closer to perfect" system, it may be helpful to think of the opposite, an imbalanced system of contrasts. If contrasts were not balanced, then the segments at the beginning of the most word types would also be at the beginning of the most frequent words, on average. Simply put, if there were no effect of balancing, both frequent and relatively infrequent words would be equally likely to begin with the same

segments, meaning that the probability of a word-initial segment would be a simple function of how many word types begin with it.

For example, consider an abstract language with three possible word-initial symbols, $\{a, b, c\}$, and 7 words which follow a Zipfian frequency distribution[5]. If word forms were created by randomly choosing a symbol to begin with, the probability of a word beginning with a particular symbol, e.g., $p(s_1 = a)$, would be solely determined by the number of word types that begin with it.

On the other hand, what relationship between type and average token frequencies would result in greater balance of contrasts? For word-initial entropy to be at its maximum, the total probability of a word form beginning with each of the three symbols would need to be balanced, $p(s_1 = a) = p(s_1 = b) = p(s_1 = c)$. A possible way for this to be the case, is if a begins the form for the most frequent word, b begins the forms for the second and third most frequent words and c begins the remaining four forms. Were this the case, there would be a negative relationship between the average frequency of words that begin with a, b or c and the number of word types that begin with that contrast.

Following from this example, I predict that there is actually an inverse relationship between type and average token frequency for contrasts in the lexicon. Consider English where few words types begin with [ð] while many

[5]I am assuming a Zipfian distribution of frequencies because all tested natural languages show this distribution for frequencies (Ferrer-i Cancho and Solé, 2003; Bentz and Ferrer-i Cancho, 2016).

more begin with [s]. However, [ð]-initial words, e.g., *the*, *then*, *there*, are more frequent on average than that [s]-initial words, causing there to be a negative relationship between type and average token frequencies for [ð]- and [s]-initial word forms. This pattern repeats with the second position in word forms and the number of [ð]-initial word types that continue with [ɛ], e.g., *then*, *there*, is greater than that of [ə], e.g., *the*, though the average frequency of [ðə]-initial words is greater. If the lexicon is structured to convey high information per contrast, the frequencies and forms of words should be organized so that at each contrast position, each possible contrast is closer to equiprobable than might be expected otherwise.

Note that this negative relationship follows directly from a balanced distribution of contrasts. Because the lexicon contains a relatively small number of very high frequency words, in order for the segments that are not found in these words to be balanced with those that are, the segments not in high frequency words must occur in more word types overall. Therefore, a negative relationship between type and token frequency is a necessary result of a lexicon organized for balanced contrasts.

Preparation

In order to test this, I will focus on a single contrast in word forms: word-initial position. Note that I am not predicting that *only* the beginnings of word forms are balanced; this should be a trend across all positions in word forms. Restricting the analysis to a single position mitigates some potential

confounds of testing all positions and allows for a simpler cross-linguistic comparison.

I have chosen to focus on word-initial position because i) ignoring sentential context, all words compete at this position, ii) word-initial positions are generally representative of the phonotactic distributions across words (Stojanović, 2013), iii) listeners preferentially focus on word beginnings during processing (Nooteboom, 1981; Salasoo and Pisoni, 1985; Marslen-Wilson and Zwitserlood, 1989; Connine et al., 1993) and iv), it is an easily comparable position across languages because all word forms have beginnings, regardless of language. If the lexicon is structured to create a more balanced contrast at word beginnings, then the contrasts that begin fewer word types should begin words that are more frequent, on average.

I measured the type count for a given distinct value for a lexical contrast, c, as the number of word types that began with it (Eq. 3.2) and the average token frequency of a contrast as the average frequency of all words, w, that began with it (Eq. 3.3).

$$\text{count}(c) = \sum_{w} 1 \text{ iff } c \in w \tag{3.2}$$

$$\text{mean\_freq}(c) = \frac{1}{\text{count}(c)} \sum_{w} \text{freq}(w) \text{ iff } c \in w \tag{3.3}$$

Whereas in other parts of this work, I use individual, contrastive phonological segments, i.e., *phonemes*, to represent lexical contrasts, here I used

biphone sequences, i.e., ordered pairs of two phonological segments. I did so because the total number of contrastive segments allowed at word-initial position can be rather small, limiting the power of statistical testing. For example, in a language with a relatively small phonemic inventory, a single unusually frequent or infrequent segment may cause the tests to fail to reach the threshold for significance.

Corrections for Zipfian distributions in the lexicon

For word-initial, as well as all positions in word forms, there is a strong Zipfian distribution of segment frequencies (e.g., Schiller et al. 1996; Kessler and Treiman 1997; Baayen 1992); the most frequent contrast by either token or type count is many times more frequent that the least frequent. This causes the majority of the total probability mass to be distributed among a small set of categories. A simple and widely applied approach to dealing with Zipfian distributions is to take the logarithm of the raw values (Manning and Schütze, 1999; Jurafsky and Martin, 2014). Considering this, I will be using the *log* type count for biphones and the *log* average frequency for words that begin with a given biphone (see Fig. 3.2).

Yet, that is not the only correction that need be made. Zipfian distributions carry with them a long *tail* of very small values and many extremely low frequency biphones occur in an equally extremely small number of word types. Put another way, in the limit of when a biphone appears once in a corpus, it necessarily appears in only one word type. On the other hand,

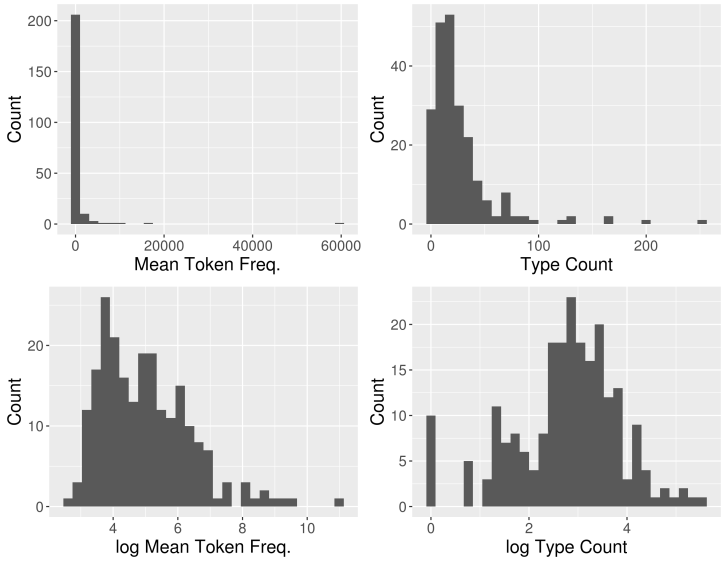

Figure 3.2: Comparison of raw token frequency and type count (top) for word-initial biphones in English, and their respective log-transformed values (bottom). The distributions for both token and type are clearly more normal after log-transformation.

when a biphone is more frequent in a corpus, there is potential for it to be distributed among a varying number of word types.

The necessary correlation caused by very low frequency biphones potentially may skew the data towards a positive relationship between type and token frequency, though this relationship would be driven by words which constitute only a small fraction of the total probability mass of the lexicon altogether. For example, only a single word type in English (according to my corpus) begins with [ɨk], *ecstatic*, and [ɨk] would equally affect the determination of the statistical correlation between type and average token frequency as [ðə] or [ɪn], both of which account for a much larger proportion of English

tokens. If the correlation was determined over only biphones that individually make up a sizable proportion of the corpus, this structural correlation weakens (see Fig. 3.3).

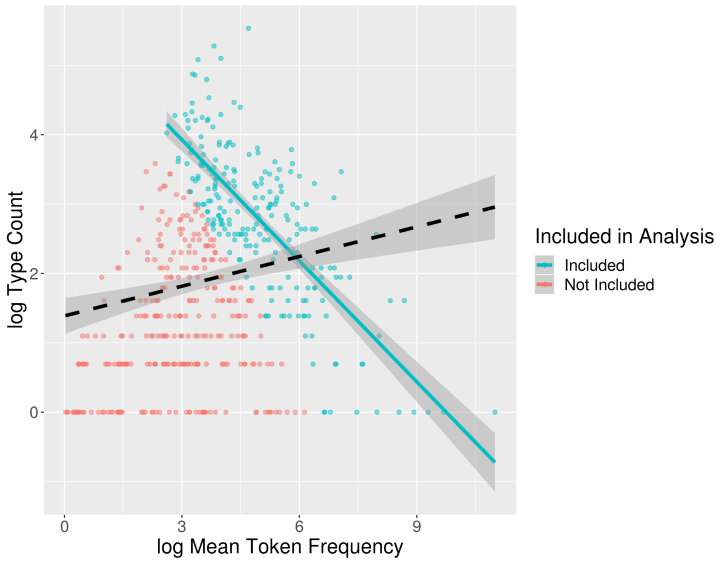

Figure 3.3: Relationship between mean token frequency (x-axis) and type count (y-axis) for word-initial biphones in English. Blue points indicate biphones present in the final analysis and red points indicate those removed (<5% total probability mass). Dashed, black line shows linear relationship for all points, solid, blue line shows the relationship for only points included in the analysis. When all points are included in determining the best-fit line, there is a clear positive slope. When only the top 95% of points are included, the slope is negative.

In order to mitigate the structural correlation caused by the Zipfian tail, I chose to remove the least probable biphones, until 95% of the total probability remained (see Fig. 3.4). To determine which biphones to remove, I first calculated the overall probability for each word-initial biphone by summing the token frequency for word forms that began with the target biphone and

then dividing by the the total token frequency for the entire lexicon. I then sorted biphones by their probability in ascending order. Starting with the least probable, I removed biphones until 5% of the total probability mass of the entire lexicon had been removed. This removed biphones such as [lu], [və] and [ʃɹ] which only occurred in a single or few low-to-medium probability word types each (*lute*, *veneer* and *shrine* respectively), while keeping biphones such as [kə] which occurred in larger sets of words of varying probability, e.g., *contractual*, *control*, etc.

I chose a threshold of 95% because in order to test across as large a section of the lexicon as possible, while minimizing the potential effects of the Zipfian tail. For example, Schiller et al. (1996) found that 85% of all syllables in a Dutch corpus were instances of roughly 50% of all Dutch syllable types. Put another way, though the word forms in the Dutch lexicon are composed of many thousand distinct syllables, the vast majority of the syllables that a Dutch listener hears are tokens of a few hundred. In this case, choosing a threshold of 95% avoids the Zipfian tail while being slightly more conservative. Though this resulted in removing a large number of biphone types, this was a way to minimize any problematic effects of the Zipfian distribution while still testing a large enough part of the lexicon[6].

Another way to visualize this is by looking at the total frequency for biphones, i.e., the sum of token frequencies for all words that begin with a

[6]Note that I continued to *log* transform values even with this correction. I did so in order to make the variables closer to normal distributions, which are more ideal for correlation tests.

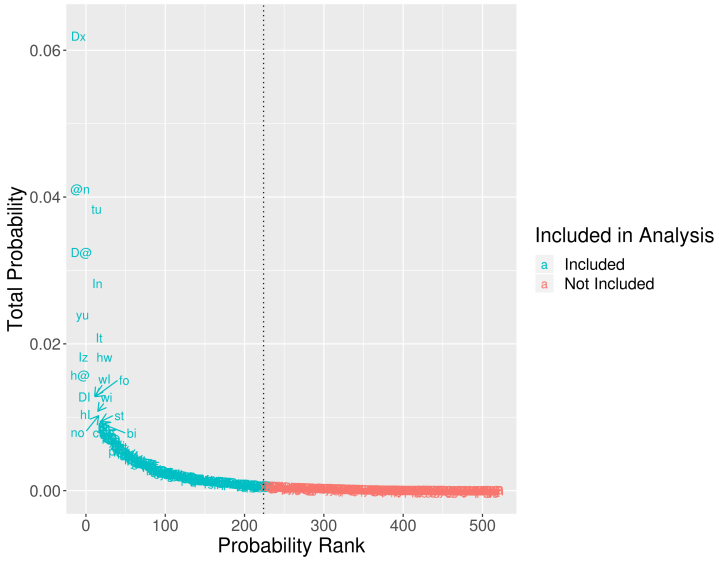

Figure 3.4: Word-initial biphones included for English. The x-axis ranks biphone contrasts by the total probability of word forms that begin with the segment in question and the y-axis indicates the probability for a given biphone. The dotted line represents a collective sum of 95% of the total probability, with biphones on the right of the line excluded.

particular biphone (Fig. 3.5). Looking at all biphones, there is a strong, positive relationship between type count and total token frequency. That is, the most probable initial biphones overall are the ones that lie at the beginning of the most distinct word types. When only the top 95% of biphones are considered, the relationship nearly disappears[7]. This shows that the positive relations ship is mainly driven by the least probable words, i.e., those that only occur one or twice in the relevant corpus.

[7]This is interesting considering that a positive relationship between type count and token frequency for material in the lexicon is a common assumption (e.g., Hockett 1967; Baayen 1992). As these data show, this relationship is indeed represented here, however it appears that the main driving force lies in the least frequent material across the lexicon.

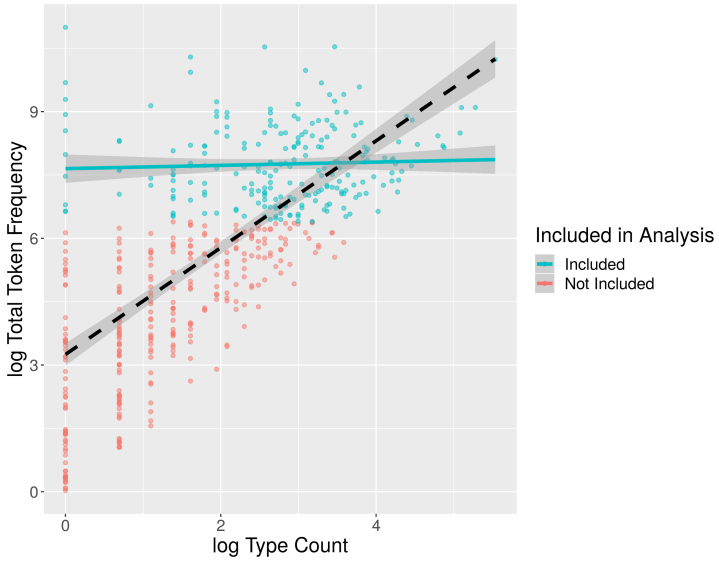

Figure 3.5: Relationship between type count (x-axis) and total token frequency (y-axis) for word-initial biphones in English. Across the entire set of biphones, there is a strong positive relationship. When only the top 95% of biphones are considered, the relationship essentially disappears. This shows that among the majority of word tokens, the probability of each word-initial contrast is more or less equal.

The astute reader may ask why I chose to investigate the relationship between type count and average token frequency, rather than with total token frequency itself. If contrasts were perfectly balanced, the total token frequency for each biphone would be identical. Similarly, there would be an inverse relationship between type count and average token frequency. These two relationships are mutually entailed by one another, though testing for the relationship with average token frequency is more straightforward.

Because the lexicon is a complex system which is shaped by many, often competing pressures, a perfectly balanced set of contrasts is unlikely.

Therefore, in order to investigate this, I would need a test for an overall statistical trend towards balance. Testing the correlation between total token frequency and type count with standard statistical tests would be difficult since the prediction would be that the two have no correlation. This means that I would need to arbitrarily specify a strength of the correlation between the two as a null-hypothesis and show that the actual data show a weaker correlation. On the other hand, by choosing to test the correlation between type count and average token frequency, I do not need to arbitrarily specify a correlation strength while still being able to test the same property of the lexicon.

Comparison to Novel Lexicons

To investigate whether the lexicons of the tested languages demonstrated the predicted properties more strongly than might be expected otherwise from a structurally similar but non-linguistic code, I chose to compare them against a more language-like baseline. It is possible that a 'significant' relationship might arise in any code of similar structure to the lexicons of the tested languages, and this served a more conservative validation of any results for the simpler tests.

To construct a baseline for each language, I created novel forms of the lexicon via nonce word forms generated from phonotactic n-gram models (for more detail, see Chapter 2.0.2). Though creating novel lexicons via probability-shuffling (see Ch. 2.0.2) is more conservative, I chose to use

comparison lexicons of nonce word forms for two reasons. Firstly, a lexicon of nonce forms allows for both type and token counts for biphones to vary with each novel lexicon, both being pathways to create a balanced set of contrasts. Secondly, probability-shuffled lexicons often resulted with far fewer remaining biphones after the removal of the Zipfian tail, making comparison to the original lexicon problematic.

Because the large majority of words posseses a frequency equal to the minimum threshold for inclusion in the corpus, shuffling frequencies causes more biphones to be associated with more minimum-frequency words, dropping the total frequency for the biphone as a whole. As an example, [ðɛ] is included in the original lexicon despite being associated with very few word types since the word types it does begin, e.g., *there*, *thereby*, are high frequency on average. When the set includes more minimum-frequency words, [ðɛ] more often finds itself in the Zipfian tail and removed. This was not an isolated example and the probability-shuffled lexicons often ended up with half as many biphones as the real-world lexicon after the Zipfian tail was removed.

In a way, this is a sign in and of itself that the lexicon is organized to create probabilistic balance, since the pressure for balanced contrast is found primarily in the top 95% of the lexicon. However, that is not sufficient evidence to make any rigorous claim and I chose to use novel word forms as an alternative method for developing a baseline.

3.1.2 Results - Linear Relationship between Token and Type Frequency

Significance Tests

I began by assessing the Pearson's correlation between *log* mean token frequency and *log* type count for word-initial biphones in each language individually. Recall that type count of a biphone was equal to the number of word types where it was at the beginning and the mean frequency was the average word frequency for all words that share an initial biphone. For all languages except Hausa, there was a significant, negative relationship between the variables, as predicted (see Figs. 3.6, 3.7 - 3.11). Note that though the Hausa data did not reach significance, it still trended in the predicted direction. Together, this shows that on a language-by-language basis, there is a strong effect of balancing between token and type frequency for word-intial biphones.

Testing all languages together, there was a significant, negative effect of *log* mean token frequency on *log* type count when tested in a linear mixed-effects model with language nested within family as random intercepts (see Tab. 3.1)[8]. This indicated that the word-initial biphones found in the most probable words were also found in fewer types, as would be expected if contrasts were closer-than-random to balanced.

[8]R formula: lmer(data = all_langs, log(count) $\sim$ log(mean) + (1|family/language), REML = F)

A. Fixed Effects:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.224	0.112	55.414	**0.001**
log Mean Token Freq.	-0.552	0.016	-34.687	**0.001**

B. Random Effects:

	Name	Variance	Std.Dev.
Language:Family	(Intercept)	0.084	0.290
Family	(Intercept)	0.004	0.064

Table 3.1: Mixed-effects model to predict *log* type count for biphones, given *log* mean token frequency, with random intercepts per language nested in language family. There is a significant effect of token frequency, showing that across all data, there is a predicted balancing between token and type frequency.

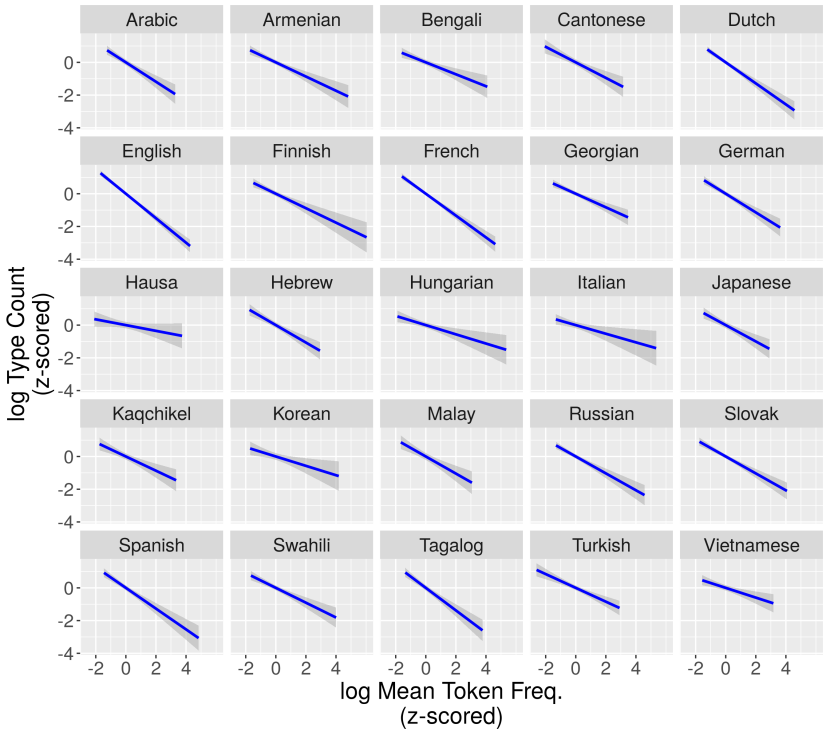

Figure 3.6: Relationship between mean token frequency and type count for all languages in the dataset. The x-axis shows *log*-transformed value for the average frequency of words that begin with a biphone and the y-axis shows the *log*-transformed number of word types per biphone.

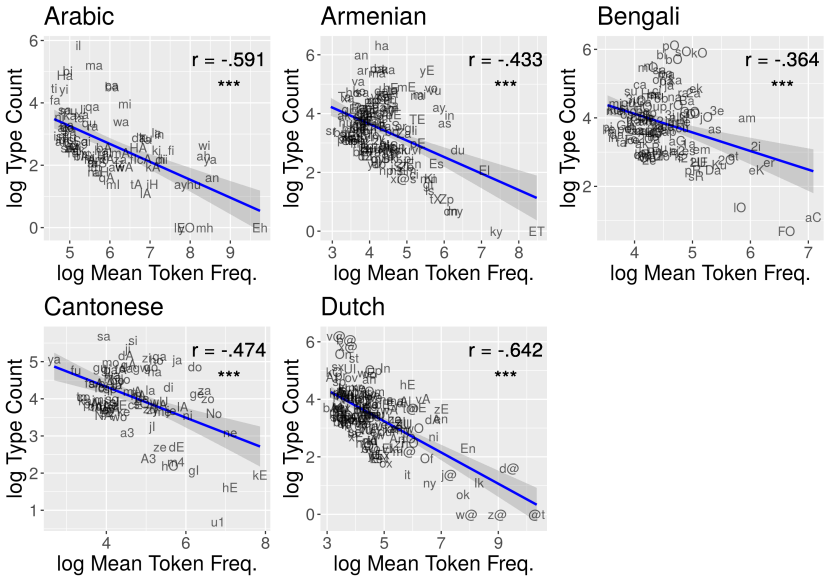

Figure 3.7: Relationship between *log* mean token frequency and *log* type count. Significant in all languages shown here.

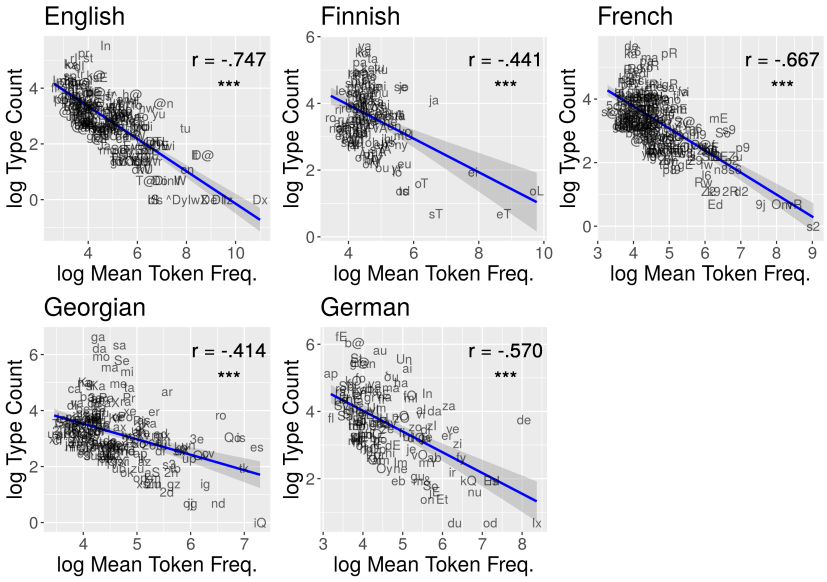

Figure 3.8: Relationship between *log* mean token frequency and *log* type count. Significant in all languages shown here.

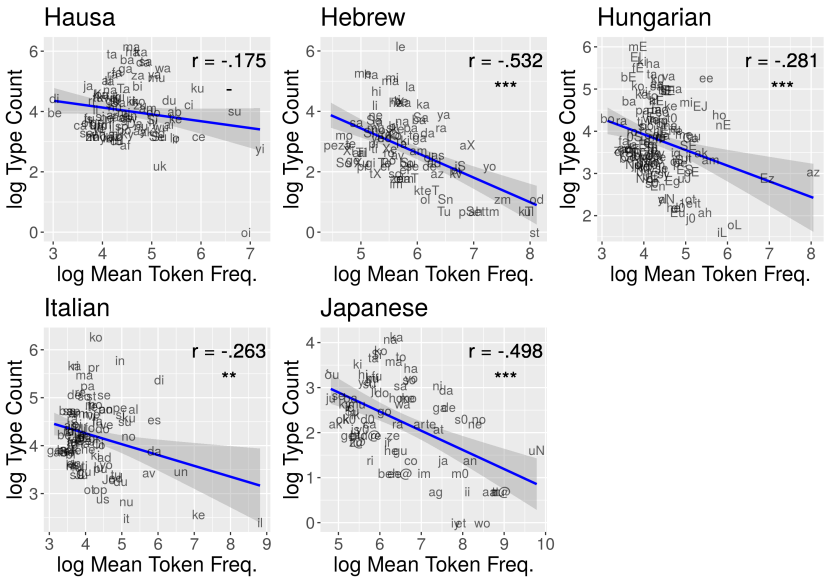

Figure 3.9: Relationship between *log* mean token frequency and *log* type count. Significant in Hebrew, Hungarian, Italian and Japanese. Not significant in Hausa, but trending towards significance ($p < .1$).

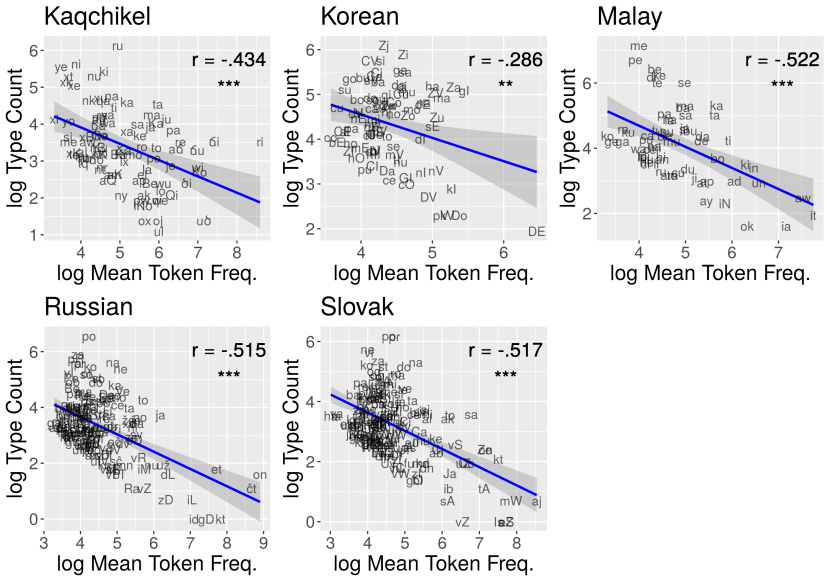

Figure 3.10: Relationship between *log* mean token frequency and *log* type count. Significant in all languages shown here.

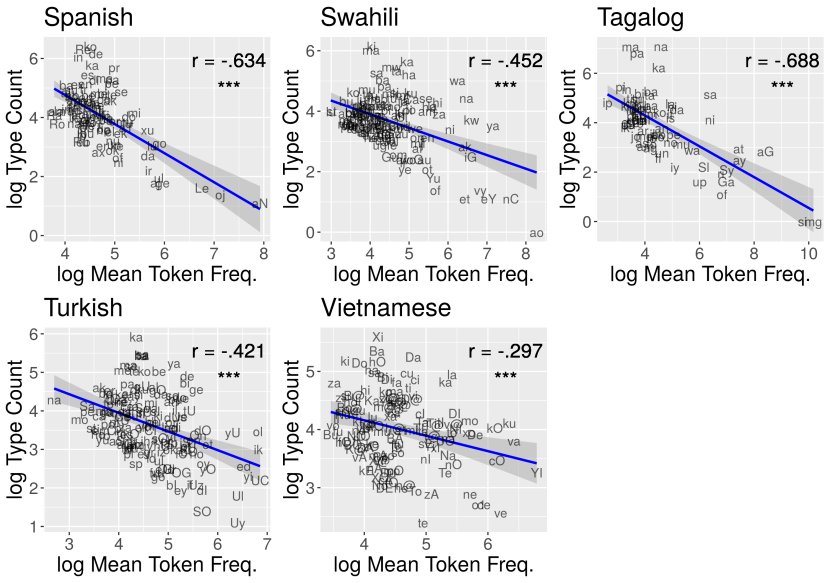

Figure 3.11: Relationship between *log* mean token frequency and *log* type count. Significant in all languages shown here.

Novel Lexicons

I then moved to comparing the correlation within languages against those of 1,000 novel lexicons. Recall that each novel lexicon was created from novel word forms for each language, thereby varying both type and token counts for biphones (see Figs. 3.12, 3.13 - 3.17). The results here were far more mixed. For each language, I first calculated the correlation between type and token counts for biphones, r, and then compared that value for each language against the same value re-calculated in the novel lexicons.

For many languages (14 of 25), the real-world lexicon showed a stronger (more negative) correlation between *log* mean token frequency and *log* type count for word-initial biphones than 95% of the novel lexicons. Interestingly, for Hausa - the sole language that failed to be significant in the previous section - the real-world lexicon had a weaker correlation than 95% of the novel lexicons, suggesting that in this language, the distribution of word-initial biphones is significantly *unbalanced*.

Though the results by language were mixed, a logistic mixed-effects model to predict lexicon type (real-world or novel), given the correlation between *log* type and *log* mean token frequency, r, showed that the real-world lexicons had significantly stronger (more negative) correlations than their novel counterparts. That is, compared to the novel lexicons as baseline, the correlation in the real-world lexicon was consistently more negative than would be expected. This suggests that though the effect may not be found in all languages individually, it is strong enough across the data to be a significant

A. Fixed Effects:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-12.282	1.684	-7.292	**0.001**
r	-13.755	2.625	-5.241	**0.001**

B. Random Effects:

	Name	Variance	Std.Dev.
Language:Family	(Intercept)	7.958	2.821
Family	(Intercept)	0	0

Table 3.2: Logistic mixed-effects model for predicting lexicon type (real-world or nonce forms), given correlation between token and type frequency and random intercepts per language per family. Model showed a significant, negative effect of r, indicating that, relative to the novel lexicons, the real-world lexicons had a stronger (more negative) correlation between the relevant factors, as predicted.

effect (Tab. 3.2)[9].

[9]R formula: glmer(data = all_langs, lex.type ~ r + (1|family/language), family = binomial)

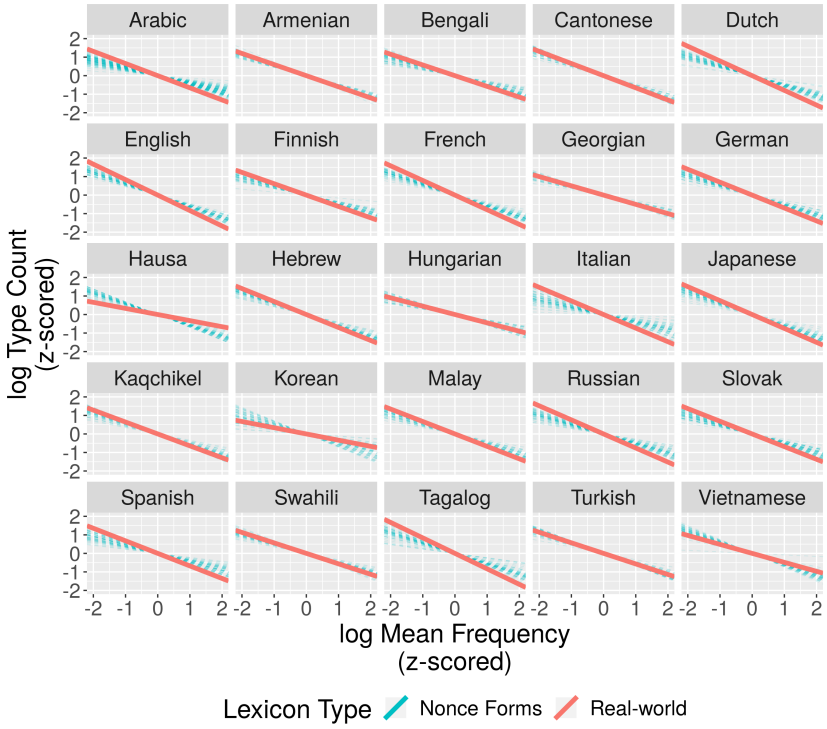

Figure 3.12: Comparison between real-world (red) and novel lexicons (blue). Each line shows the correlation between type and token frequencies for bi-phones in respective form of the lexicon. If the real-world lexicon falls below the distribution of the novel lexicons, then it shows a significantly stronger relationship than expected.

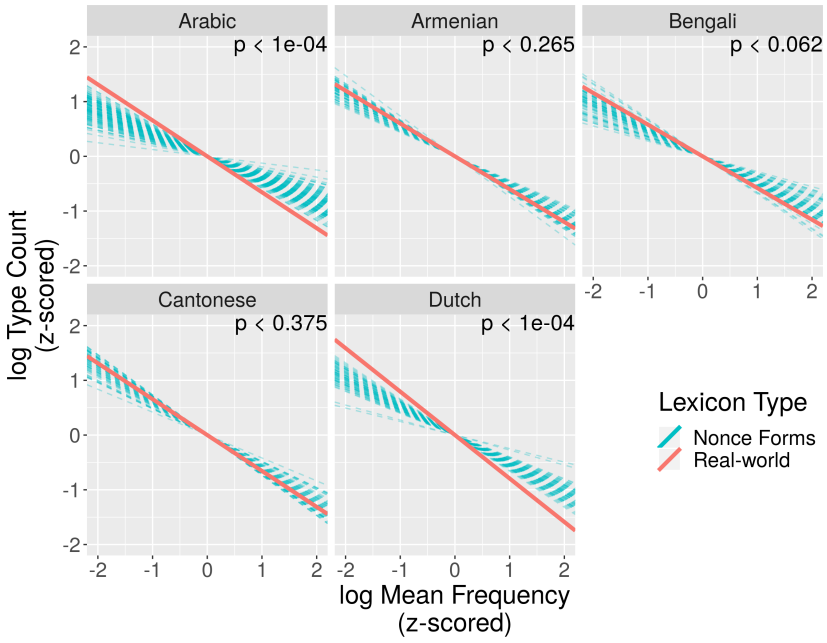

Figure 3.13: Comparison between correlation between token and type frequency in the real-world and novel lexicons. Significant in Arabic and Dutch. Not significant in Armenian, Bengali and Cantonese.

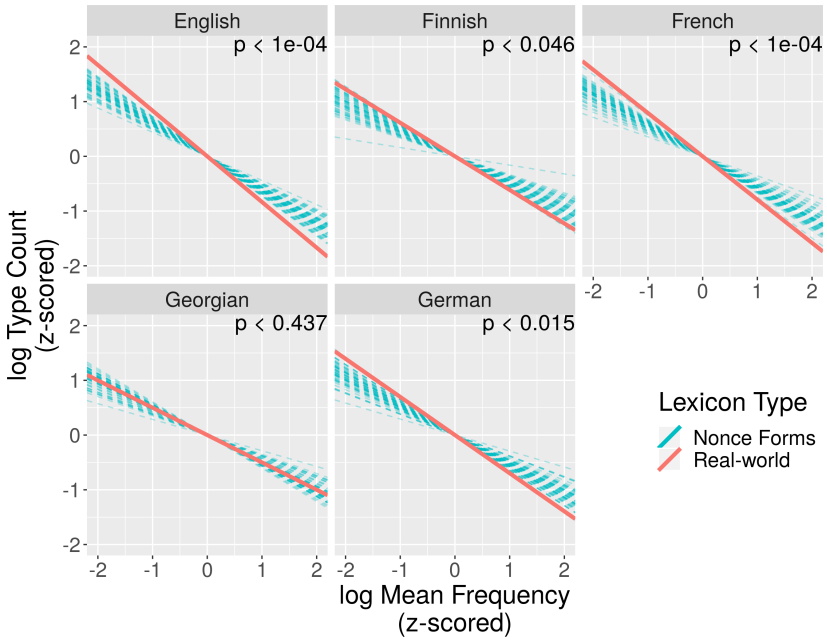

Figure 3.14: Comparison between correlation between token and type frequency in the real-world and novel lexicons. Significant in English, Finnish, French and German. Not significant in Georgian.

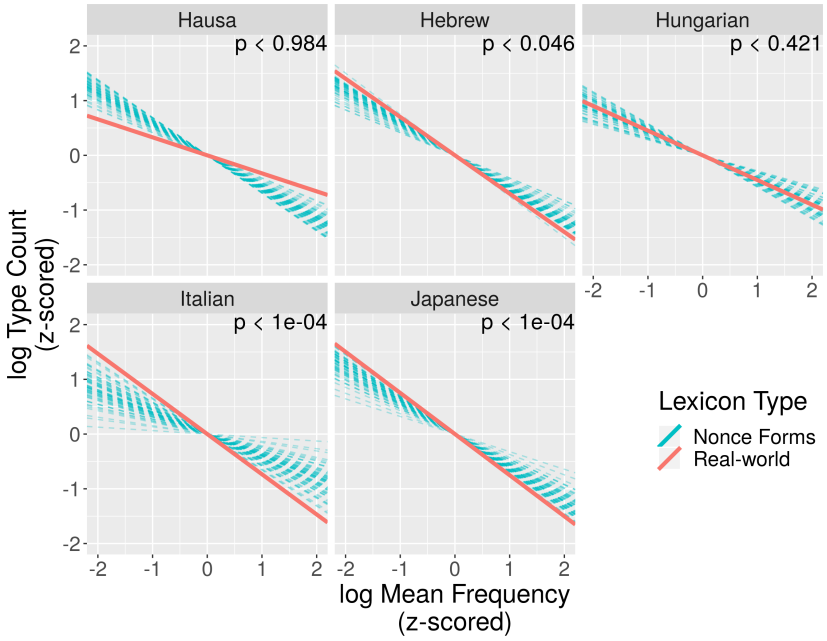

Figure 3.15: Comparison between correlation between token and type frequency in the real-world and novel lexicons. Significant in Hebrew, Italian and Japanese. Not significant in Hungarian. Significant in opposite direction in Hausa.

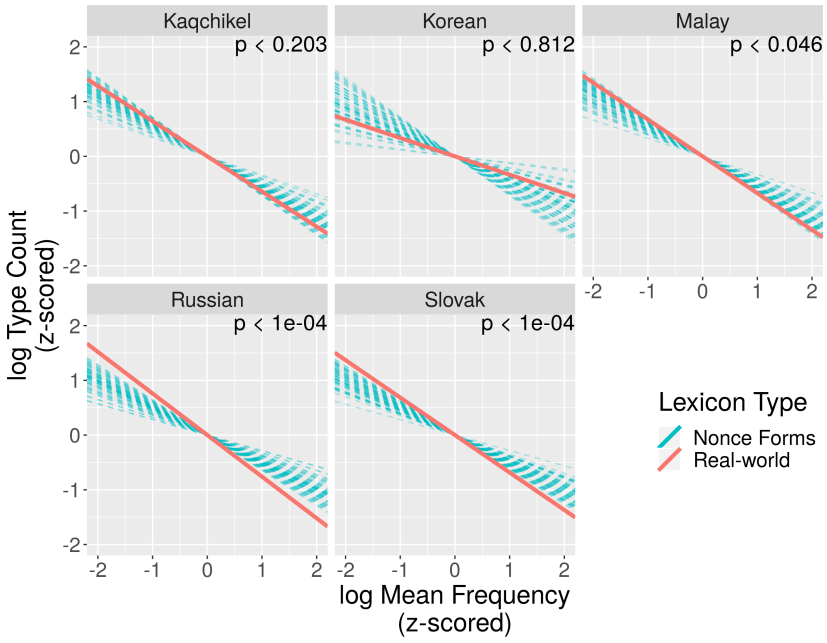

Figure 3.16: Comparison between correlation between token and type frequency in the real-world and novel lexicons. Significant in Malay, Russian and Slovak. Not significant in Kaqchikel and Korean.

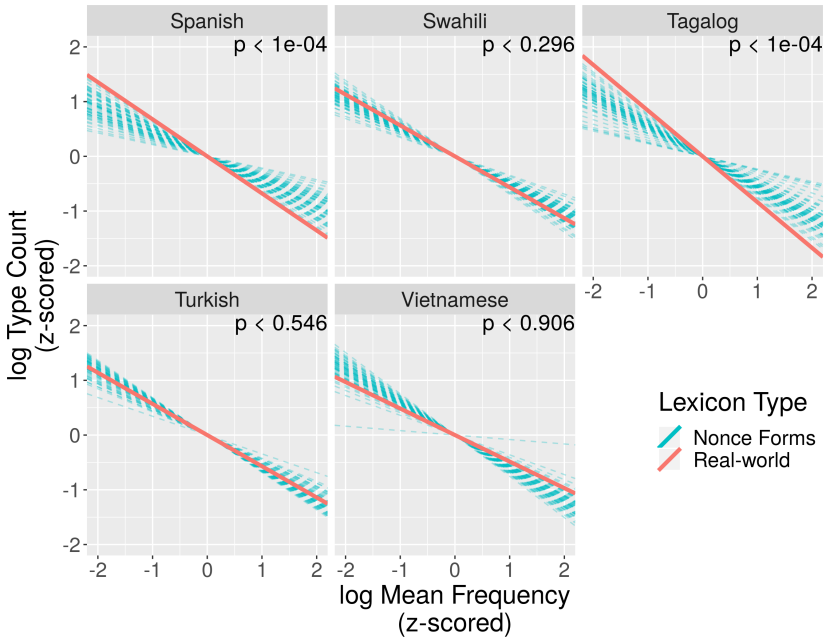

Figure 3.17: Comparison between correlation between token and type frequency in the real-world and novel lexicons. Significant in Spanish and Tagalog. Not significant in Swahili, Turkish and Vietnamese.

3.1.3 Word-initial entropy

In the second set of tests of word-initial contrasts, I will compare the entropy word-initial contrasts against a baseline distribution for the language at other positions. As mentioned earlier, there are other pressures for a non-balanced distribution of contrasts throughout the lexicon and this may bias all contrasts away from equiprobability. Put simply, word-initial contrasts may not demonstrate a clear, linear relationship between type and token probabilities because the lexicon-wide distribution as a whole is greatly asymmetric. However, because of the privileged status of word-initial material in word identification, the word-initial position may be relatively closer to balanced compared to the lexicon as a whole.

Consider the system of stop consonants in Armenian, which has a 3-way voicing contrast (voiceless aspirated, voiceless unaspirated, voiced) for each place of articulation. However, voicing contrasts collapse word-finally and word-internally when adjacent to other consonants (Vaux, 1998). Because of Armenian's phonology, the distribution of stops at word-beginnings (where the segments are not subject to neutralization) is more balanced than it is at later points, meaning that the entropy of the word-initial system is greater than the entropy of the same segments at other positions in the word (see Fig. 3.1.3)[10]. Regardless of whether contrasts at the beginning of word

[10]Interestingly, entropy generally decreases at later positions in word forms (King and Wedel, 2018), which may in part lead to the cross-linguistic development of a greater proportion of contrast-neutralizing phonological processes at later positions in word forms (for more, see Wedel et al. 2019).

forms reaches a significant negative relationship, it is clear to see that the distribution is far *more* balanced at the beginning of words when compared to other positions.

Preparation

If the lexicon is structured to have a more equiprobable distribution at the beginning of word forms, then the entropy at the first position will be greater than the entropy of a lexicon-wide distribution for those same types of contrast. Because I am testing contrasts across different positions, for this test, I will be using individual segments rather than biphones. I do so as the use of a more abstract representation allows for more comparable level of phonological specificity.

To prepare the data, I first removed the segments which comprised the bottom 5% in order to avoid the effects of the Zipfian tail (see above for more description) and I summed the frequencies for the word forms that began with each segment (Eq. 3.4).

$$\text{freq}_{\text{initial}}(s) = \sum_{w} \sum_{s_1 \in w} \text{freq}(w) \text{ iff } s_1 \in w \tag{3.4}$$

I then divided each by the total frequency of the word forms that remained to create a probability distribution for word-initial segments (Eq. 3.5). Word forms that began with removed segments were not included in the total frequency of the corpus, in order to ensure the probability distribution would

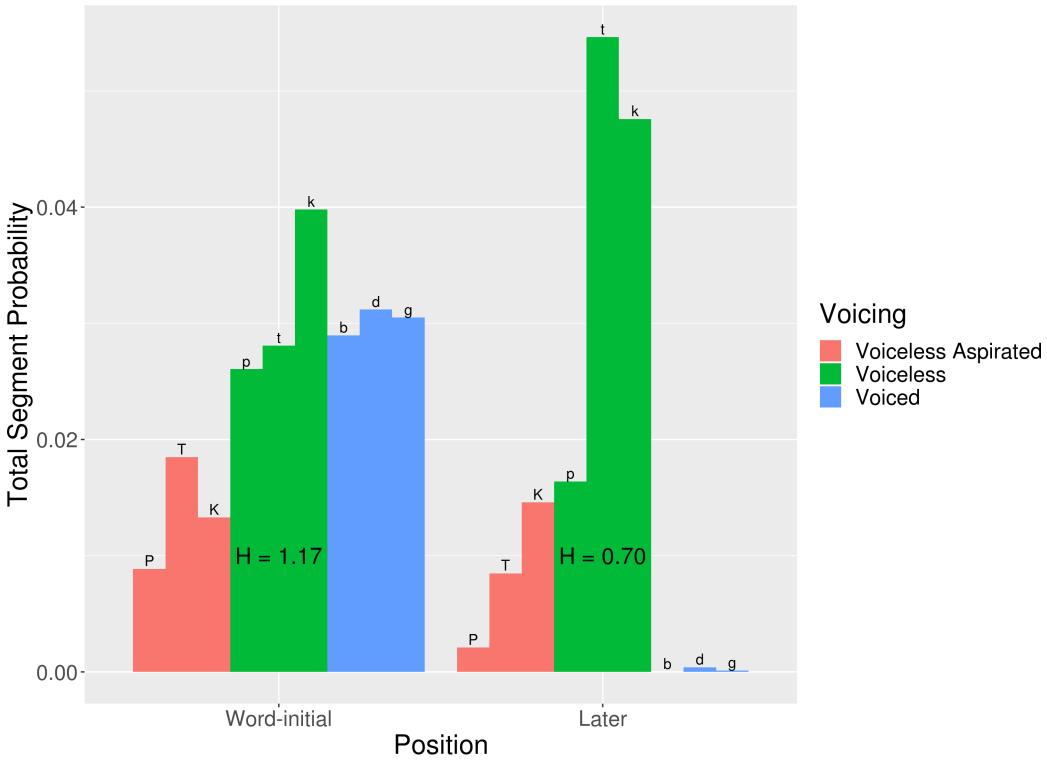

Figure 3.18: Distribution for total probability of Armenian stop consonants. The group on the left represents the total probability of words that begin with each stop segment, the group on the right shows total probability of words that contain the stops at any position after first. The H values represent the overall entropy for stops at word-initial and non-initial positions. Overall, the word-initial stops are much more balanced and as a result have higher entropy.

sum to 1.

$$p(s) = \frac{\text{freq}(s)}{\sum_{s'} \text{freq}(s')} \tag{3.5}$$

I determined entropy as the sum of the expected information of each segment at word-initial position (Eq. 3.6).

$$H(s) = \sum_{s} p(s) * -\log_2 p(s) \tag{3.6}$$

For the segments that remained, I constructed a similar probability distribution based off of frequencies of segments at later positions in words. In this case, I grouped all segments that fell after the first position of a word form to build a baseline count for each of the segments in the analysis (Eq. 3.7). I then used Eqs. 3.5 and 3.6 to determine a baseline entropy for these segments at later positions.

I chose to maintain the same segments in the word-initial and non-initial distributions to ensure that the total number of possible contrasts to determine entropy was constant for both; the entropy of a variable strongly scales with number of distinct values. For example, I did not include [ŋ] in the calculation of non-initial entropy for English because [ŋ] was not included in the set of word-initial segments.

If a segment occurred multiple times in a word form, it was counted multiple times, e.g., the frequency of *habitat* would contribute twice to the relevant count for [t] in the distribution for non-initial segments.

$$\text{freq}_{\text{non-initial}}(s) = \sum_{w} \sum_{s_i \in w} \text{freq}(w) \text{ iff } s_i \in w \text{ and } i > 1 \qquad (3.7)$$

However, I did not compare entropy values directly. When comparing rough, aggregate values of a complex variable as I am doing here, it is possible that results may reflect idiosyncrasies of the corpus rather than the language as a whole. For example, the results for English may be very different, whether the corpus includes names like *Jean* or *Jacques*, i.e., word-initial [ʒ], or not. To some extent, removing the lowest 5% of the probability mass mitigates this, but to be more conservative, I created 1,000 variants of each lexicon to create a distribution for both word-initial and non-initial entropy. To make each variant lexicon, I randomly removed 10% of word types and re-calculated entropy values with this new, smaller version of the corpus. This created novel variants of each language that contained a random set of 90% of the original words. In many respects, this was similar to 'bootstrapping' data, but I opted for this method over the standard practice of sampling with replacement in order to avoid having multiple copies of the same word in the bootstrapped samples. In non-lexical datasets, having identical data points is acceptable, while in this analysis it seemed problematic.

Note that this method does, in a way, create novel forms of the lexicon, but in very different ways from other lexicon generating strategies. I am using this method rather than probability-shuffling, for example, as I am interested in testing whether a lexical property is robust in a lexicon, not

whether said property is *stronger* than in a non-linguistic code. That is, I am not claiming that a larger difference between entropy at word-initial and later positions is evidence of lexical structuring for efficient communication. Rather, I am simply claiming that some difference should exist, and should exist independent of the effect of a single word form in that lexicon.

3.1.4 Results - Word-initial vs. Non-initial Segment Entropy

The previous tests showed a general trend for word-initial contrasts to be balanced to some degree, though there were several languages for which the relationship failed to be significant. A possible reason that the failure to reach the requisite threshold for significance in some languages is that the language-wide distribution of contrasts is itself unbalanced to a degree that word-initial contrasts may be relatively balanced for the language, but still not to a degree where they satisfy the requirements for a linear relationship.

To compare the entropy of word-initial segments to that of the same segments at other positions in word forms, I performed a two-tailed t-test between the distribution of entropy values at word-initial position and those at non-initial position. I found that entropy was significantly greater in all languages except Cantonese (see Figs. 3.19, 3.20 - 3.24).

It is important to point out that all but one of the languages which failed to reach significance in previous tests (i.e., Cantonese) were now found to

significantly show the predicted relationship. This suggests that even if these languages fail to demonstrate a significant relationship, it may be because their lexicon-wide distribution of contrasts is so unbalanced that it would be infeasible for a single contrasting position to show effects of a trade-off between token and type frequency.

A. Fixed Effects:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-161.595	3.638	-44.415	**0.001**
H	43.964	0.533	82.509	**0.001**

B. Random Effects:

	Name	Variance	Std.Dev.
Language:Family	(Intercept)	43.47	6.593
Family	(Intercept)	93.15	9.651

Table 3.3: Logistic mixed-effects model to predict contrast position (non-initial or word-initial), given the entropy, H, of the contrast. Model fit with random intercepts per language, nested within language family. The model showed a significant, positive effect of H, via two-tailed t-tests, indicating that word-initial contrasts have higher entropy per language.

To test the effect across the dataset, I constructed a logistic mixed-effects model with random intercepts for language nested within family. The dependent variable for this model was the position (non-initial or word-initial position) with entropy, H, as a fixed effect (Tab. 3.3)[11]. The model showed a significant, positive effect of H indicating that across the data as a whole, greater entropy correlated with word-initial position. This indicated that, as predicted, the probability distribution for word-initial contrasts was significantly more balanced than a lexicon-wide baseline for the same segments.

[11]R formula: glmer(data = all_langs, position ∼ H + (1|family/language), family = binomial)

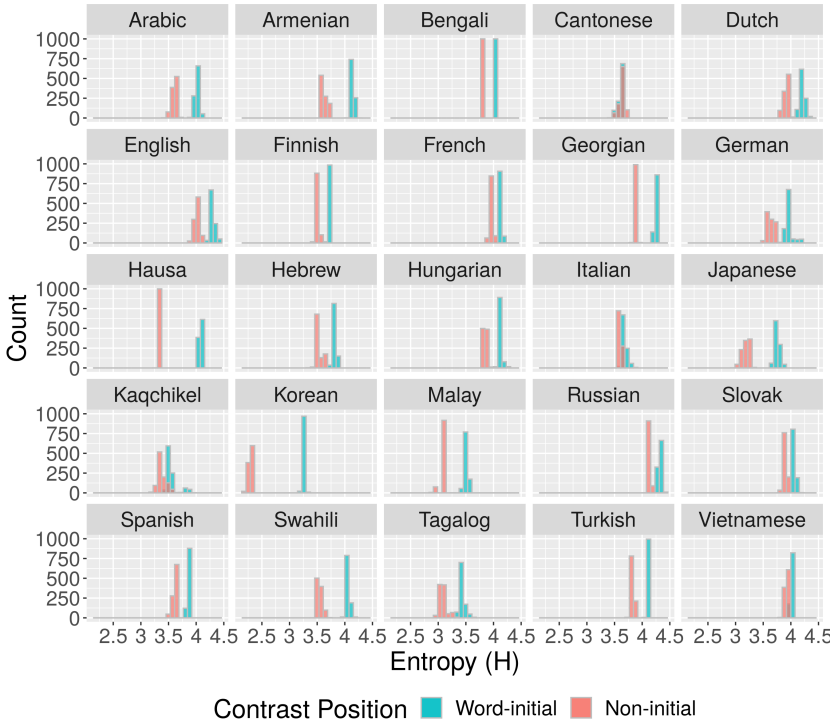

Figure 3.19: Histograms of entropy (H) for word-initial (blue) and non-initial (red) segments for 1,000 random samples of 90% of words from each real-world lexicon in the dataset. In the center of each graph is the average difference in entropy between word-initial and non-initial positions and significance, judged from a paired, two-tailed t-test.

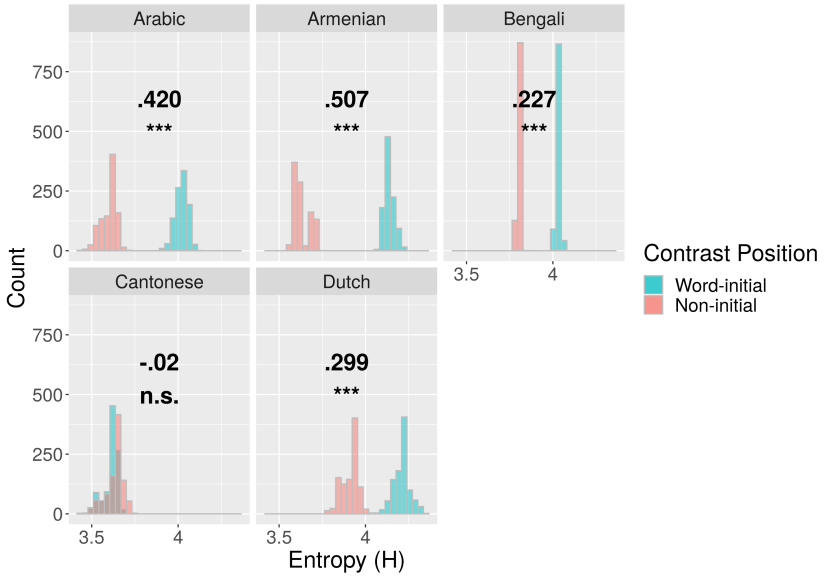

Figure 3.20: Histograms of H for word-initial and non-initial segments. Significant in Arabic, Armenian, Bengali and Dutch. Not significant in Cantonese.

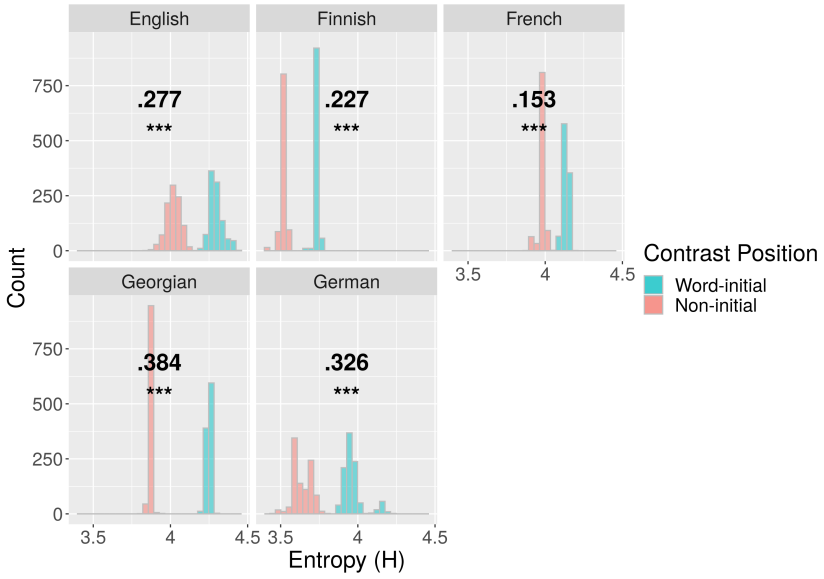

Figure 3.21: Histograms of H for word-initial and non-initial segments. Significant in all languages shown here.

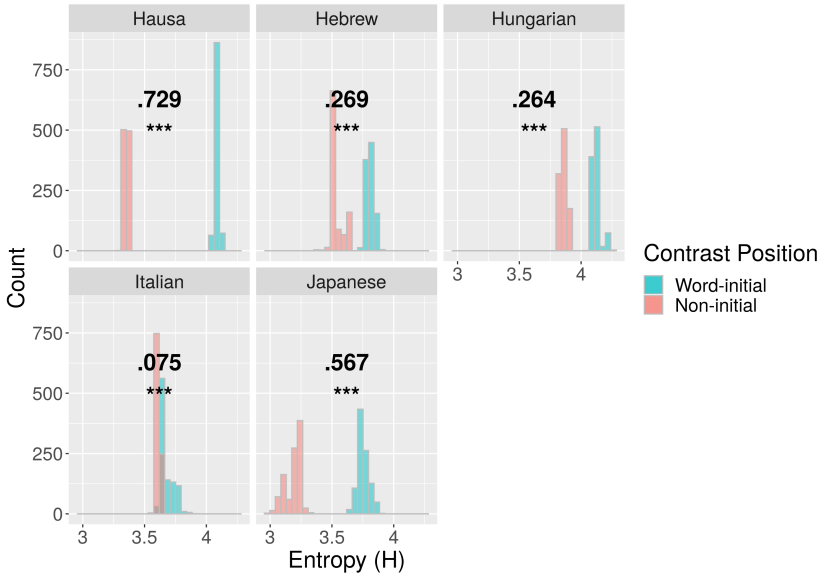

Figure 3.22: Histograms of H for word-initial and non-initial segments. Significant in all languages shown here.

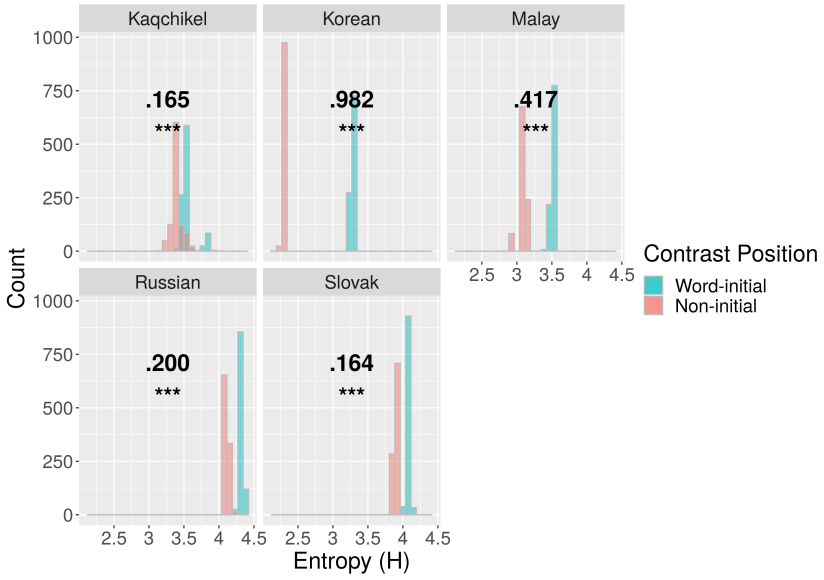

Figure 3.23: Histograms of H for word-initial and non-initial segments. Significant in all languages shown here.

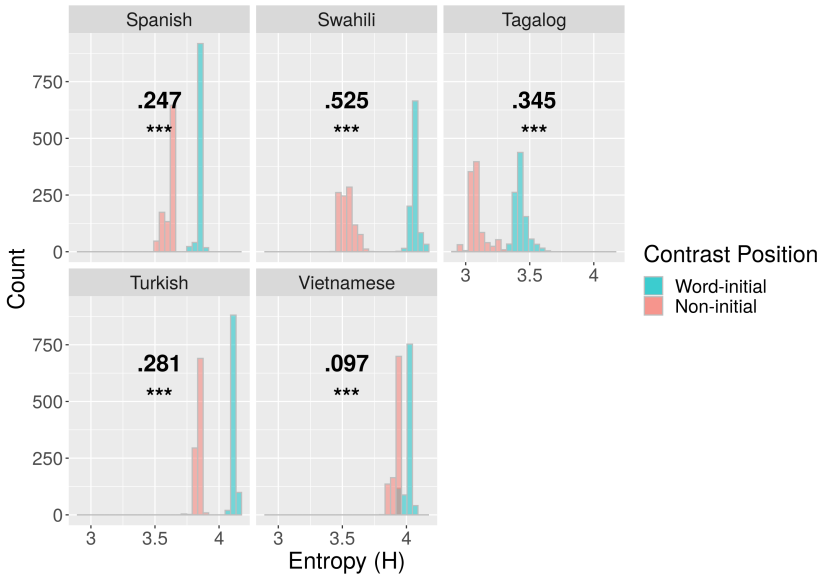

Figure 3.24: Histograms of H for word-initial and non-initial segments. Significant in all languages shown here.

3.1.5 Discussion

The results of this section showed that the distribution of contrasts at word-initial position showed evidence of probabilistic balance, both when tested using standard significance tests, when compared to more language-like baselines and when compared to a baseline distribution for other positions in the lexicon's word forms. As such, this suggests that lexical contrasts are structured in similar ways to an optimally efficient code which is a key part of building an efficient communication system.

Though the overall results supported my hypothesis, there were many languages that failed to show significant properties of balancing when tested for a linear relationship between type and token frequency for word-initial biphones. Within this section, I continued by comparing the distribution of word-initial contrasts to a baseline of the same contrasts at other positions, which showed a significant difference for all languages except Cantonese.

Note that for many of the novel lexicons, there remained a negative correlation between type- and token-frequency for word-initial biphones. Recall that if there were no pressure for probabilistic balance, this correlation should instead be positive. This is likely a result of the closeness of the novel lexicons to the original. When creating novel word forms via a phonotactic model, it is possible to generate word forms that more and less similar to the original word forms. In this case, I chose to use segment trigrams, which resulted in word forms that were remarkably similar to those in the real-world lexicons. In fact, more than 99% of novel lexicons contained at least 10% of word forms

that were identical to forms in the real-world lexicons. Because of this, the weak negative relationships of the novel lexicons are likely residual effects of balancing in the original lexicon. Nevertheless, the majority of languages did show the predicted effects in the later tests of word-initial, suggesting that the reason for failing to show a significant difference in this test may stem from the nature of the novel lexicons themselves.

Nevertheless, the results for the tests of word-initial entropy were also non-uniform. Looking more closely into this, the degree of difference between word-initial and non-initial contrasts varied between languages. If the reason that certain languages failed to show a linear relationship was that their baseline distribution was relatively more skewed than languages that showed a linear relationship, these languages should show a larger difference between the entropy of word-initial and non-initial contrasts[12]. For example, recall that Armenian possess a 3-way voicing contrast for stops at word-initial position and this 3-way contrast is all but lost at non-initial positions due to various phonological processes within the language. Because of this, the word-initial system of stops will likely be more balanced a priori when compared to later positions, even if it fails to significantly demonstrate the desired inverse relationship between type and token.

To test this, I first calculated the difference between word-initial entropy

[12]The mathematical reason for this has to do with the logarithmic function of information. To put it simply, because entropy is a sum of log-transformed values that all sum to 1, it becomes more difficult to increase entropy when entropy is already near maximum. On the other hand, when the entropy of a variable is sufficiently below its maximal value, i.e., when the distribution is skewed, small changes towards balance show large increases.

and non-initial entropy for each of the 1,000 'bootstrapped' variants of each lexicon. I then split the languages in the dataset into two groups: those that showed a significant effect in the linear tests against novel lexicons (14 of 25) and those that did not (11 of 25). I fit a logistic mixed-effects model to predict which group a language fell in (whether it demonstrated a significant, negative relationship in the previous tests), given the difference between word-initial and non-initial entropy (see Tab. 3.4)[13].

A. Fixed Effects:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	29.124	4.297	6.777	**0.001**
Initial/Non-initial Diff.	103.390	0.002	49680.736	**0.001**

B. Random Effects:

	Name	Variance	Std.Dev.
Language:Family	(Intercept)	97.85	9.891
Family	(Intercept)	1835.	42.84

Table 3.4: Logistic mixed-effects model to predict whether the language reached significance in linear tests, given the difference between word-initial and non-initial entropy for each of the 1,000 tests per language. The model found a significant, positive effect of entropy difference indicating that languages that languages that did not show a significant linear relationship had a greater difference, on average, between word-initial and non-initial entropy.

The model showed a significant, positive effect of entropy difference, indicating that languages which failed to reach significance in the linear tests had a greater difference between word-initial and non-initial contrast balancing. This suggests that the reason these languages failed to reach significance was an overall skew towards imbalance across word forms.

[13]R formula: glmer(data = all_langs, not.sig.in.prev $\sim$ diff +(1|family/language)

In any case, the results found here are relatively straightforward, with there being ample evidence of the lexicon being organized for balanced contrasts. Though I limited my investigations to word-initial positions, I do not predict that more or less balanced contrasts should be restricted to this position. Rather, I chose word beginnings because I felt it to be the most likely place to find probabilistic balance and because it was a position that was easily compared cross-linguistically. Nevertheless, I predict that a trend for balance can be found across word forms and I leave it to future work to investigate the degree of probabilistic balance at other positions in the lexicon and how it might vary from language to language.

3.2 Entropy across word forms

Moving past word beginnings, are the other contrast positions of word forms also structured to be part of a high entropy code? Clearly, there is a benefit to high entropy, though there are many competing factors that affect the shape of the lexicon (for discussion, see Köhler 1987). As a result, it is unlikely that high contrast entropy is the only driving factor in lexical structuring. For example, frequently repeated articulatory gestures are less effortful, all things being equal (e.g., Vitevitch et al. 2004; Tomaschek et al. 2018), and there is evidence that lexicons repeat material more than would be expected otherwise, showing that effort reduction plays a role in the structure of the lexicon, beyond the effect of length (Dautriche, 2015; Dautriche et al., 2017;

Meylan and Griffiths, 2017; Mahowald et al., 2018). Thus, because there are multiple pressures affecting the evolution of the lexicon, it is unlikely that all contrasts are probabilistically balanced, though there is likely to be a gradient effect, particularly for contrasts that contribute the most to the overall communicative efficiency of the lexicon as a whole.

Given the competitive and incremental nature of word processing, some positions in the word are better loci for high entropy contrasts. This predicts that certain contrasts in word forms should possess higher entropy than others. For example, Piantadosi et al. (2009) found that stressed syllables posses higher entropy than unstressed syllables, which they attribute to the greater likelihood that these syllables will be be understood by listeners correctly. In this respect, positioning high entropy contrasts in stressed syllables increases the likelihood of the word as a whole being understood correctly, increasing the overall efficiency of the lexicon, while not significantly increasing effort[14].

Because processing is incremental, the information of early material may be incorporated in the processing of subsequent material which altogether allows for more efficient processing overall (Hawkins, 2004; Levy, 2008; Gibson et al., 2019). Interestingly, this fact appears to affect how speakers organize sentences and larger discourses, structuring messages to maximize the predictive power of early material (Genzel and Charniak, 2002; Hale, 2003, 2006; Christiansen and Chater, 2016; Ferrer-i Cancho, 2017). Therefore, if

[14]Stress does correlate with greater production effort (e.g., Morton and Jassem 1965), though by positioning the additional effort to high information material, the impact of efficiency is lessened. For more discussion, see Hall et al. (2016).

the lexicon itself is structured to efficiently provide listeners with the information required to identify words, the earliest material should contain relatively more information, on average (for discussion, see King and Wedel 2020).

As such, it appears that there is ample evidence of early material begin generally more informative. Harris (1955, 1970) showed evidence that the greatest number of contrasts in a language occur at the beginnings of words, suggesting this observation could be used as a cross-linguistic method to parse and differentiate words in a language where little is known of the grammar. Strauss and Magnuson (2008) found that there is a greater set of competitor words that vary by a single dimension (i.e., neighborhood size) for the first syllables of words, compared to later positions, again showing a trend for greater entropy early in word forms. Treiman and Kessler (1995) found a greater distribution of initial clusters compared to final clusters in English CVC words and Hayes and Wilson (2008) showed that the entropy for word-initial contrasts is greater than those of later positions. Most directly, King and Wedel (2018) show that the least probable words possess higher information segments, with this pattern being strongest at early positions.

In the psycholinguistic literature, word-beginnings are often granted a special status and word-initial material was fundamental in many of the earliest models of word recognition (e.g., Forster 1976; Marslen-Wilson and Welsh 1978; Marslen-Wilson 1987; McClelland and Elman 1986; Norris 1994; Norris and McQueen 2008. Moreover, Nooteboom (1981); Salasoo and Pisoni (1985); Connine et al. (1993), among others, demonstrated that these effects

go beyond lexical distribution and participants in experiments are, in fact, more apt at identifying words when given only early material compared to only final material of the words, even when the size and total information of sub-word material was held constant. At a phonological level, Hoolihan (1975) and Wedel et al. (2019) found that contrast-neutralizing and potential recognition-effecting phonology is more likely to occur at the end of words, compared to the beginnings. Also, Steriade (1994) theorized that certain harmony processes, such as vowel rounding, were more likely to spread towards the beginning of the word than towards the end of the word, arguing that spreading towards the beginning provides listeners with early cues for hard-to-perceive late contrasts. In sum, there seems to be ample evidence that lexicons allocate more informative material to the earliest parts of words and that the human perceptual system is in some way keyed into this.

Contrast entropy beyond position

At a glance, this seems to support the general hypothesis until this point, in part. Namely, that the sets of word forms that constitute the lexicon are arranged such that the average information of a given contrast is high. This was the motivation for the focus on the probalistic balance for biphones at word-initial position in the previous segment. That being said, word-initial position is by no means the only part of word forms where a pressure for high entropy might display itself. This leads to a simple question: is the greater entropy of word beginnings an arbitrary fact of language per se or is it part

of a greater pressure towards a maximally efficient code, given the strategies of human language processing?

Of course, there is a benefit to early informative cues, as information presented early in processing can aid in the processing of later material. However, this does not necessarily entail that the decrease in contrast entropy at later positions need be monotonic and uniform across the lexicon. As the number of word types that match the given phonological material, or *cohort* (e.g., Marslen-Wilson and Welsh 1978), shrinks as more of the word is processed, the entropy of contrasts as a whole is likely to decrease with the decrease in overall competition. Yet, there is almost guaranteed to be some variability in entropy among the various cohorts across at a particular position in word forms, e.g., all contrasts at the second segment, third segment, etc. Are there other factors that might predict the entropy of a cohort?

For example, consider the entropies of the contrasts that follow [kʊ̂n] (e.g., *council*, *county*) and [kən] (e.g., *control*, *condition*) in English, which have a relatively low (450 per million) and high (1300 per million) total probability respectively among words in either cohort. In this case, the contrast that follows [kən] has more information on average than that of [kʊ̂n] (2.3 bits vs .93 bits). Because the word forms which belong to the [kən] cohort are more likely all together, there is a greater likelihood of this contrast contributing to the overall information of a message. Put another way, the probability that a listener will encounter a word in the [kən] cohort is greater than that of [kʊ̂n], all things being equal, meaning that the information

present in the segment that follows [n], regardless of what it is, contributes more to the overall informativeness of the lexicon as a whole.

In other words, what is most important for efficient communication is not that a particular contrast has high entropy, but rather the expected information for a contrast is high. That is, the contrasts that are themselves the most likely, i.e. found in more, high probability words, should possess more entropy on average. A lexicon structured like this is built in such a way as to maximize the overall information across possible messages.

Abstractly, a probabilistically balanced contrast in the lexicon would require some amount of evolutionary *work* over time to achieve. Therefore, it would be expected that parts of the lexicon that themselves *do the most work* would be subject to whatever processes create high entropy contrasts. As an equally abstract but more grounded in efficient communication example, consider a *Huffman code* (for more, see MacKay 2003). Huffman encoding is an algorithm for creating a maximally efficient code ignoring noise for a given set of semantic objects and contrastive symbols, which can be thought of as words and phonemes respectively in modeling linguistic communication. A Huffman code reaches maximal efficiency because each contrasting point is as close to balanced as possible, given other constraints of the code, meaning that the overall entropy of each symbol is maximal. However, if it is necessary for one reason or another to have an unbalanced contrast, it is best to place that contrast in a position of the network where it is least likely to occur.

For example, placing an unbalanced contrast at the beginning of words would be severely disadvantageous to the aggregate communicative efficiency of the language altogether because every work token would be affected. On the other hand, placing an unbalanced contrast in a position where it affects only a few words, such as in a cohort with relatively few, relatively low probability words, would have less of a negative effect. Extracting from these, if the lexicon is structured for efficient transmission of information related to word recognition, there should be a relationship between the total probability of a contrast, or *cohort probability* (Eq. 3.8), and the entropy for that cohort, *cohort entropy* (Eq. 3.9). In other words, the contrasts that themselves *do the most work* are the most informative, on average.

$$p(cohort) = \sum_{w \in cohort} p(w) \tag{3.8}$$

$$H(cohort) = \sum_{s \in cohort} p(s|cohort) * -\log_2 (s|cohort) \tag{3.9}$$

It important to point out that this subsumes the previously described importance of word-initial positions and early material. As cohort probability decreases at later positions, earliest positions are predicted to have greatest entropy, which aligns with the assertions of early material being privileged in word processing. The difference here is that I predict that the variance outside of position to also be a partial product of the lexicon evolving for efficient communication, with the lexicon organizing itself so that word compete

with others that are more or less balanced in probability.

Of course, the size and linear position of a contrast is certain to play a role. All things being equal, larger cohorts where more word types compete are likely to possess a greater entropy. Furthermore, cohort size monotonically shrinks at later positions - once a word has left the cohort, it cannot be regained - meaning that later cohorts are likely to possess lower entropy, independent of other factors. If the relationship between cohort probability and entropy is an aspect of lexical structuring for efficient communication, the effect of cohort probability should be independent of both cohort size and position in the word form.

3.2.1 Preparation

If the lexicon is structured to be an overall high information code on average, then the positions that are themselves most likely should have greater entropy. Explicitly, there should be a relationship between the probability of a contrast and its entropy, between *cohort probability* and *cohort entropy* respectively.

Recall that a cohort (e.g., Marslen-Wilson and Welsh 1978) is the set of word forms that share phonological material until a point. For example, *contamination*, *constant* and *cat* all belong to the [k] cohort because they share the initial segment [k]. On the other hand, *contamination* and *constant* both belong to the [kə] and [kən] cohorts as well, though *cat* /kæt/ does not. Here, I use the term *cohort probability* to represent the probability that a

given contrast will be relevant in word processing, e.g., the probability that the contrast that follows [kən] occurs across the tokens of a corpus. To calculate this, I summed the individual word probabilities of all words that belonged to that cohort (Eq. 3.10).

$$p(cohort) = \sum_{w \in cohort} p(w) \tag{3.10}$$

I use the term *cohort entropy* to represent how much information that the possible *continuing segments* for a cohort contain, on average. Here, I use the term *continuing segment* to mean the set of phonological segments that can immediately continue to create any of the word forms of the language. For example, in the [kən] cohort, [t] and [s] are among the the set of continuing segments because words like *contamination* and *constant* exist in my English corpus, while [h] is not because no word in the English corpus begins with [kənh].

To determine cohort entropy, I first calculated the conditional probability of each possible continuing segment given the current cohort, e.g., the probability of [s] after [kən], by finding the total frequency for all words that began with the segment in question, divided by the total frequency of all other possible next segments (Eq. 3.11).

$$p(s_i|cohort) = \frac{p(cohort \wedge s_i)}{p(cohort)} = \frac{freq(s_1 ... s_i)}{freq(s_1 ... s_{i-1})} \tag{3.11}$$

To be clear, it is possible to measure segment probability with both a

token-based and type-based measure. I opted for a token-based measure here (e.g., van Son and Pols 2003; Cohen Priva 2017; King and Wedel 2020) because it can capture a relationship between the total token frequency and type counts for continuing segments in a way a type-based measure could not. That is, type-based segment information (c.f. Meylan and Griffiths 2017) measures the number of word-types that continue with a given segment compared to other possible word-type continuations. A key component of the investigation here is the relationship between word probability and type counts of segments, meaning that means to compare the relative probability between words in cohorts, e.g., token frequency, is needed.

To calculate cohort entropy itself, I found the expected information across all possible continuing segments for that cohort (Eq. 3.12). As I did in the previous section, I removed segments that together comprised the bottom 5% of the total probability mass per cohort, to mitigate the effect of Zipfian distributions of conditional segment probabilities.

$$H(cohort) = \sum_{s_i \in cohort} p(s_i|cohort) * -\log_2 p(s_i|cohort) \tag{3.12}$$

With a measure for both cohort probability and entropy in hand, it was now possible to investigate a relationship between the two. However, the nature of the data is slightly more complicated, meaning that certain cohorts should be excluded from the analysis while additional control variables should be included.

Cohort Selection

Given the definition of a cohort, a cohort can range in size from all words in the lexicon, i.e., word beginnings, to a single word form. For example, following a word form's *uniqueness point* (e.g., Marslen-Wilson and Welsh 1978), or segment where the form becomes unique in the lexicon judged from the beginning of the form, the size of the current cohort is exactly one. As might be expected by this definition, a proportionally large section of all cohorts in the lexicon consist of a single word, e.g., roughly 76% of all cohorts in English. Because these cohorts possess a single word, they often possess relatively low probabilities (recall that in a Zipfian distribution, a majority of words possess very low probabilities). Furthermore, because these cohorts only consist of a single continuing segment, by definition they possess an entropy of 0.

Together, the fact that a large proportion of the total cohort structure consists of these low probability, '0 entropy' values could inadvertently create an uninteresting statistical relationship. To avoid this possibility, I removed all cohorts that fell after a word form's uniqueness point from the analysis. For instance, I removed cohorts such as [evieʃ], [evieʃɪ], [evieʃɪn] which all contain only a single word, *aviation* (according to my English corpus). That said, the cohort [evie] is still included in the analysis because it contains two words, *aviation* and *aviator*. Note that doing this did not totally exclude '0 entropy' cohorts altogether; there are cases where two or more word types remain in a cohort but there is only a single segment, e.g., *dormitory* and

dormant both belong to the [doɹ] cohort but for this cohort, there is a single continuing segment, [m]. However, it greatly reduced the chance that any discovered effect would be solely driven by single-word cohorts.

In addition, the distribution of cohort probability - like many linguistic probability distributions - has a long Zipfian tail, due in part to many cohorts that share several segments possessing very few word types. Again, because of the Zipfian nature of word probabilities, many of these very low probability cohorts consisted of just two word types, that themselves had word probabilities that barely made the cut off threshold (1 per million) during corpus preparation. This, by itself, creates an artifact where the least probable cohort possible (2 per million) in the data consisted of two words which were of equal probability. To avoid this, I removed the least probable 5% of remaining cohorts from the analysis, leaving the top 95%. Doing so, I was able to focus the tests on a vast majority of contrasts, while avoiding the possible danger that thresholds used in the preparation of the data might affect the results.

Note that these two controls do overlap but are not necessarily redundant. It is possible that some single-word cohorts (especially in smaller corpora) may still find themselves in the top 95% of total cohort probability. Because of this, it was necessary to use both measures of pruning the data for the analysis.

Control Variables

Generally, the entropy of contrasts decreases at later positions in words (King and Wedel, 2018), primarily because cohorts can only lose words as position increases. To account for this, I collected the *cohort position*, the position where the contrast occurs in word forms for each cohort in the analysis. Because cohort position was one of multiple independent variables, I chose to include it as a separate factors in linear models in tests.

As an additional control variable, I collected the total number of word types for each cohort. Generally, a cohort with more word types should have a larger entropy, all things being equal, as there are likely to be a greater number of possible continuing segments. Crucially, because cohort probability is correlated with the number of word types within the cohort, any relationship between the cohort entropy and probability might be reporting the simple fact that smaller cohorts have lower entropy.

An astute reader may ask why I did not include the number of continuing segments itself as a control variable. I did not include the number of continuing segments for a cohort because of its overly strong correlation with both word probability and cohort count. In point of fact, the correlation between *log* cohort probability and number of segments was far stronger than with cohort count (see Fig. 3.25) and, for each language, the two variables had a Pearson's correlations of .46 on average. Moreover, VIF scores in linear models to predict cohort entropy, given *log* cohort probability and number of segments, were above a comfortable threshold (4+ for many languages)

meaning that both factors share much of their explanatory power for the dependent variable (for more on VIF, see O'brien 2007).

Together, these controls suggest that if all three variables were included as separate independent variables in a regression model, the multi-collinearity could possibly interfere with the interpretation of the model, e.g., *suppression effects* (Baayen, 2008)[15]. To avoid the dangers of multi-collinearity, I chose to not include number of possible segments in a cohort as a control variable.

Furthermore, it is not clear to me that the high entropy need be an independent effect concerning number of contrasting segments. All in all, the hypothesis of this section is that word forms within the lexicon are structured such that the most informative contrasts overall occur with the greatest likelihood in use. Entropy can increase from either an increased number of possible values and from a more balanced distribution for those values. It could be the case that a relatively large set of contrasting segments for a cohort is an early stage of that cohort's evolution towards greater entropy. Here, I am not making any claims with respect to which strategy is likely to be employed and for these reasons, I chose to not include the number of continuing segments as a control.

Before continuing, I would like to point out that this is but a single of

[15]This is also the reason I chose to use raw cohort count in lieu of a *log* transformed version. When *log* transformed, the strong correlation with *log* cohort probability made interpreting models difficult. That said, the correlation between cohort size and entropy was not greatly affected by not log-transforming the variable. Though imperfect, including raw cohort count allowed for more conservative tests, while not introducing collinearity problems.

Figure 3.25: Correlations between *log* cohort probability and cohort size (red) number of continuing segments (blue). For all languages, the correlation between cohort probability and number of segments is greater than with cohort size (in addition to a very strong correlation with cohort entropy and size themselves), meaning that including number of segments as a factor increases the chance of multi-collinearity affecting the results.

many ways to deal with the complexities latent in the data, though it was one that best accounted for complexity between variables, while not excluding a large proportion of the total dataset and not adding to the overall complexity of the analysis. More sophisticated controls will allow for clearer results in the future. Even so, by including these control variables, my results are more conservative than those that would emerge from simpler tests that did not include these controls in any form. Furthermore, because I will use the same

control variables in constructing language-specific baselines (see below), if the lexicon does show the predicted effects, it is not likely to be a result of the inclusion/lack/treatment of a particular control.

Comparison to Novel Lexicons

In addition to standard significance tests, I will also be comparing each real-world lexicon against a distribution of 1,000 novel lexicons, made from word forms that were generated from a phonotactic model from each language (for more, see Chapter 2.0.2). I chose to create novel lexicons via phonotactic models because this method varies the both the number of word types in each cohort and token-based segment probabilities, whereas probability-shuffling only varies the distribution of the token-based segment probabilities. Varying both is important since they are equally valid pathways to creating balanced contrasts. To ensure that the fit models for both types of lexicon will be easily comparable, I will re-calculate the same control variables for each novel lexicon.

If the real-world lexicon falls significantly beyond the distribution (>95%) of novel lexicons, it will be evidence that the real-world lexicon is structured to position probabilistically balanced and therefore higher entropy contrasts in more likely positions at a greater degree than might be expected in a structurally similar but non-linguistic code. For simplicity in the results of these tests, I will only report the results for the effect of cohort probability; the effect of the control variables of position and cohort size are not relevant

to the hypothesis.

3.2.2 Results - Significance Tests

To begin, I fit linear models per language to predict cohort entropy, given a) *log* cohort probability, b) cohort size and c) position. As expected, for all languages, there was a significant, negative effect of position and a positive effect of cohort size. This indicates that later cohorts and larger cohort possess greater entropy, all things being equal. Most relevant here, for all languages, there was a significant, positive effect of *log* cohort probability, indicating that more probable cohorts had greater entropy, on average. (see Figs. 3.26, 3.27 - 3.29). This suggests that the lexicon is structured to position the most informative contrasts where they are most likely to contribute information in the average message, a key aspect of an efficient code.

To ensure that the results remained across the data as a whole, I fit a linear-mixed effects model to the entire dataset, with the same fixed factors as the by-language linear models, including random intercepts per language, nested within language family. The model found a significant, positive effect of *log* cohort probability, indicating that the most probable cohorts in the lexicons of the tested languages were also those with the greatest entropy, on average. As expected, the model also showed a positive effect of cohort size and negative effect of position, replicating the results in the models over individual languages (see Tab. 3.5)[16].

[16]R formula: lmer(data = all_cohorts, entropy ∼ log.prob + cohort.count + position +

A. Fixed Effects:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.857	0.044	19.331	**0.001**
log Cohort Prob.	0.210	0.003	79.289	**0.001**
Cohort Size (type)	0.081	0.002	33.226	**0.001**
Cohort Position	-0.149	0.002	-93.604	**0.001**

B. Random Effects:

	Name	Variance	Std.Dev.
Language:Family	(Intercept)	0.004	0.070
Family	(Intercept)	0.012	0.113

Table 3.5: Linear mixed-effects model to predict cohort entropy given *log* cohort probability, cohort size and position, with random intercepts per language nested within family. The model found a positive, significant effect of cohort probability, suggesting that for all languages, more probable contrasts are more informative on average, as predicted.

(1|family/language), REML = F)

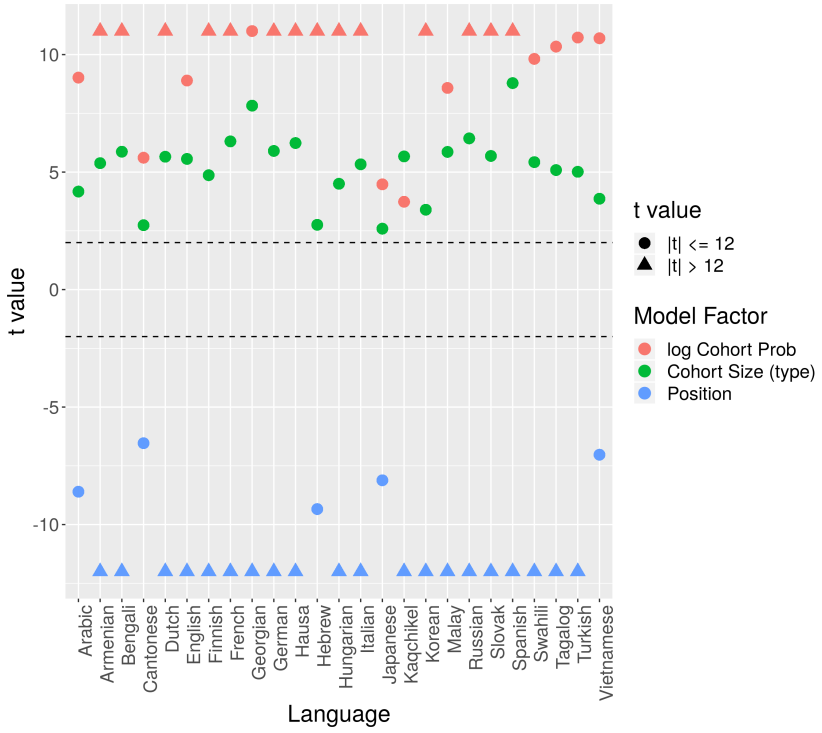

Figure 3.26: Results of linear models to predict cohort entropy for all languages in the dataset. The y-axis shows the t-value of the specified factor. The t-value is a measure of the amount of variance that a particular factor in a linear model explains (for more, see Baayen 2008), with values greater than 2 or less than negative 2 suggesting significance. Extreme values (+/- |12|) have been truncated for visual readability. For all languages shown, there was a significant, positive effect of cohort probability, indicating that more probable contrasts had greater entropy.

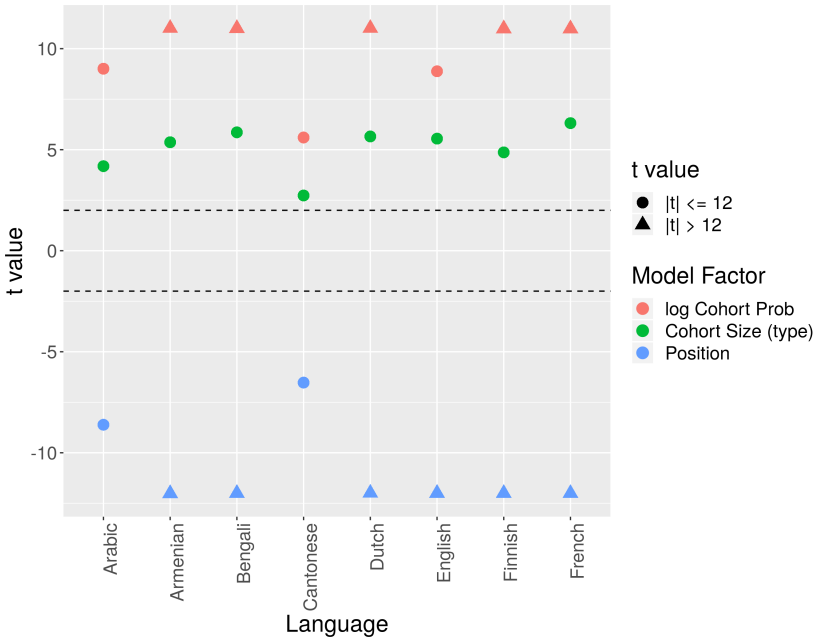

Figure 3.27: Results of linear models to predict cohort entropy. For all languages shown, there was a significant, positive effect of cohort probability, indicating that more probable contrasts had greater entropy.

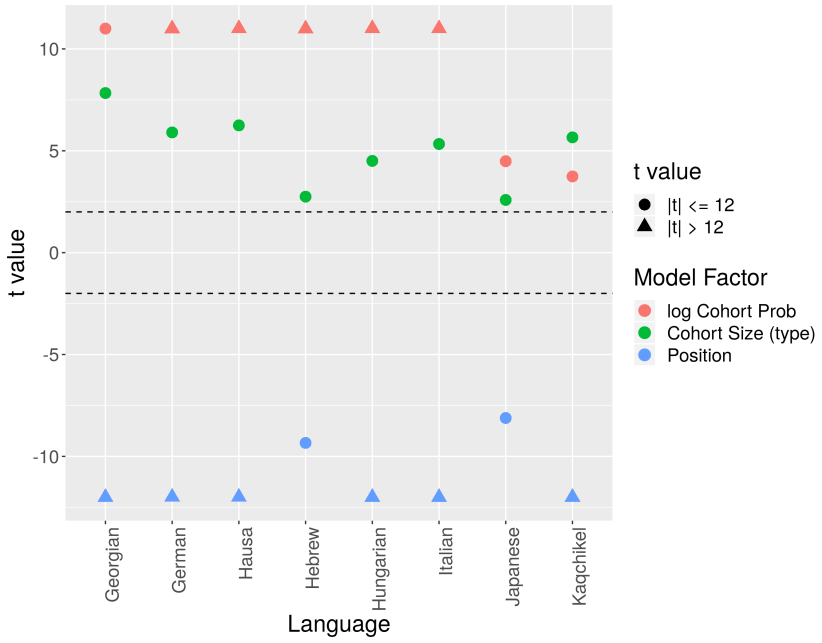

Figure 3.28: Results of linear models to predict cohort entropy. For all languages shown, there was a significant, positive effect of cohort probability.

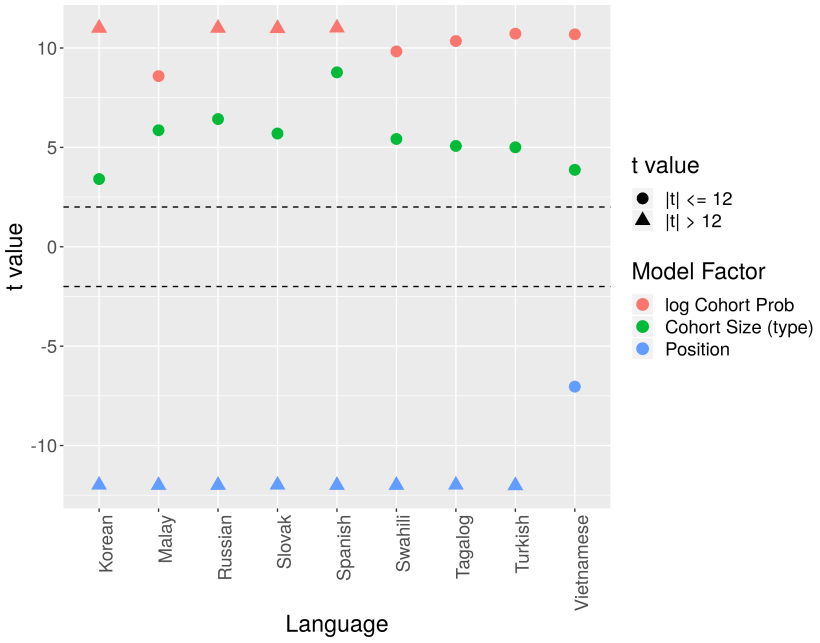

Figure 3.29: Results of linear models to predict cohort entropy. For all languages shown, there was a significant, positive effect of cohort probability.

3.2.3 Results - Novel Lexicons

Thus far, results have suggested that the more probable cohorts also possessed the greatest entropy, controlling for total size and position of the cohort. Yet, this result might arise in any language-like code and its mere presence may not be indicative of lexical structuring for efficient communication. To investigate this, I compared each of the real-world lexicons against a distribution of 1,000 novel lexicons.

For each of the 1,000 novel lexicons per language, I fit a linear model with an identical architecture as the ones in Figs. 3.26, 3.27 - 3.29. I then extracted the estimated effect of *log* cohort probability for the real-world and all novel lexicons. In all languages except Georgian, Hebrew, Russian, Slovak and Spanish the estimate for the real-world lexicon fell above 95% of those of the novel lexicons (Figs. 3.31 - 3.35), suggesting that, as predicted, the correlation between cohort probability and entropy is stronger in these language's lexicon than might be expected otherwise. Note that for Spanish, there appeared to be a significant, negative, indicating that Spanish demonstrates the reverse effect (see discussion).

To test the strength of the effect across the data as a whole, I constructed a logistic mixed-effects model to predict lexicon type, given the model coefficient for *log* word probability, b, from each of the linear models per lexicon, with random intercepts per language nested within language family. The model found a significant, positive effect of b, indicating that though results were not uniform across languages individually, the trend across the dataset

showed the real-world lexicons had a stronger effect of word probability than their respective comparable novel lexicons (see Tab. 3.6)[17]. This aligns with the original hypothesis, suggesting that the lexicon organizes word forms to place high entropy contrasts where they are more likely and that the lexicon does so more than would be expected given the computed baseline.

A. Fixed Effects:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-20.618	2.978	-6.924	**0.001**
b	222.908	34.675	6.428	**0.001**

B. Random Effects:

	Name	Variance	Std.Dev.
Language:Family	(Intercept)	24.95	4.995
Family	(Intercept)	1.594	1.262

Table 3.6: Logistic mixed-effects model for predicting lexicon type (nonce forms vs. real-world), given the coefficient corresponding to *log* word probability in linear models to predict cohort entropy, *b*. The mixed-effects model found a significant, positive effect of *b*, indicating that across the tested languages, the real-world lexicons had relatively stronger effects of cohort probability on predicting cohort entropy, as predicted.

[17]R formula: glmer(data = all_langs, lex.type ∼ b + (1|family/language), family = binomial)

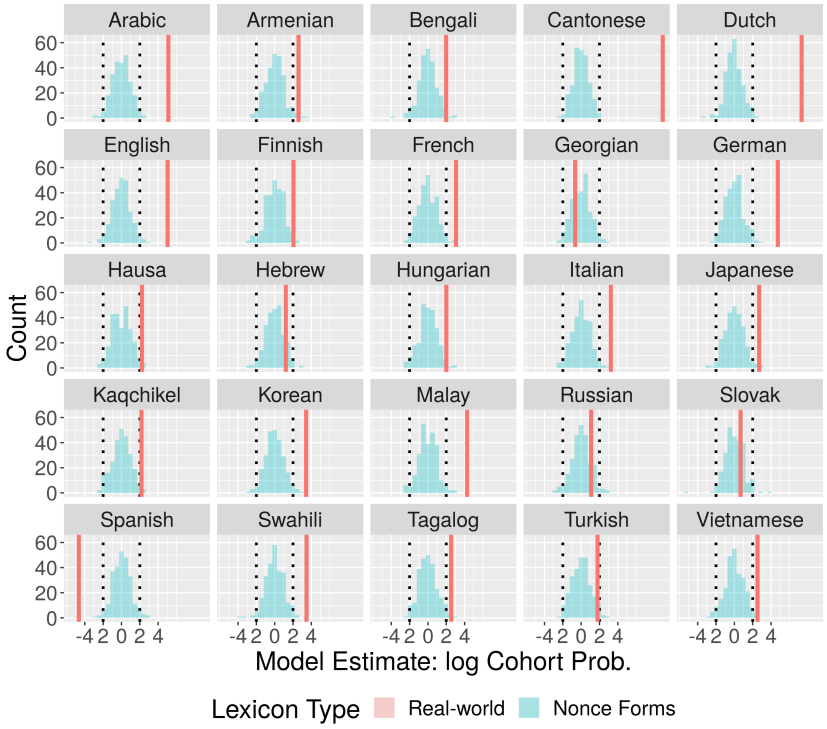

Figure 3.30: Comparison of model coefficients for effect of *log* cohort probability on cohort entropy for all languages in the dataset (cohort size and position included in models, though not shown) for real-world lexicon (red) and 1,000 novel lexicons (blue). The coefficients have been z-scored and the dotted lines indicated +/- 2. If the red line falls outside of the blue distribution, it indicates that the real-world lexicon shows a stronger correlation between cohort probability and entropy than would be expected otherwise.

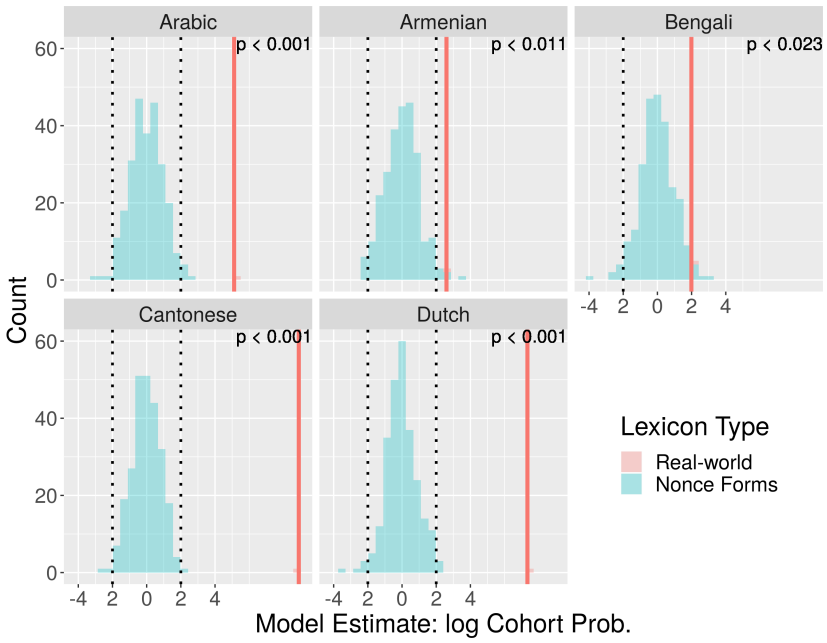

Figure 3.31: Comparison of model coefficients for effect of *log* cohort probability on cohort entropy in the real-world and novel lexicons. For all languages shown here, the effect of cohort probability is significantly greater than in the novel lexicons.

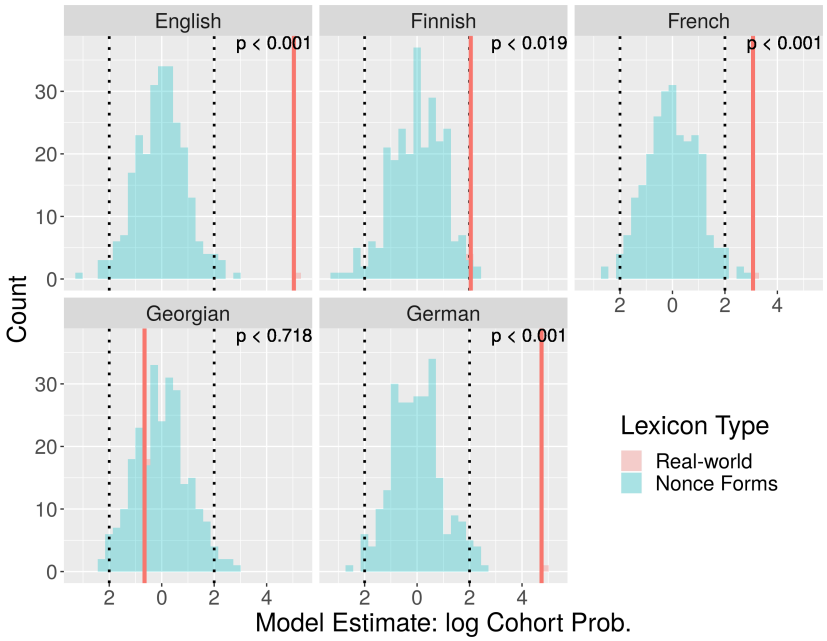

Figure 3.32: Comparison of model coefficients for effect of *log* cohort probability on cohort entropy in the real-world and novel lexicons. For English, Finnish, French and German, the real-world lexicon had a greater effect than 95% of the novel lexicons.

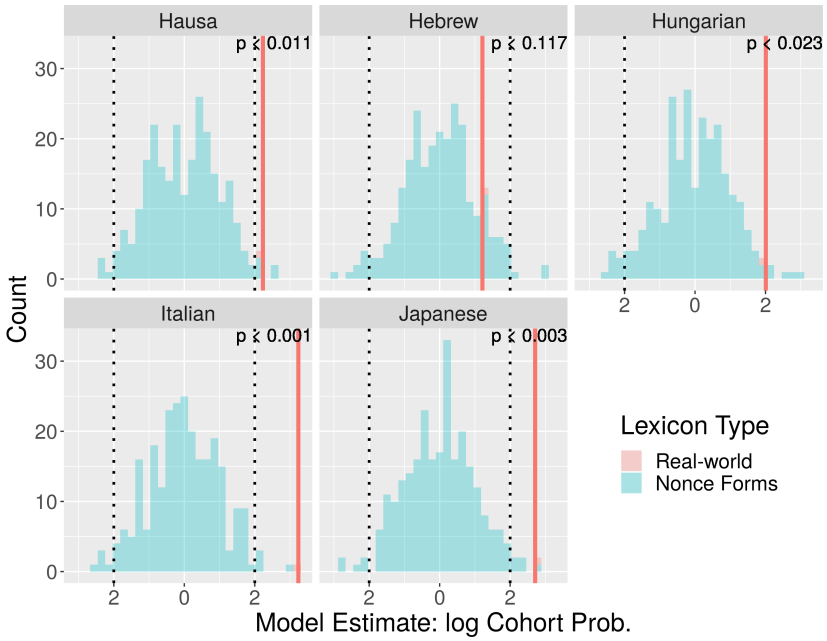

Figure 3.33: Comparison of model coefficients for effect of *log* cohort probability on cohort entropy in the real-world and novel lexicons. For Hausa, Italian, Hungarian and Japanese, the real-world lexicon had a greater effect than 95% of the novel lexicons.

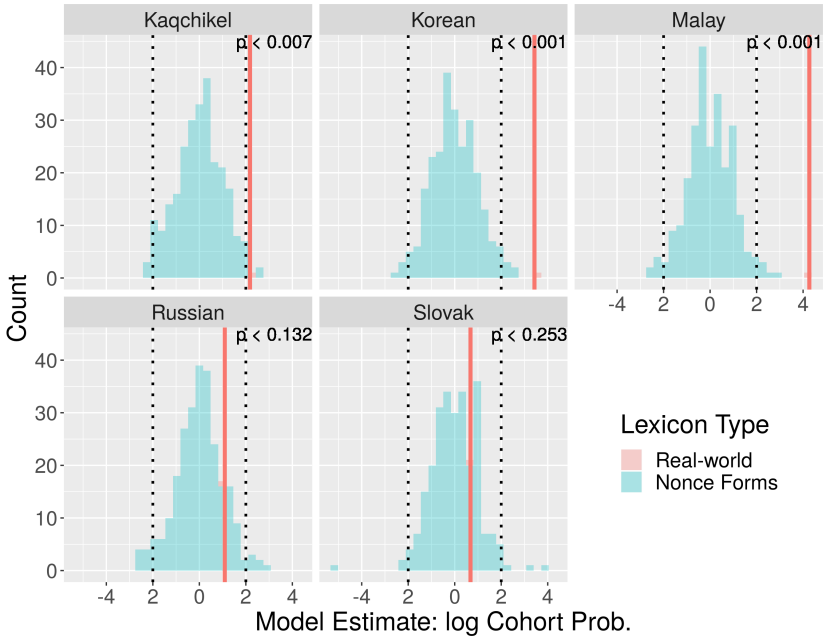

Figure 3.34: Comparison of model coefficients for effect of *log* cohort probability on cohort entropy in the real-world and novel lexicons. For Kaqchikel, Korean and Malay, the real-world lexicon had a greater effect than 95% of the novel lexicons.

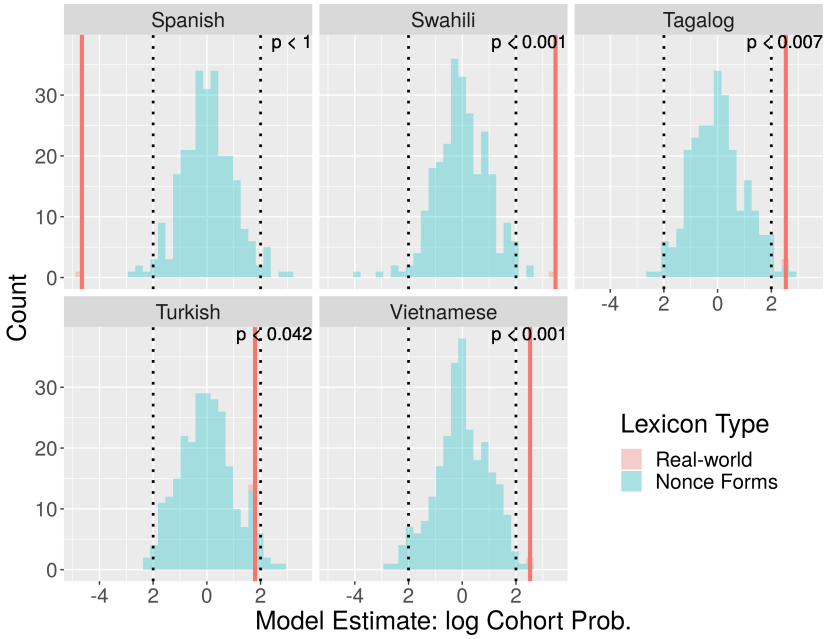

Figure 3.35: Comparison of model coefficients for effect of *log* cohort probability on cohort entropy in the real-world and novel lexicons. For Swahili, Tagalog, Turkish and Vietnamese, the real-world lexicon had a greater effect than 95% of the novel lexicons. For Spanish, the real-world lexicon had lesser effect than the novel lexicons (see Discussion).

3.2.4 Discussion

Across the dataset, I found compelling evidence that the most likely contrasts in the lexicon are also those with the greatest entropy. This would be expected in a code that has been structured for efficient information transmission, assuming that each sub-word part is processed incrementally and relative to earlier material within the word form. Ultimately, the true gage of the efficiency of a communication relates to the the average information conveyed by each part of the message, weighted by how likely that part is. Because the most likely parts of the lexicon were found to also be those of that are most informative on average, this suggests that the specific contrasts that build word forms are organized under pressures of efficient communication.

That said, the results were not universal across the dataset. Particularly, for Georgian, Hebrew, Russian and Slovak there was no effect found beyond what would be expected by structurally similar code. For Spanish, the effect went opposite of the predicted direction significantly. The question now turns to why these languages failed to demonstrate the predicted effects of efficient encoding.

One likely reason for this lies in the structure of the languages themselves. Hebrew is a Semitic language and has a strict root-and-pattern word formation strategy (Ussishkin, 2005). This in part restricts the possible positions of high information contrasts, e.g., the final consonant of the lexical root is often late in the word form whereas the final segment for a lexical root in less restricted language may occur earlier. Russian and Slovak are both Slavic

languages and have a complex set of word-final inflections, varying across dimensions of gender, number, etc. (Dryer and Haspelmath, 2013). Because of this, many permutations of nouns and verbs have condensed into a single form during the creation of the corpus, causing possibly high information contrasts to not appear in the uninflected word stems of the data. Georgian is an agglutinative language, with many word forms consisting of long sequences of distinct morphemes (Aronson, 1990). It could be the case that Georgian word forms are structured with specific, high frequency, morphologically complex forms in mind. These forms would be lost in lemmatization, meaning that they would not occur in the analysis lexicon. Of course, these claims require an in-depth investigation into these languages separately, with more finely detailed corpora.

Be that as it may, even if lemmatization is the cause for those languages to fail to show the predicted effects, it does not explain why Spanish shows a reversed effect. Spanish is no more morphologically complex than French or Italian, and these languages robustly demonstrate the predicted effects.

Upon a closer look at the actual Spanish corpus used here - which was a collection of television transcripts - I found that the words for *president*, *government*, *minister* and a word that is homophonous between *political party* and *sports match* were among the 20 most frequent in the corpus[18]. This suggests that frequency information represented in the corpus may not be

[18]I mention the 20 most frequent terms here not because of a special status for these words, but as an anecdote for a likely lexicon-wide skew of word probabilities.

an accurate representation of Spanish as it is spoken every day. It goes without saying that no corpus is a perfect representation of its language but an obvious skew towards political terms for the most frequency words words may have affected the results seen here.

As a post-hoc test, I re-ran the tests for Spanish using a corpus of one million Spanish Wikipedia pages (Goldhahn et al., 2012). This source was not phonetically transcribed as my original Spanish corpus was. Be that as it may, Spanish orthography is relatively transparent and the lack of phonetic transcription should not play as big a role as it might for English or French.

When compared to a sample of 1,000 novel lexicons[19], the real-world Spanish lexicon demonstrated a significant, positive effect of the cohort probability on cohort entropy, contrasting with the significant, negative effect found in the other Spanish source. This suggests that a likely reason the language failed to show an effect in earlier tests earlier was due to the properties of the specific corpus used[20].

To me, this stands as a sign of the potential influence that using one corpus over another might have on the results of any testing and as a testament to the benefit of using rigorous baselines when testing complex linguistic properties. That is, even the original Spanish source with problematic word

[19]The comparison lexicons here were generated from an n-gram model trained on the Spanish Wikipedia data.

[20]To be sure of results from previous chapters of this dissertation, I re-ran tests using the Spanish Wikipedia data and found near identical results. I left the results from the original Spanish corpus as-is because the inclusion of this secondary source was a post-hoc addition and I did not want to create a pattern of post-hoc substitutions of data sources when certain tests failed.

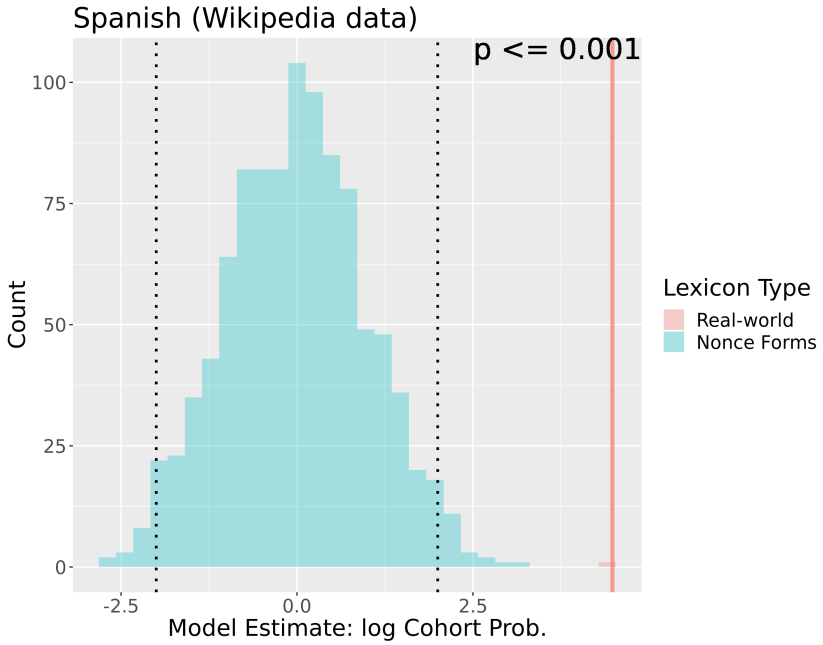

Figure 3.36: Comparison of Spanish word forms taken from Wikipedia and 1,000 novel lexicons generated from those forms. In this case, the real-world Spanish lexicon shows a greater effect of cohort probability on cohort entropy, suggesting the early reverse effect may have been due to idiosyncrasies of the original Spanish corpus, e.g., disproportionately high frequency political terms.

frequency data showed a significant relationship between type and token probabilities. It was only when compared to novel lexicons that the opposite effects emerged. I see this as additional example of the benefit of using language-like baselines, as suggested by Moscoso del Prado Martin (2013), a practice that I hope becomes standard in computational linguistics.

Be that as it may, altogether, the results here supported the original hypothesis of the section: there is a strong pattern among word forms where high entropy contrasts are located in positions where they themselves are

more likely, controlled for the number of competing words and position in the word. This makes the lexicons of the tested languages in line with abstract high entropy codes, such as Huffman codes. Of course, segment entropy is but one means to qualify the overall informativeness of a contrast.

Recall that treating segments as atomic, contrastive symbols ignores the fact that many segments are perceptually more similar than others, meaning that some contrasts are more or less informative, all things being equal. For example, hearing a [t] provides a listener with information to discount all words that do not continue with a [t]. As it is coded here, words that continue with a [d] are equally discounted compared to a [g], which is perceptually more distinct. It could be the case that for the languages that fail to show the predicted relationship between cohort entropy and probability do so because they maximize information at contrasts relative to the actually perceptual similarity among the phonological segments at each contrast.

For example, Georgian has a very dense set of stop consonants, with 3-way stop voicing and four places of articulation (Aronson, 1990)[21]. It may be the case that the relationship between contrast probability and entropy in this language may only arise when the phonological nature of the contrasts are considered with the others in the cohort. If the Georgian lexicon is sensitive to the nature of contrasts, it may show the predicted relationship between cohort probability and entropy when only perceptually similar segments are considered.

[21]The uvular position only includes a single value for voicing.

As another post-hoc test, I re-calculated cohort entropy for the real-world and novel lexicons using only stops, i.e., ignoring vowels, liquids, fricatives and affricates as continuing segments, finding that the real-world Georgian lexicon begins to display a stronger relationship between cohort probability and entropy when only stops are considered (see Fig. 3.37).

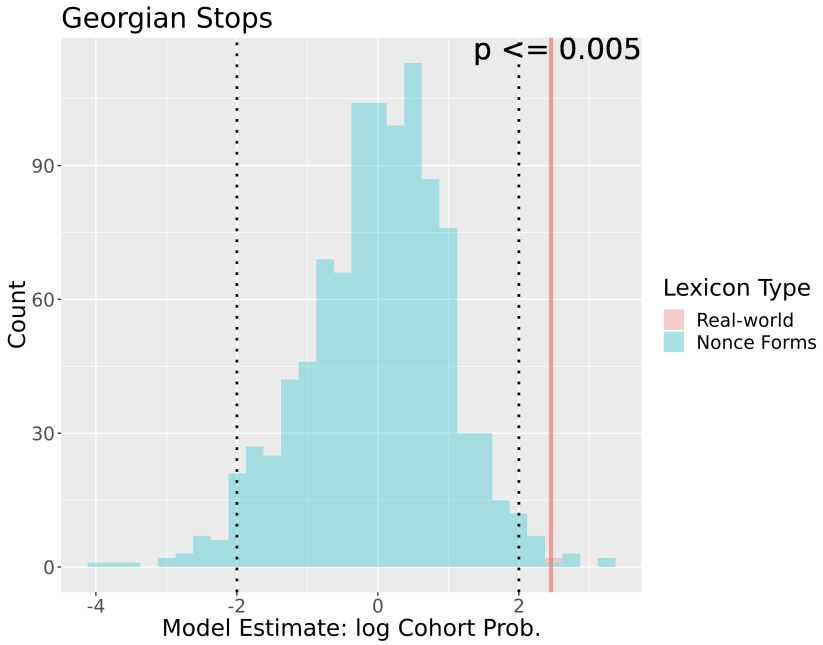

Figure 3.37: Comparison of Georgian stops in the real-world lexicon and 1,000 novel lexicons. In each, cohort entropy and size is considered only for stop consonants. When a smaller set of more perceptually similar segments are considered, Georgian begins to display a significant difference between the real-world lexicon and the baseline, suggesting that the phonological nature of contrasts may play a role in lexical optimization.

This is not to suggest that lexical structuring behaves differently in Georgian than other languages and that optimization applies only to stops, merely I hope to show that considering phonological nature of contrasts may eluci-

date some questions raised here. Regardless, the findings presented here offer compelling evidence that the lexicon is structured for efficient information transmission, given that it is processed incrementally with the information contributed by parts of words relevant to the context of earlier word form material and the set of lexical competitors.

3.3 General Discussion

In this chapter, I investigated the lexicons of a diverse set of languages, showing that word forms are partially arranged so that individual phonological contrasts are a) between probabilistically balanced alternatives and b) that there is a relationship between the average information, i.e., entropy, of a contrast and its likelihood overall. Both of these are properties of efficient communicative codes (Cover and Thomas, 2006), codes that are designed to transmit maximal information over the shortest amount of time.

For both parts of this overarching prediction, I demonstrated the predicted properties, using both standard statistical tests and via comparisons of the tested languages against more realistic baselines, i.e., novel lexicons generated from a phonotactic n-gram models. Together, the results of both families of tests supported the original predictions, indicating that the evidence for efficient lexical structuring of the tested languages is not likely an artifact of the methods or data used here.

Data Issues

At this point, I would like to address the specific metrics employed here, particularly their coarseness. To measure balanced contrasts, I focused on a single position. To measure the the average information for all contrasts, I assumed strict cohorts of competitors. In both cases, I assumed that contrasts are between equally perceptually dissimilar atomic symbols, and not between phonological units that have a systematic perceptual similarities (e.g., Flemming 2004; Mielke 2012). I made these assumptions to reduce the requirements for the level of processing of the data and extend across multiple languages and levels of data quality.

A more sophisticated metric for segment information may elucidate some of my results further. For example, I assume that as soon as a word differs from the current cohort, it is excluded completely, e.g., *barracuda*, which shares a great deal of phonological overlap with *parakeet*, is not a competitor following the first segment. Furthermore, I assume that each segment is equally perceptually distant from each other, e.g., [b] equally excludes words that continue with [k] as one that continues [p]. In both cases, there is evidence that listeners are sensitive to gradient differences (McMurray et al., 2009) and do not totally exclude competitors (Levy et al., 2009; Toscano et al., 2013), meaning that - as might be expected - there is more to the story than simplified methods can glean.

Moreover, there is evidence that the lexicon is sensitive to specific kinds of contrasts and this may affect the shape of word forms independently. When

looking at the lexicons of various lexicons, Graff (2012) found significantly fewer minimal pairs across perceptually similar dimensions, e.g., stop voicing, than would be expected given other properties of the tested language's lexicons, suggesting that lexicon avoids perceptually similar dimensions for the sole contrast between word forms. In addition, there is a evidence that segments that are the single difference in minimal pairs show heightened, contrastive articulation, with the effect strongest when the competitor differs across a perceptually similar dimension (Nelson and Wedel, 2017). Together, if not directly related, these suggest that lexical structuring is at least partially affected by the actual featural make-up of contrasts, which should play some part in the maximization of entropy for contrasts.

Though I do expect that a future analysis that incorporates more sophisticated measures to provide clearer and insightful results, I feel that the fact that the results here were as clear as they were with such rough representations to be an indication of the strength of the effect. That is, the fact that such patterns could be found with little more than simple phonetically transcribed lemma lists is an indication of the effect and its strength in the lexicon, in and of itself.

Broader Impacts

When considered with other work, these results suggest that not only are word forms organized to be efficient in terms of length, e.g., Zipf's law of abbreviation (Zipf, 1949; Piantadosi et al., 2011; Bentz and Ferrer-i Can-

cho, 2016), and ease-of-production, e.g., Dautriche et al. (2017); Meylan and Griffiths (2017); Mahowald et al. (2018), but that the specific contrasts and their position in word forms are such that the lexicon is a high entropy code. As mentioned in the beginning of this chapter, communicative efficiency is defined as information over time, in other words, as a ratio between entropy and length. As such, showing that word forms are relatively short is a single part of the overall argument that they are efficient vessels for information. The results here suggest that, in fact, length is not the only aspect of the lexicon that is shaped for efficient communication. This lies along the claims made by Köhler (1987) that the lexicon is an ever-changing optimal solution to different pressures of use. On one hand, the lexicon is under pressure for short, easy-to-produce forms. On the other, the lexicon is shaped to provide as much information as possible on average, given the types of contrasts available. These pressures together drive language towards an optimal communicative code, tailored to human communication.

At a broader level, the results found here for individual word forms align with work on higher scopes of linguistic structure, namely syntactic word orders. Genzel and Charniak (2002) show that large discourses are organized to efficiently transmit information, considering that earlier sentences are processed before later ones. Hale (2003, 2006) and Levy (2008) argue that sentences are organized so that the order of words in sentences is optimum given that earlier words are processed before later ones. Ferrer-i Cancho (2017) goes one step beyond and argues that incremental word processing is

partially to blame for word-order (e.g., subject, object, verb) patterns across languages, showing that the most common word orders sit as an optimal balance between efficient information transmission on dependency minimization (Futrell et al., 2015; Gildea and Jaeger, 2015).

In all, the word-level effects that I have shown here and the broader syntactic effects seem to point to a general strategy for linguistic organization, namely one that is higher sensitive to the linear order of constituents. At a syntactic level, words are organized so that they provide a constant and uniform amount of information (e.g., Jaeger 2010, more discussion of *uniform information density* will follow in the next chapter), which is arguably ideal for processing. At a word level, contrasts are organized such that they provide high information as to the identity of the word, i.e., balance contrasts, and are organized so that the least informative contrasts are positioned where they minimally affect the aggregate entropy of the lexicon altogether.

Chapter 4

Synergy between Redundancy and Efficiency

4.1 Background

Thus far, I have shown that the lexicon is structured to be a high information code, given incremental word processing. However, an efficient communication system does not simply transmit as much information as possible, but transmits information so that as much as possible is received accurately. For example, messages of a Huffman code contain the maximum amount of information possible given the number of possible contrasts available (for more detail, see (Cover and Thomas, 2006)). However, Huffman codes distinctly lack any redundancy, meaning that if a single symbol of a transmitted message is incorrectly received, the entire message will be lost. It is therefore

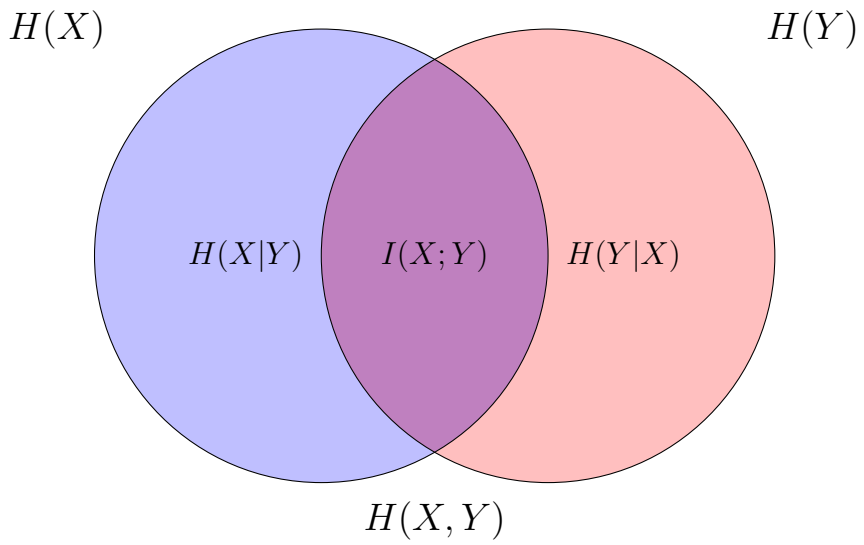

Figure 4.1: A diagram of mutual information. The left-most circle, $H(X)$, represents the entropy of the transmitted message, the right-most circle, $H(Y)$ represents the entropy of the received message. The intersection, $I(X;Y)$, represents the mutual information, or the amount of information that is successfully transmitted.

better to design a code with the receiver in mind, even if it comes at the cost of efficiency.

Abstractly, an efficient communication system should maximize the *mutual information* between transmitted messages and received messages (Shannon, 1948; MacKay, 2003). Mutual information is defined as the information that shared between the original transmitted message and how that message is received (see Fig 4.1). Another way to think of this is as successfully transmitted information, despite possible forms of interference.

There are two main sources of interference in transmission. The first is

the abstract concept of *noise*. Put simply, noise in this case any probability that one segment is understood by a listener as something other than what was intended by the speaker. For example, a spoken [p] in the word *pat* could be received as [b] or lost entirely, causing a listener to incorrectly recognize *bat* or *at* respectively. The confusion between [p] and [b] in this case could be a result of actual acoustic noise, such as a loud background, or it could be the result of something else, such as the listener expecting the word *bat* more than *pat*. Regardless of the actual source, the term noise simply means possible confusion between contrastive units. The second type of interference is the *channel capacity* for language processing. The channel capacity is the limit of the amount of information that can be processed by a listener in a given amount of time. For example, it is possible - though perhaps difficult - to understand a recorded sentence if it is played back at twice speed. It is nearly impossible to understand the same sentence if it is played back at five times the normal speed. That said, channel capacity is not restricted to speed of transmission. A sentence that is very densely encoded with information, e.g., complex syntax and low probability words out of context, is more difficult to process, all things being equal, than multiple sentences that convey the same information with simpler parts. The difficulty in processing a dense sentence may also result in interference with accurate communication.

Therefore, in order to accurately convey as much information as possible, the maximally efficient lexicon should be sensitive to these sources of interference and to be constructed to minimize their effects as much as possible.

I will discuss each of these sources of interference and strategies to mitigate their effects below.

4.1.1 Noisy channel and redundancy

Language exists in a *noisy channel* (Shannon, 1949). To offset the potential effect of noise and ensure accurate communication, messages should include a certain amount of redundancy. This redundancy can be applied ad hoc to a message, e.g., producing words with greater duration when there is a greater chance of confusion, or it can be encoded into the lexicon itself. For example, if part of the first syllable of *chimpanzee* is distorted by noise, the word can still be correctly understood because of the redundant information in *-anzee*. In this case, I will refer to the redundancy that is encoded into a word form as *form-internal redundancy*, that is the material in a word form that goes beyond the minimum needed to uniquely distinguish it from others in the lexicon. As the example of *chimpanzee*, longer words generally possess more redundant material, as more of them must be distorted to be confused with a lexical competitor.

Yet, length is not the only means to encode redundancy in a word form. Though both word forms are three segments long, *chat* possess more form-internal redundancy than *bat*; if the [æ] in *chat* were to be distorted by noise, a listener would still have more information to identify the word than if the same were to happen to *bat* because fewer word forms in English fit the /t͡ʃ-t/ skeleton than /b-t/. Considering this, one measure of redundancy can be

thought of as the distance to other forms in the lexicon, with length being strongly correlated with this.

In the same vein, because word processing is incremental, all material after enough of the word has been produced to distinguish a word form from others in the lexicon is present, i.e., its *uniqueness point* (c.f. Marslen-Wilson and Welsh 1978), can be considered redundant. For example, the uniqueness point of the word *aviation* /evieʃɨn/ in English is [ʃ] since no other word form begins with /evieʃ-/ (according to my English corpus). As might be expected, longer words generally will have more material after their uniqueness point, which again strengthens the relationship between word length and form-internal redundancy.

Nevertheless, redundancy does have clear drawbacks. Increasing the length of word forms increases production effort as, all things being equal, longer word forms require more effort to produce. Similarly, longer word forms spread their information over a longer duration, reducing the average information per part of a message. As discussed in the previous chapter, the information within part of a word form is relative to how well it disambiguates other forms of the lexicon. If a word form is longer, then each part of it will exclude fewer competing word forms on average, meaning that the average information per part is less. Furthermore, increasing the distance between forms independent of length, also increases production effort across the lexicon as a whole. At its core, speech and the act of producing words is a form of motor movement which means that regularly repeated gestures,

i.e., the most common segment sequences in a language, require less effort on average (Vitevitch et al., 2004; Dautriche, 2015; Tomaschek et al., 2018). As a function of this, a lexicon with a relatively high proportion of repeated material, such as a lexicon with a Zipfian distribution of segments, is less effortfull to produce all things being equal. Given these pressures, an efficient lexicon for communication in a noisy channel should not include large amounts of redundancy across the board, but rather include redundancy only where needed. As it happens, certain words benefit from redundancy more than others, making the forms of these words better candidates for inclusion of form-internal redundancy.

At an abstract level, spoken word identification is a type of inference (e.g., Gibson et al. 2013; Kleinschmidt and Jaeger 2015), where a listener attempts to find the word in the lexicon that is most likely given both acoustic information and the word's prior probability (Eq. 4.1)[1]. For word identification to be accurate, the product of a word's prior probability, $p(word)$, and the probability of that word given the spoken acoustics, $p(acoustics|word)$, for the intended word should be higher than all others in the lexicon.

$$p(word|acoustics) \approx p(acoustics|word) * p(word) \tag{4.1}$$

Importantly, word forms that are more dissimilar from the rest of the lexicon, i.e., those with the greatest form-internal redundancy, will yield a

[1]I leave out the marginal prior, $p(acoustics)$ in the Bayesian reversal of conditional probabilities because, simply put, $p(acoustics)$ is the same for all word forms.

greater value for $p(acoustics|word)$ for the average production of the intended word and a yield a much lower value for all forms in the lexicon. This is particular beneficial to words that are less probable on average as they will often be relatively weaker competitors, compared to words that are more probable on average. That is, if a word's prior probability is much less that that of a competitor which is perceptually similar, the word may be ruled out as the intended word, regardless of the exact acoustic information[2]. To mitigate this, the lexicon can allocate greater redundancy to the forms of less probable words, offsetting the effects of a lower average probability while still limiting the addition of redundancy to forms where it has the most benefit.

Interestingly, there is evidence that the lexicon is structured to allocate greater form-internal redundancy to less probable words. Across many languages, less probable words are generally longer (Zipf, 1935; Piantadosi et al., 2011; Bentz and Ferrer-i Cancho, 2016), with length being strongly correlated with the amount of form-internal redundancy in a word form. More specifically, there is a robust trend for less probable words to be more dissimilar from others in the lexicon than more probable words (Landauer and Streeter, 1973; Frauenfelder et al., 1993; Siew, 2013; Mahowald et al., 2018) and to possess more segments that fall after a word form's uniqueness point (King and Wedel, 2018, 2020). Put together, this shows that the lexicon is at least partially sensitive to the benefits of form-internal redundancy, particularly

[2]As empirical evidence of this, listeners are more likely to mistake a less probable word for a more probable competitor when subject to noise or acoustic distortion (Ganong, 1980; Hilpert, 2008; Gibson et al., 2013).

for less probable words.

Note that neighborhood size and number of post-uniqueness point segments are two of many possible ways of measuring redundancy in word forms. To some extent, phonotactics and higher-level constraints on word formation encode redundancy into word forms themselves. The identity of a segment after a word-initial [b] in English is partially redundant since English phonotactics limits the number of legal consonant clusters. The roundness of the second vowel in a Turkish word is partially redundant since all vowels share that feature. The number of root consonants in a Hebrew is partially redundant since words are primarily based on a system of 3 consonant root. Nevertheless, neighborhood size (Vitevitch et al. 2008; Chan and Vitevitch 2009; Siew and Vitevitch 2016, for discussion see Luce and Pisoni 1998; Chen and Mirman 2012) and relative position of uniqueness point (Marslen-Wilson and Welsh 1978; Marslen-Wilson 1987; Radeau et al. 1989; Grosjean 1996; Radeau and Morais 1990, for discussion see Dahan and Magnuson 2006; Weber and Scharenborg 2012) stand as easily quantifiable metrics for form-internal redundancy that have also been shown to play roles in word recognition. It is for these reasons that I will focus on them to measure the amount of redundancy in word forms. I suspect that additional measures of form-internal redundancy will lead to more nuanced discussion of the role of redundancy in the lexicon.

4.1.2 Channel capacity and a smooth signal

Though redundancy does allow information to be transmitted through noise, there is still a limit to the amount of information that a listener can reasonably process over a given time (for discussion, see Levy 2008; Christiansen and Chater 2016). In the terminology of Information Theory, this is the *channel capacity*, or the maximum amount of information that can be communicated by a single part of a message. Regardless of the exact value for this limit (e.g., Mollica and Piantadosi 2019), a more uniform distribution of information within the message is ideal.

Consider Fig. 4.2, where each shape represents a method to encode information. In this diagram, the rectangle is the optimal shape for the information in the message.

If it were shorter in duration, the height would need to increase in order to still encode the same total information and part of the area would fall above channel capacity be lost. Were the rectangle longer in duration, all information would still be communicated successfully, though it would be less efficient because it would take more time. Now, consider the curved shape. Though of equal duration to the rectangle, this shape is not uniform and part of it falls above the channel capacity where information is lost and part of it falls below capacity where communication is less efficient. Though abstract, this diagram conveys the benefit of uniform encoding of information.

Interestingly, this is the exact prediction of *uniform information density* (e.g., Jaeger 2010; Jaeger and Tily 2011) which argues that a steady and

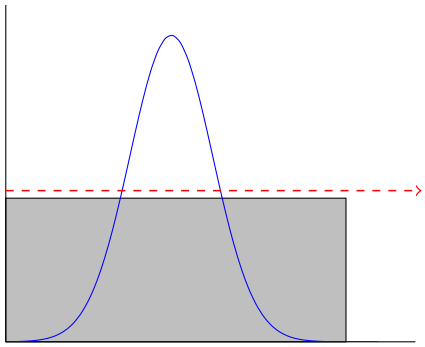

Figure 4.2: The rectangle and parabola represent two possible ways to abstractly shape the information content of a message, the x-axis represents time, the y-axis represents information at a given moment and the dotted red line represents the channel capacity. Though each shape has equal area, the rectangle with its uniform information rate is the optimal shape because it is closer to channel capacity as much as possible.

uniform stream of high information parts of the message is optimal and that speakers regularly modulate how they speak to create a more or less uniform distribution of information. For example, Levy and Jaeger (2007) showed that speakers are more likely to produce the optional complementizer *that* in English sentences if the beginning of a relative clause is less contextually likely. By doing so, the potential 'jump' in information from a low probability relative clause is mitigated, creating a smoother distribution throughout the sentence. On the other hand, by not including *that* when a relative clause is more likely, speakers avoid adding redundant material when it is less useful for smoothing out the signal. As a result, the average sentence is closer to uniform than if this strategy was not employed and this is by no means the only evidence of speakers actively smoothing out the distribution of information in utterances.

Bell et al. (2003) found a link between contextual probability of function words and their duration and degree of reduction. Aylett and Turk (2004, 2006) found a significant relationship between sentence-level contextual probability of words and the duration and phonetic specificity (i.e., extremeness of F1/F2) of the token. Mahowald et al. (2013) showed that both in corpora and experimental settings, speakers were more likely to truncate words, e.g., *chimpanzee* → *chimp*, when the word was contextually likely. Galati and Brennan (2010) found that when words were more likely given an extralinguistic context (i.e., visible pictures), they were more reduced. Uther et al. (2007) and Pate and Goldwater (2015) found that speakers even modulated production depending on their listener, only reducing words that the listener was likely to know given the listener's experience with the language (though see Turnbull 2015; Tomaschek et al. 2018).

At a phonetic level within word forms, van Son and Pols (2003); Van Son and Van Santen (2005) found that segments that were had a lower contextually probability had greater durations. Tang and Bennett (2018) demonstrated a similar effect in the morphologically complex language Kaqchikel, showing that these effects are not limited to a small set of well-tested and relatively morphologically simple languages like English and Dutch. Seyfarth et al. (2016) found that the single difference in minimal pairs showed greater duration, i.e., voice-onset time for [t]/[d] minimal pairs, when the other form in the pair was also contextually likely, suggesting that listeners modulate pronunciation to explicitly exclude likely and similar competitors. Consider-

ing that listeners are better able to identify contextually likely words from sparser acoustic information (e.g., Van Berkum et al. 1999, 2003, 2005; Gibson et al. 2013), this seems an ideal strategy to balance between entropy and effort. In other words, the production of a segment and its duration is a function of the information it provides to identifying the containing word form (see Hall et al. 2016 for a review).

At a syntactic level, Genzel and Charniak (2002) showed that sentences become more complex, i.e. sentence-level entropy increases, for later sentences in journalistic articles. This, they argue, is a product of later sentences being able to rely on context provided by earlier sentences, making these more complex sentences simpler to process than they would be without context. Frank and Jaeger (2008) found similar results, showing that speakers are more likely to avoid using contractions, such as *you are* → *you're*, when the contracted words are less likely. Kurumada and Jaeger (2015) found that optional object marking in Japanese was more likely to be absent when it was contextually likely, avoiding more or less redundant material when it contributed little to a listener's comprehension and Fedzechkina et al. (2012) found that this may play into the diachronic development of morphology in a language. Pellegrino et al. (2011) and Coupé et al. (2019) demonstrated a correlation between speech rate (syllables per second) and information (information per syllable) across a varied set of languages. Put simply, the syllables of languages which are spoken more quickly possess less information on average and vice versa, which creates a more or less equivalent rate

of information transmission across languages.

Altogether, this shows that language users often modify messages to make information more uniformly distributed throughout word forms. This process can have residual effects on the word forms themselves. For example, word forms that are more often articulated with greater duration due to local context are produced with greater duration for all productions (e.g., Bybee and Hopper 2001; Seyfarth 2014; Nelson and Wedel 2017; Sóskuthy and Hay 2017), which has been proposed to be evidence of a pathway for the lexicon to become shaped for efficient communication (Piantadosi et al., 2009, 2011; Kanwal et al., 2017). As such, this stands as strong evidence that the lexicon is not only structured to be sensitive to the benefits of redundancy, but to the benefits of uniform information density as well.

4.1.3 Mutual information and an efficient lexicon

In the previous chapter, I have shown evidence that the lexicon is structured so that inter-word contrasts are higher information than might be expected otherwise. In the previous sections, I have discussed evidence that the lexicon is structured for both an optimal assignment of form-internal redundancy and uniformness of information over the average message. The next question pertains to the synergy of these properties. That is, do these properties work together in the lexicon to make it an overall more efficient code for communication?

Often, efficiency and redundancy are seen as opposing ends of a single

dimension. If this were the case in language, any structuring for redundancy would damage the lexicon's capacity for efficiency and vice versa. However, language is too complex to be represented as a single dimension. It may be the case that certain parts of the lexicon are optimized for efficient encoding of information while others are designed for the ideal amount of form-internal redundancy, all the while being sensitive to the importance of a uniform distribution of information.

Consider again the simple model of lexical contrasts of English in Fig. 4.3. Because of the frequency distribution in this model lexicon, each contrast is balanced and thereby has maximal entropy. In addition, there is a relationship between form-internal redundancy and word probability across the lexicon, e.g., *the* [ðə] has one segment after its uniqueness point while *sphinx* has three. Moreover, because each contrast has the same entropy, the flow of information is closer to uniform. Put together, this toy lexicon demonstrates structuring for efficient transmission of information, while being sensitive to both noise and channel capacity.

Now, what would happen if the association between word forms and word probability was altered? If *the* became the least frequent word in the lexicon and *sphinx* the most, this toy lexicon would no longer be as efficient. Specifically, a) contrasts would become less balanced, b) the benefit of redundancy would be weakened and c) the average information per part of the message would be further from uniform. This leads to a testable prediction. Is the lexicon organized such that all three of these pressures are satisfied more

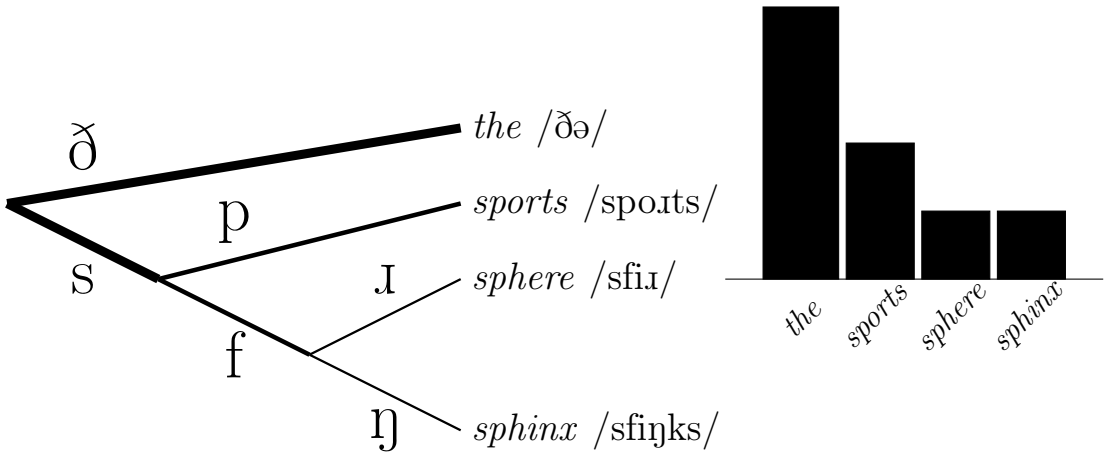

Figure 4.3: A maximally efficient toy version of English. The bar chart on the right represents the relative probability of each word. Because of the assignment of word probability to form, each contrast has maximum entropy and the least probable word forms have the most form-internal redundancy. Because the lexicon has high information contrasts, while still being structured for being robust to noise, this toy lexicon is an efficient lexicon.

than would be expected otherwise?

Comparison Lexicons

To test this, I will compare the real-world lexicon against a distribution of structurally similar but randomized variants, which will all consist of phonotactically licit word forms for the language. That said, testing whether one form of the lexicon is more efficient than another is not straightforward. In this case, the overall efficiency of the lexicon is related to how well it satisfies the three pressures previously mentioned, each of which is guaranteed to share a complex relationship with the others. For example, one way for a lexicon to possess greater information per contrast than another is to include

no redundancy at all. However, this lexicon, though more efficient, is less robust to noise, making it difficult to determine which of the two is more efficient overall.

On the other hand, if one form of the lexicon possess higher information contrasts than another and both are equally structured for beneficial redundancy, then the high information one is, put simply, more efficient, as it satisfies all pressures at an equal or better level. Following from this, a simple but feasible means to evaluate whether the real-world lexicon represents a better synergy of the different aspects of optimization is to compare it against variant forms of the lexicon where certain aspects have been kept the same.

I will compare the real-world lexicon against a baseline set of randomized variants of the original lexicon where the relationship between word probability and form-internal redundancy has been kept constant (for more detail, see Chapter 2.0.2). By doing so, each of the randomized variants will not vary in the amount of beneficial redundancy. The variant lexicons will vary, however, in the average amount of information per contrast and the distribution of information throughout word forms. If the real-world lexicon possesses higher information contrasts that are more uniformly distributed within word forms than this baseline, it will show that the lexicon is more efficient, without sacrificing robustness to noise. As such, it will be evidence that the lexicon is structured to be more efficient, given multiple pressures for communication.

The next question is how to capture the average information per contrast with a single, comparable measure. Recall that the information of a segment is related to its contextual probability, given the earlier material of the word. For example, the [f] in *sphinx* is a relatively high information segment because few word forms that begin with [s] also begin with [sf]. However, even though this [f] is relatively high information, it has a relatively small contribution to the overall informativeness of the lexicon. As it stands, *sphinx* is a relatively low probability word meaning that the information contributed by any of its segments is unlikely to be included in a sentence, all things being equal. A simple way to account for both the information of segments and their overall likelihood in messages is to represent the average information of the lexicon as the joint entropy of word and segments. This can also be thought of as a weighted average of segment information values, where the weight is equal to the probability of the word that contains the segment (e.g., Eq. 4.2).

$$H_L = \sum_{w} \sum_{s_i \in w} p(w) * h(s_i|s_1...s_{i-1}) \tag{4.2}$$

However, this metric by itself does not factor in uniformity. Because of this, I will be using a slightly modified form which does (Eq. 4.3). Crucially, in this modified equation, each segment's information is multiplied by its contextual probability. This has the effect of penalizing sharp peaks or drops in information across a word form; the sum of a word form's token-based segment information is equal to its *word information* or $-\log_2 p(word)$, re-

gardless of how information is distributed (for discussion, see King and Wedel 2020). However, with this modified form of the equation, word forms that have a more uniform distribution will contribute more to the overall sum for the lexicon.

$$H_L^* = \sum_{w} \sum_{s_i \in w} p(w) * p(s_i|s_1...s_{i-1}) * h(s_i|s_1...s_{i-1}) \tag{4.3}$$

Therefore, using this metric allows for comparison of average information overall as well as a general uniformity throughout each word. If the real-world lexicon shows a greater value for Eq. 4.3 than the randomized variants, it will suggest that the lexicon is structured for high information contrasts, while still being sensitive to the communication in a noisy channel and the benefits of a uniform information.

4.2 Methods and Results

4.2.1 Preparation

To prepare the data for this section, I first calculated token-based segment probability and information for all words of the lexicon (Eqs. 4.4 - 4.5). I chose to use a token-based mesaure of segment information over a type-based one in order to allow variation when shuffling word probabilities; type-based segment information would be identical for probability-shuffled variants of the lexicon. I did not remove any words or segments from the analysis, e.g., the

5% least probable, because the modified formula for lexical entropy includes multiplication by segment probability; the contribution of segments with a relatively low conditional probability were already discounted, minimizing the need to remove them.

$$p(s_i|s_1...s_{i-1}) = count(\frac{s_1...s_i}{s_1...s_{i-1}}) \tag{4.4}$$

$$h(s_i|s_1...s_{i-1}) = -\log_2 p(s_i|s_1...s_{i-1}) \tag{4.5}$$

Comparison lexicons

To create variant lexicons that maintain the same amount of beneficial redundancy, I shuffled word probabilities of words that possess the same length and the same amount of form-internal redundancy. For example, *drink* /drɪŋk/, *bronze* /bɹanz/ and *thwart* /θwaɹt/ are all five segments long, though *drink* and *bronze* have two redundant (post-uniqueness point) segments while *thwart* has four. Here, only *drink* and *bronze* are candidates for switching as they possess the same proportion of redundant material. Shuffling word probability in this way ensured that the relationship between word probability and length, i.e., Zipf's law of abbreviation, was identical for all randomized variants, as was the relationship between word probability and form-internal redundancy.

In order to reduce the chance of the results being skewed by idiosyncrasies

of one measure of redundancy or the other, I chose to use two means to determine form-internal redundancy. The first was the number of segments that fall after a word form's uniqueness point. A word form's *uniqueness point* (c.f. Marslen-Wilson 1987) is the segment where the word becomes unique in the lexicon, e.g., [j] in *vacuum* because no other word form begins with [vækj]. The second was the number of neighbors a word form had. A lexical *neighbor* (c.f. Luce and Pisoni 1998) is a word form that differs by a single segment, e.g., *pat*, *bat* and *at* are all neighbors.

I chose these measures specifically because they are discrete and allow for large enough groups of words in each probability-shuffling group (see Fig. 4.4). For the probability-shuffling paradigm to yield enough permutations of the original lexicon to be successful, there must be a large set of word forms that are candidates to shuffle probabilities with one another. Were I to use a more continuous measure of form-internal redundancy, e.g., mean edit distance, there would unlikely be a sufficiently large number of word forms to shuffle, meaning that there would not be a sufficient number of permutations of the lexicon as a whole.

Note that I removed word forms from the analysis if there were no other words in the lexicon that shared both length and form-internal redundancy, as these words would be unchanged in probability-shuffled lexical variants.

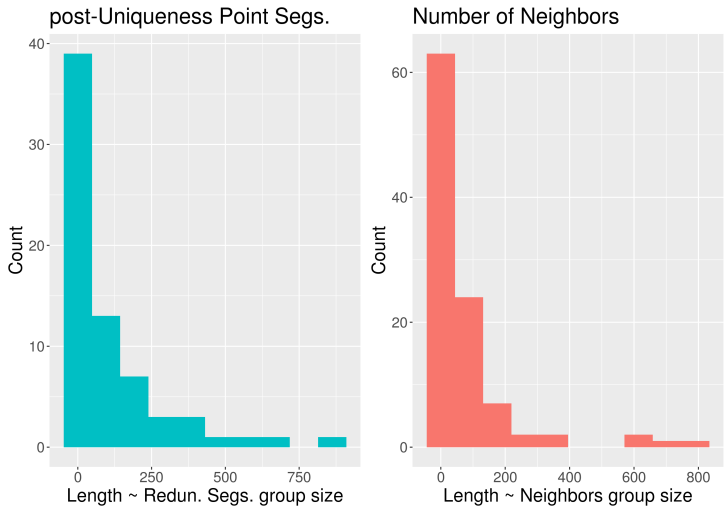

Figure 4.4: Distribution of group sizes (word forms that share both length and form-internal redundancy) for both measures of redundancy. The x-axis shows the size of the group of word forms that share both length and form-internal redundancy, e.g., words of length six with two segments after their uniqueness points. The y-axis shows the number of distinct groups for each size, e.g., there are five groups with roughly 200 members in the right graph. Though the distribution is strongly right-tailed, a large part of the English lexicon has sufficiently sized groups for a large possible space of word probability permutations.

4.2.2 Results

Shuffling within Neighbor groups

For each language, I created 1,000 probability-shuffled lexicons and compared the value for H_L^* in the real-world and novel lexicons. Recall that H^*L is a measure of the overall entropy of segment contrasts in the lexicon, modulated how uniform the distribution of segment information is within a word form. For this value to be higher in one form of the lexicon than another,

segments must have both higher information on average and the segments of a particular word form should possess more similar information values.

For many languages (20 out of 25), H_L^* in the the real-world lexicon fell significantly (greater than 95% of probability-shuffled alternates) above the baseline distribution created by the randomized lexicons, indicating that the word forms for these languages more efficiently and uniformly encode the necessary information to disambiguate other forms in the lexicon, while maintaining the same relationship between word probability and the values for both length and the number of neighbors (see Figs. 4.5, 4.6 - 4.8). This suggests that the segment-level contrasts within word forms are organized to be both high information as well as similar in information value to other segments in the same word form. Crucially, the greater value of segment entropy and uniformity does not come at the expense of robustness, i.e., lowering the amount of beneficial form-internal redundancy in the lexicon, suggesting that the lexicon is structured to be both efficient and accurate, as predicted.

When tested in a mixed-effects logistic model with random intercepts per language per family, lexicon-type (probability-shuffled vs. real-world) was found to predict a significantly greater total lexical entropy for real-world lexicons compared to their probability-shuffled counterparts (see Tab. 4.1)[3]. This indicates that the effect carried across the dataset, though it might not

[3]R formula: glmer(data = all_langs, lex.type ~ diff.from.mean + (1|family/language), family = binomial)

A. Fixed Effects:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.608	0.616	-15.604	**0.001**
H_L^*; difference from mean	2.069	0.244	8.478	**0.001**

B. Random Effects:

	Name	Variance	Std.Dev.
Language:Family	(Intercept)	0.051	0.227
Family	(Intercept)	0	0

Table 4.1: Mixed-effects logistic model for predicting lexicon type (probability-shuffled vs. real-world) given relative difference in H_L^* from the mean of the probability-shuffled alternatives. Language is given as a random factor, nested within family. The model a significant, positive effect of mean difference, indicating that the real-world lexicons had greater H_L^* across the dataset.

have manifested in some languages individually.

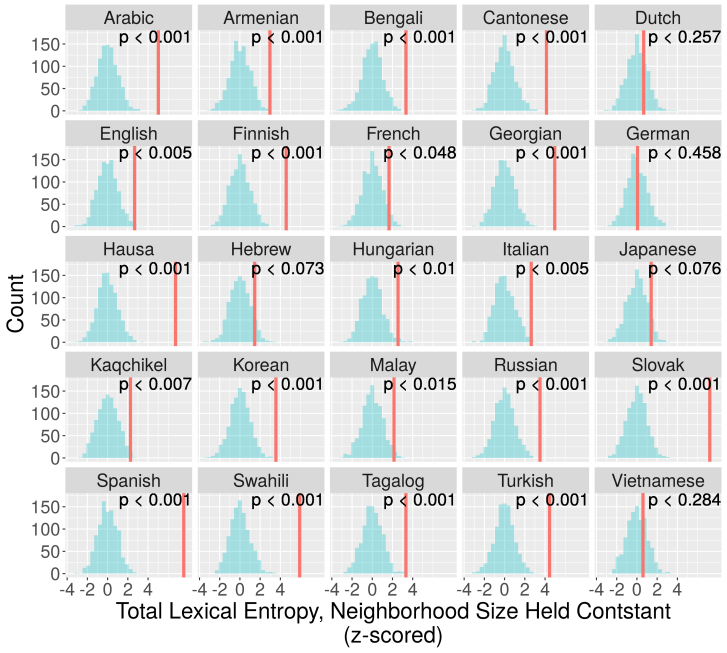

Figure 4.5: Comparison for total lexical entropy, H_L^*, in the real-world and probability-shuffled lexicons, where number of neighbors has been held constant across shuffles. The blue histogram shows the distribution for the probability-shuffled variant lexicons and the red line indicates the value for the real-world lexicon. For all languages shown except Dutch, the real-world lexicon possess a greater value than 95% of probability-shuffled lexicons.

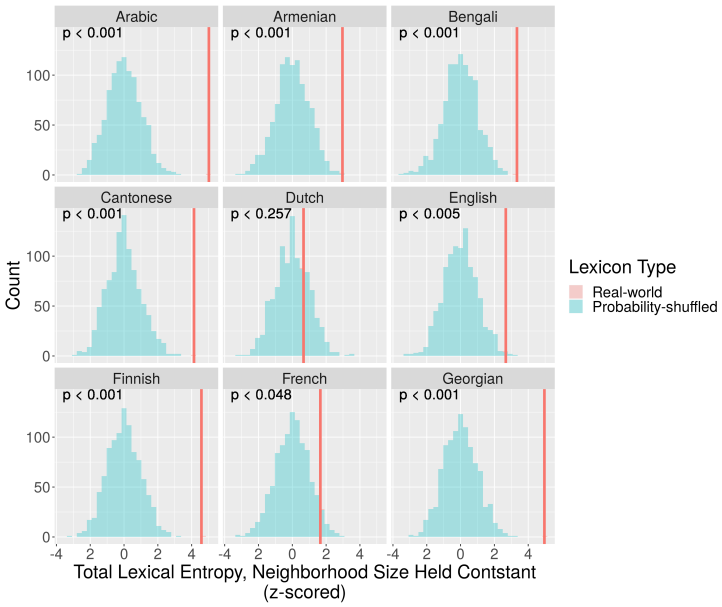

Figure 4.6: Comparison for H_L^* in the real-world and probability-shuffled lexicons. For all languages shown except Dutch, the real-world lexicon possess a greater value than 95% of probability-shuffled lexicons.

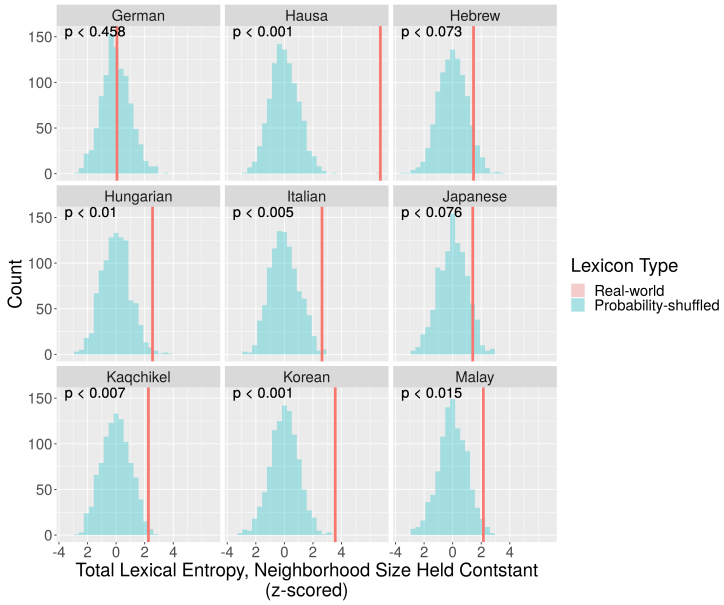

Figure 4.7: Comparison for H_L^* in the real-world and probability-shuffled lexicons. For all languages shown except German, Hebrew and Japanese, the real-world lexicon possess a greater value than 95% of probability-shuffled lexicons.

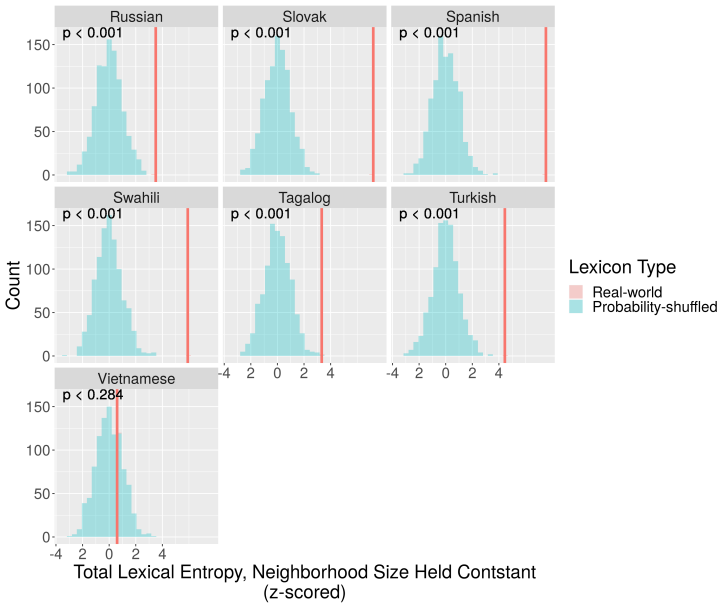

Figure 4.8: Comparison for H_L^* in the real-world and probability-shuffled lexicons. For all languages shown except Vietnamese, the real-world lexicon possess a greater value than 95% of probability-shuffled lexicons.

Shuffling within Tail length groups

I continued by performing similar tests, comparing the real-world lexicon against a distribution of 1,000 probability-shuffled lexicons where shuffling occurred within groups of word forms that shared the same number of post-uniqueness point segments. As with the previous tests, the majority (23 of 25) languages showed a greater value of H_L^* in the real-world lexicon when compared to the distribution of probability-shuffled novel lexicons (see Figs. 4.9, 4.10 - 4.12). Together, the results from the tests of individual languages suggest an overall trend for the tested lexicons to be organized such that word forms posses higher information and more uniformly informative segments than would be expected from alternative codes that maintained the same relationship between word probability and both length and tail-length[4].

When tested in a mixed-effects logistic model, I found that H_L^* in the real-world lexicons were significantly greater, i.e., had a greater difference from the mean for all values of total lexical entropy (see Tab. 4.2)[5]. This suggests that though the results were not uniform in the per language tests across the entire dataset, the pattern is strong enough to meet the criterion for significance.

[4]The fact that segments in the tail possess 0 information does not affect the overall *smoothness* of segment information because across both the real-world and novel lexicons, the same number of '*0-information*' segments is constant.

[5]R formula: glmer(data = all_langs, lex.type ~ diff.from.mean + (1|family/language), family = binomial)

A. Fixed Effects:				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.331	0.644	-14.479	**0.001**
H_L^*; difference from mean	1.956	0.262	7.462	**0.001**
B. Random Effects:				
	Name	Variance	Std.Dev.	
Language:Family	(Intercept)	0.547	0.739	
Family	(Intercept)	1.924	4.387	

Table 4.2: Mixed-effects logistic model for predicting lexicon type (probability-shuffled vs. real-world) given relative difference in H_L^* from the mean of the probability-shuffled alternatives. Language is given as a random factor, nested within family. The model a significant, positive effect of mean difference, indicating that the real-world lexicons had greater H_L^* across the dataset.

4.3 Discussion

In this chapter, I have presented evidence that the lexicon is structured to balance between the benefits of form-internal redundancy and efficient encoding. Specifically, I showed that the real-world lexicons of a diverse set of languages are structured to more efficiently encode disambiguating information in word forms, when compared against structurally similar but randomized codes. As such, the real-world lexicon better satisfies the pressures for both efficiency and redundancy, important features of a communicative code given how language is used to communicate. This stands as an important contribution to the growing body of work on the systemic effects of efficiency on language (for review, see Gibson et al. 2019), by providing synchronic evidence for two distinct pressures affecting the lexicon while minimally affecting each other.

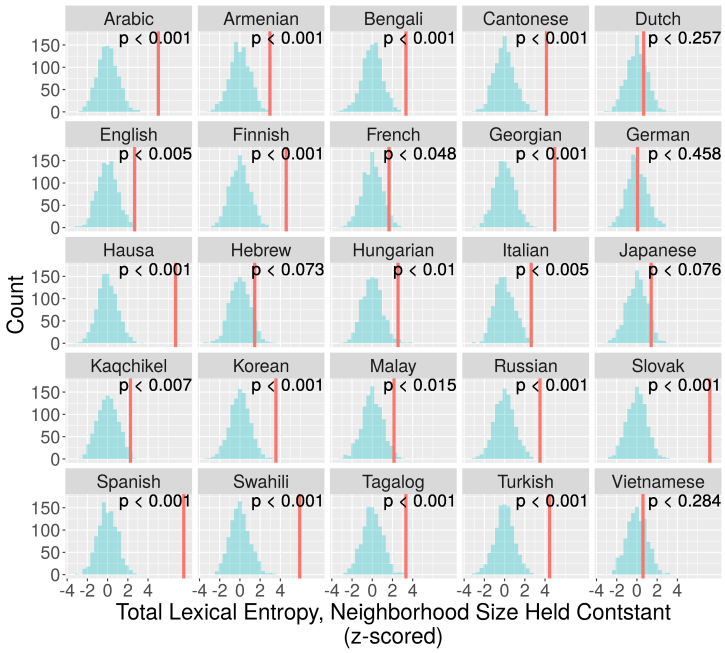

Figure 4.9: Comparison for total lexical entropy, H_L^*, in the real-world and probability-shuffled lexicons,where number of post-uniqueness point segments has been held constant across shuffles. The blue histogram shows the distribution for the probability-shuffled variant lexicons and the red line indicates the value for the real-world lexicon. For all languages shown except Dutch, the real-world lexicon possess a greater value than 95% of probability-shuffled lexicons.

4.3.1 Post-hoc analysis

An interesting result of this section was that Dutch failed to show any difference from the randomized baseline for both sets of tests. This may be due to an idiosyncrasy with the Dutch corpus used, given that all other Indo-European languages showed the predicted effects for at least one of the tests. Given that the Dutch data was lemmatized, it could be possible that the exact means of lemmatization may be responsible for lack of observed differ-

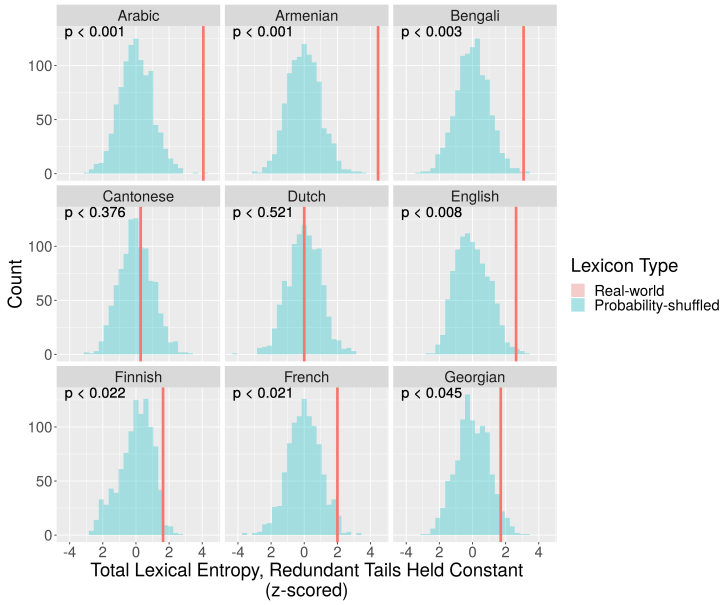

Figure 4.10: Comparison for H_L^* in the real-world and probability-shuffled lexicons. For all languages shown except Cantonese and Dutch, the real-world lexicon possess a greater value than 95% of probability-shuffled lexicons.

ence between real-world and probability-shuffled lexicons. I am not stating that a lexicon of lemmas should not show the predicted effects, but rather the particular choices made by the corpus curators when lemmatizing the data may have had undesired effects. For example, verbs in the Dutch corpus are given in their infinitive form, rather than as an uninflected stem, and the fact that all verbs end in the same infinitive ending may drive the lack of effect.

To investigate this further, I collected word forms from 1 million Wikipedia pages (Goldhahn et al., 2012) and used that as an alternate Dutch corpus. In this data source, the word forms were not lemmatized but I did use regular expressions to reduce multi-character sequences to single characters to give

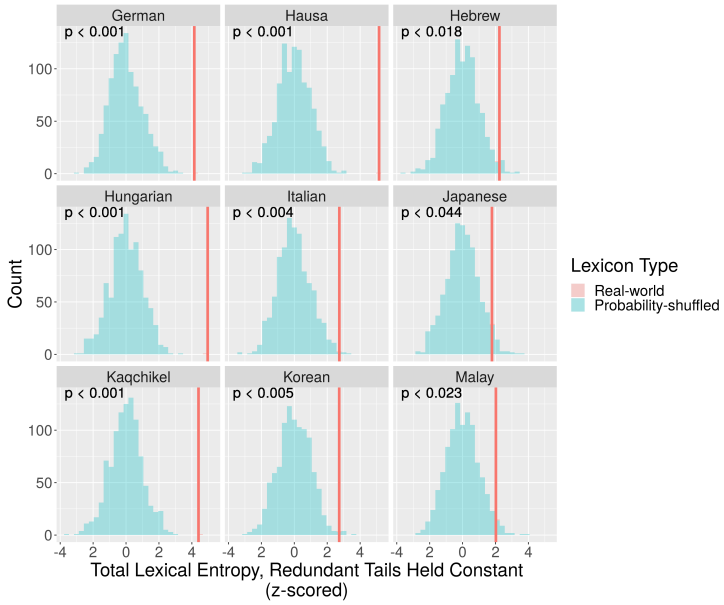

Figure 4.11: Comparison for H_L^* in the real-world and probability-shuffled lexicons. For all languages shown, the real-world lexicon possess a greater value than 95% of probability-shuffled lexicons.

the word forms a 1-to-1 correspondence between symbol and segment. I then compared the total modified cohort entropy, H_L^*, for this set of word forms against two sets of 1,000 probability-shuffled variants of the unlemmatized lexicon, where I kept the amount of form-internal redundancy constant as I had in the original tests (see Fig. 4.13).

Interestingly, using raw Dutch word forms, there was a significant difference for the real-world lexicon for one means of determining form-internal redundancy (segments in tail) but not for the other (neighborhood size), though it did trend towards significance ($p < .1$). This suggests that part of the failure for a significant difference to emerge in earlier tests result from

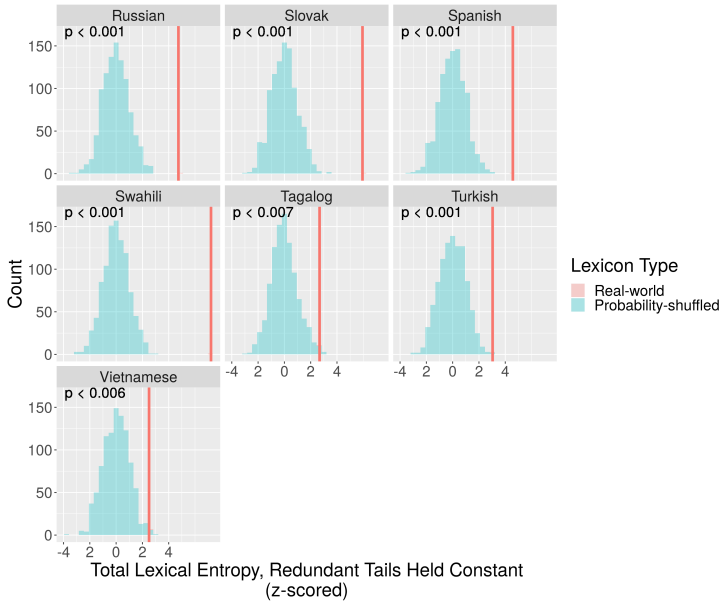

Figure 4.12: Comparison for H_L^* in the real-world and probability-shuffled lexicons. For all languages shown, the real-world lexicon possess a greater value than 95% of probability-shuffled lexicons.

the lemmatization strategy for the Dutch corpus, though it may not explain everything. In this light, further investigation is merited with respect to how Dutch word forms differ from those of other Germanic languages, specifically in how lexically contrastive information is encoded in the segments of word forms.

4.3.2 Trade-offs in the Lexicon and Language

Nevertheless, these results offer strong support that the lexicon evolves as a balance between disparate pressures. By far, this is not the first argument for or evidence of multiple pressures affecting language in tandem. In fact, many

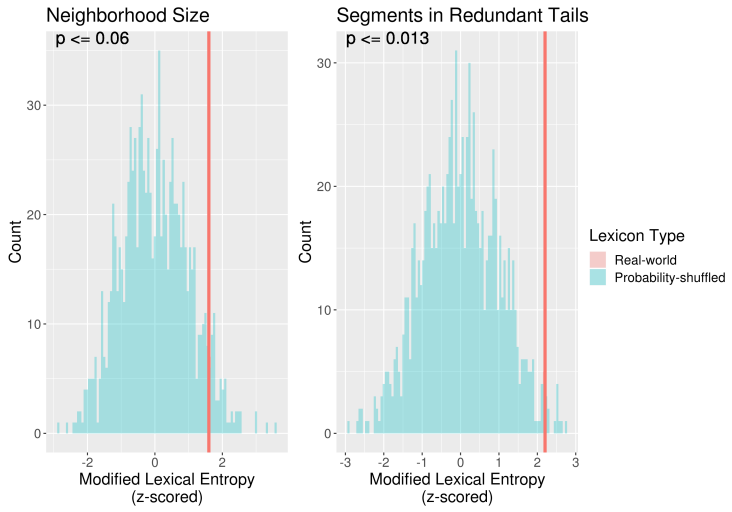

Figure 4.13: Comparison between real-world and probability-shuffled lexicons for unlemmatized Dutch corpus. For both measures of form-internal redundancy, the real-world lexicon possessed a relatively greater value of H_L^* than with the lemmatized data, suggesting that the methods of lemmatization may have affected earlier results.

argue that language broadly is a product of multiple pressures to benefit communication (e.g., Zipf 1949; Köhler 1987, 1993; Ferrer-i Cancho and Solé 2003; Piantadosi et al. 2009, 2011, 2012; Dautriche 2015; Dautriche et al. 2017; Meylan and Griffiths 2017; Mahowald et al. 2018; King 2018; King and Wedel 2020), though they do so without explicit evidence of the balance between high information contrasts and redundancy in the lexicon.

In his *Principle of Least Effort*, Zipf (1949) argued that word forms evolve to balance between ease-of-production and accuracy, i.e., the likelihood a word is correctly identified, with word length being a simple and informative correlate of these. In a theoretically maximally efficient code, message length should be a function of probability Shannon (1948); MacKay (2003), with

word frequency being a good proxy for probability and word length being a good proxy for production effort. Köhler (1987, 1993) argued that production ease and accuracy can be broken down into several different properties, with the target of evolutionary optimization not being a single static state per se, but rather an optimal relationship between various properties. Ferrer-i Cancho and Solé (2003) showed evidence that language-like word frequency distributions provided both benefits to speakers and listeners, using abstract proxies for both both production effort and accuracy in mathematical simulations.

Empirically, there is strong evidence that the lexicons of real-world languages are closer to optimal codes than might be expected otherwise. Piantadosi et al. (2009, 2011) found a stronger correlation between average contextual probability of a word, i.e., *word informativity*, and length, showing that length is partially a function of word informativity and not simply frequency, as would be expected in a code optimized for both speakers and listeners. Of course, word length itself is a proxy for production effort and Dautriche (2015); Dautriche et al. (2017); Meylan and Griffiths (2017); Mahowald et al. (2018) showed that beyond number of segments, more probable words are phonotactically simpler and therefore less effortful to produce (e.g., Vitevitch and Luce 1999). Together, these provide even more compelling evidence that word forms are structured to be efficient to produce.

Yet, optimal communication still requires that efficiently-produced word forms are understood correctly. King (2018) provided evidence that the ex-

act contrasts that make up word forms in English are structured such that confusion between word forms is less likely than might be expected in a random code. King and Wedel (2020) showed that distribution of word-internal information is ideal for how words are processed by listeners. Together, these provide more evidence that word forms are structured to increase the probability of correct identification, though independently of any structuring for efficient production.

Here, I explicitly show how both aspects of lexical optimization co-exist in the lexicon. This represents a crucial piece of evidence that the lexicon is shaped for both speakers and listeners, structuring words to benefit speakers while not adversely affect listeners and vice versa. As such, this offers a means to make predictions for an *optimal* lexicon more explicit. Zipf's *Principle of Least Effort*, for example, can possibly be enhanced, considering that I have presented a more explicit means to measure the amount of information in a word form to be used in identification. With this in hand, the predictions for the shape of the lexicon with respect to theoretically optimal code will be easier to test and hopefully allow for more comprehensive explorations into evolutionary effects on the lexicon.

Chapter 5

Conclusion

In this dissertation, I have shown several statistical properties of a diverse set of languages that together argue for the lexicon as being optimized for communication. In Chapter 3.1, I showed that lexical contrasts in word forms are closer to being balanced than would be expected otherwise. In Chapter 3.2, I showed that the lexicon is organized to position to most informative contrasts where they will have the greatest impact on the aggregate informativeness of the lexicon altogether. In Chapter 4.1, I showed that the structuring for high information contrasts does not interfere with the lexicons assignment of form-internal redundancy or sensitivity to the benefits of a uniform information distribution throughout word forms. Together, these findings offer support that the lexicon is shaped to be a high information code, while being sensitive to the specific pressures of human language use, making the lexicon more efficient than would be expected otherwise. When considered

with other work on the evolution of the lexicon (e.g., Zipf 1949; Köhler 1987; Piantadosi et al. 2011; Kanwal et al. 2017; Mahowald et al. 2018; King and Wedel 2020), this offers strong evidence that language more closely resembles an efficient communicative system, suggesting that communicative efficiency is a driving force of linguistic evolution.

5.1 Novel lexicons

One difficulty for this and any other work on the evolution of the lexicon is to show that the observed properties in real-world lexicons go beyond what might be expected from chance. A good example of this is the fact that random strings appear to display Zipf's *law of abbreviation* (Miller et al. 1958 but also see Ferrer-i Cancho and Elvevåg 2010), making the mere presence of a relationship between word probability and length less interesting. To rigorously prove any aspect of lexical optimization, it is very important to construct a suitable, language-like baseline to compare against (for more discussion, see Moscoso del Prado Martin 2013).

In this work, I compared the real-world lexicons against a baseline created by a large set of novel lexicons. These novel lexicons simulate possible alternative ways to structure the lexicon, while being sensitive to per-language word formation requirements, e.g., phonotactics. It is my hope that the results presented here are made more compelling because of this conservative baseline, as they show aspects of the lexicon that would be found in a large

random set of strings.

I chose to use two different methods to create baselines, since each test focused on a different aspect of lexical optimization. Phonotactic n-gram models (e.g., Dautriche 2015; Meylan and Griffiths 2017, for more detail, see Chapter 2.0.2) are able to create a much larger set of novel lexicons, in that both the relationship between word probability and form is altered and the set of forms. On the other hand, probability-shuffling (e.g., King and Wedel 2020, for more detail, see Chapter 2.0.2) creates a relatively smaller potential set of alternative lexicons, though it is more conservative in that it only alters the association between probability and form; all word forms in a novel probability-shuffled lexicon will be found in the original, real-world lexicon.

In the tests in Chapter 3, I chose to use phonotactic n-gram models to generate novel word forms. Though this method was less conservative, it was the better choice for these tests. A primary reason for this was that the statistical models fit on probability-shuffled lexicons were saliently different than those for the real-world lexicons (see Fig. 5.1), making comparison problematic. Specifically, far fewer contrasts remained following the removal of the Zipfian tail, meaning that the resulting correlation tests were performed over fewer data points[1].

As an aside, this is perhaps evidence in and of itself that the lexicon

[1]Recall that for these tests, I removed the least probable contrasts until 95% of the total probability mass remained, mitigating a structural correlation between type and token frequency found primarily in the Zipfian tail.

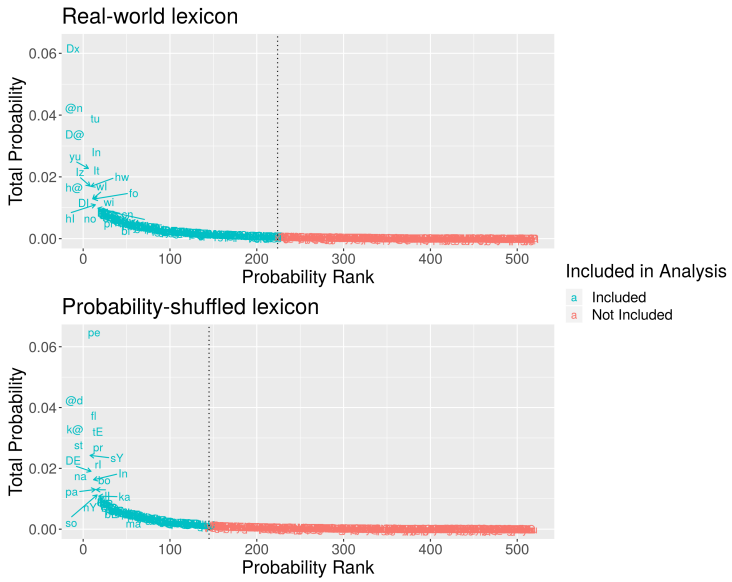

Figure 5.1: Comparison between the number of biphones removed in the real-world lexicon and one example probability-shuffled lexicon. On average, probability-shuffling resulted in more biphones needing to be removed, making it less ideal for creating a baseline.

is structured for more balanced contrasts overall. The fact that a greater number of contrasts are falling in the bottom 5% after probability-shuffling means that disrupting the specific relation between type and token frequencies creates a *more* Zipfian distribution. In other words, the resulting lexicon possesses a larger disparity between high and low frequency contrasts. On the other hand, because the original lexicons are 'less' Zipfian, they are by definition more balanced.

In addition, the less conservative nature of creating comparison lexicons via phonotactic models was lessened by the nature of the tests where they were used. In Chapter 3.1, I looked specifically at word-initial contrasts. For

these tests, each generated novel form does not need to be completely 'legal' given the language's morpho-phonological system. Rather, only the overall set of word-initial contrasts needs to satisfy the word-formation constraints of the language, e.g., English words cannot begin with [ŋ]. Because all novel forms were generated by type-based models for each language, the set of word-initial contrasts were guaranteed to be more or less equivalent to the real-world lexicons. In Chapter 3.2, I looked at the relationship between cohort probability and entropy. For all novel lexicons, I re-calculated the same control variables of cohort size and position as I did for the real-world lexicon. Because the linear model for each novel lexicon was fit with the same control variables, any possible effects that would result from each lexicon possessing different sets of word forms were lessened. That is, if the effect of cohort probability in the real-world lexicon was a product of collinearity with a control variable, the same relationship should occur in the alternative lexicons. That was not the case.

Put together, I felt that it was better to compare models fit over equal numbers of contrasts in the real-world and comparison lexicons, rather than have a more conservative method for creating novel lexicons, especially considering that the specifics of each test mitigated any effect of having a less conservative comparison lexicon.

5.2 Different shapes of information in words

One way to interpret the overall results here is that the lexicon preferentially constructs word forms to have smooth information 'slopes' on average. What this means is that overall, each segment (or any contrastive material for that matter) in the forms of the lexicon reduces the set of competing words at more or less the same rate. At a glance, this appears to contrast with the findings in King and Wedel (2018, 2020), which demonstrate that less probable words preferentially distribute proportionally more information to early positions. Here, I show that overall, word forms have a more uniform distribution of word-internal information than would be expected otherwise. Yet, these are not actually contradictory, and rather further evidence for the word forms being shaped for their use in communication.

For the sake of an example, consider the words *thwart* and *story* in English (see Fig. 5.2). Because of the rarity of the initial [θw] cluster in English, the entirety of the segment information in *thwart*[2] is condensed in the onset while there is a more gradual, uniform distribution for *story*. As it happens, *story* is a much more probable word, meaning that it is more likely to occur in communication than *thwart*. Therefore, the information in the segments of *story* has a greater effect on the aggregate distribution of information in the lexicon as a whole. Put another way, though *thwart* does have an

[2]This is when segment information is determined strictly as the $-\log_2$ contextual probability of a segment given the previous segments in the word, compared to the rest of the lexicon.

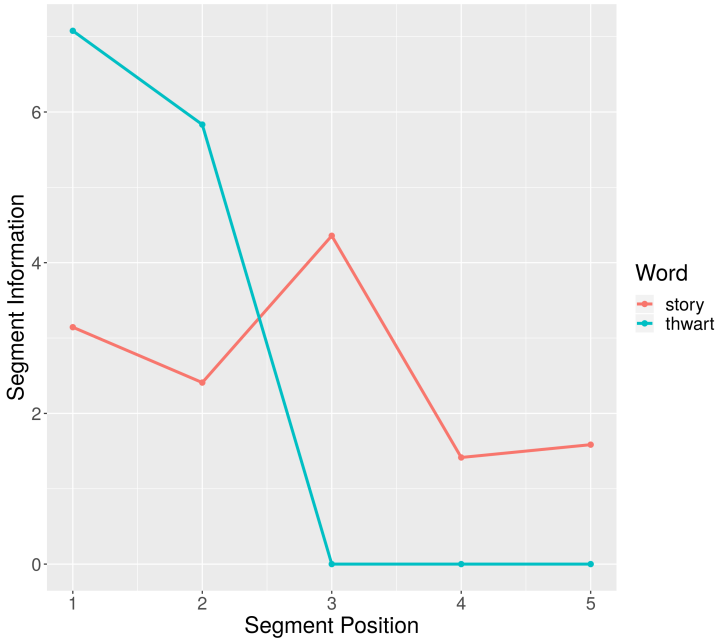

Figure 5.2: Distribution of type-based segment information in two English words. Notice that the less probable word *thwart* has a much steeper slope of information than the more probable *story*.

abnormally nonuniform distribution of information, because it is a relatively low probability word, it does not interfere with the lexicons overall uniformity as much as a more probable word would.

In addition, the disproportionately early distribution of information in less probable words has a benefit by itself in that it can potentially aid in recognition of these otherwise less expected words (for discussion, see King and Wedel 2020). This greater proportion of early information gives less probable words steeper 'slopes' of word-internal information. This predicts that probable words should have smooth information 'slopes', while less probable words have steeper 'slopes', which is supported by the results found here.

5.3 Uniformity beyond word forms

Perhaps one of the most interesting results that comes out of the work here is that the assumption of word-internal information being more or less uniformly distributed within words is accurate. For example, Piantadosi et al. (2011) expand Zipf's law of abbreviation, noting that word length is more strongly correlated with average word-level contextual probability than with raw word frequency. They argue that this causes larger linguistic sequences to be more uniformly informative, as would be expected from theories like *uniform information density* (e.g., Jaeger 2010), assuming that each part of words are themselves equally informative. As shown here, the lexicon is structured so that word-internal information is more uniformly distributed throughout a form, which stands as an interesting element of support to their and other works that assume that information is distributed uniformly through words.

That being said, though word forms have a more uniform distribution of information throughout their segments than might be expected, many words have more informative segments on average than others (see Fig. 5.3). Assuming that language is constructed so that sequences of word forms have a more or less uniform distribution of information, the fact that there is such variance between words seems problematic. However, the particular equation for segment information used here and in other works (e.g., van Son and Pols 2003; Cohen Priva 2017) is restricted to only word-internal context. Put

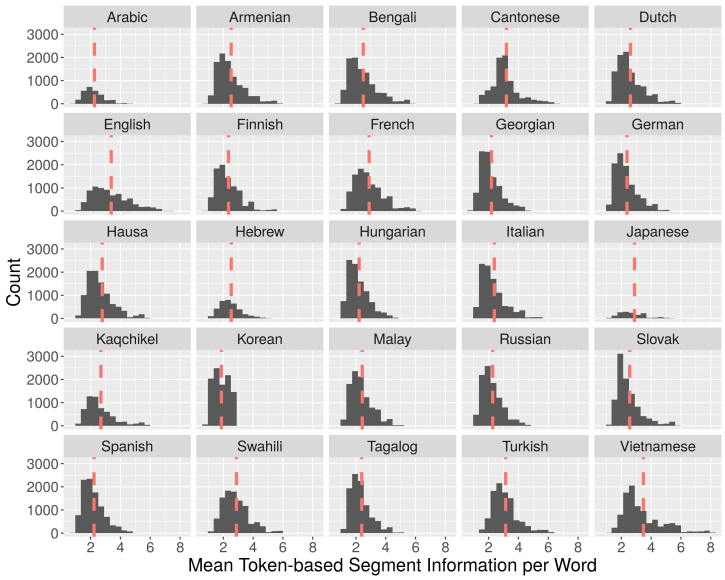

Figure 5.3: The distribution of average token-based segment information in bits per word for each language. Dotted red line shows mean per language. Each language shows a wide range of values.

another way, for these equation the grander linguistic context does not play a role in determining how much information a segment imparts on a listener. This is of course a simplification that makes it more feasible to perform a cross-linguistic study for languages of very different resources.

I suspect that if segment information were to be determined using larger contexts, particularly those outside of the word, the variance in segment information across words would be reduced, which would in turn yield a smoother distribution of segment information across sequences of words, as would be expected by the theory of *uniform information density*.

5.4 Why 'optimized' but not 'optimal'?

Within this dissertation, I have argued for the lexicon to show effects of optimization for efficiency and to be *closer* to optimal, but not 'optimal' per se. By its very definition, an optimal communication system is better than any sub-optimal system, even if the sub-optimal systems are significantly closer to the optimum than would be expected otherwise. This raises the interesting question of why languages are not at that optimal point.

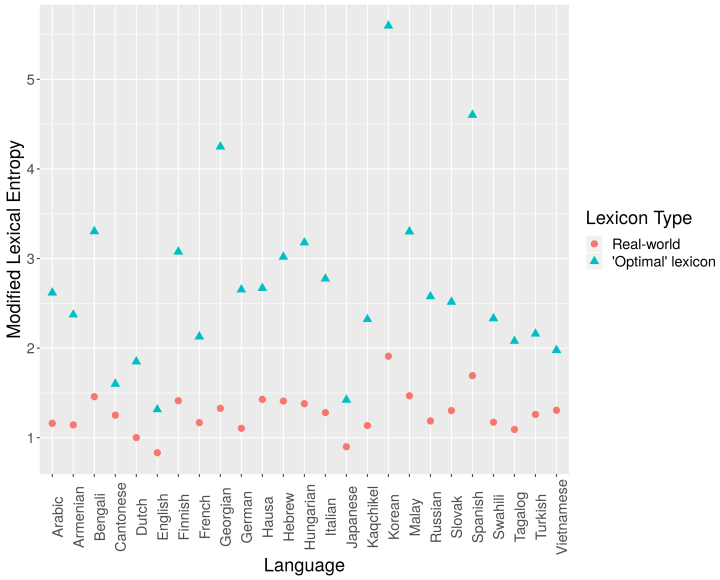

Figure 5.4: Comparison of modified total lexical entropy in the real-world lexicon against an 'optimal' form, i.e., one with perfectly uniform distributions of segment information within words. All languages fall below the theoretical optimum.

Firstly, my model of what constitutes optimality, like all models, is 'wrong, but useful' (Box, 1976). It is a simplified account of how languages are ac-

tually used by speakers and listeners and, as such, is an imperfect (though accurate) representation of linguistic communication that makes it possible to study the lexicon. Because there is some difference between the actual and simplified model representation of words, even if the actual lexicon were 'perfect', it would unlikely be equivalent to the theoretical optimum for the simplified versions of the lexicon.

Secondly, the lexicon is a constantly evolving code and forms of words at a particular moment in time are results of several pressures over several generations. Because of this, if language fails to be the most optimized code at a given moment in time, it may be because it is in the process of approaching optimality in one aspect or another. As an abstract example, imagine if a language's lexicon were at the theoretical optimal stage at a given moment. Because of the ever changing social and technological environment for that language's speakers, words are continually being added, lost or having their relative frequency change over time. As soon as the set of word forms changes, the optimality of the lexicon will cease and the lexicon will move into a sub-optimal state.

The process of optimization entails that the lexicon move *closer* to the abstract optimum with every generation. Put another way, were a the lexicon subject to optimization and it had reached at optimal state, the lexicon would rigidly refuse to add, lose or change word forms, simply because doing so would adversely affect the overall systemic optimality. If this were the case, the language would not be able to adapt to changing communicative needs of

its speakers. In a way, this can be compared to the concepts of *bias-variance trade-off* or *over-fitting* (Friedman et al., 2001). Perhaps, it is better that the lexicon be ostensibly sub-optimal, i.e., more *generalized* than *fit*, as that allows flexibility for change.

This reiterates the predictions of Zipf (1949) and Köhler (1987) who argue that the lexicon is a constantly evolving balance of multiple pressures, a balance that is constantly shifting because the communicative needs of speakers is constantly shifting. It may be the case that very fact that language is sub-optimal - by at least the metrics used here - is a sign that it is structured to be a robust system, able to be changed ad hoc when a new communicative need arises. Such a prediction very much merits further investigation.

Bibliography

Allopenna, P. D., Magnuson, J. S., and 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(4):419–439.

Aronson, H. I. (1990). *Georgian: A reading grammar*. Slavica.

Aylett, M. and 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. and Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, 119(5):3048–3058.

Baayen, H. (1992). Statistical models for word frequency distributions: A linguistic evaluation. *Computers and the Humanities*, 26(5-6):347–363.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The celex lexical database (release 2): Linguistic data consortium. *University of Pennsylvania, Philadelphia, PA, USA*.

Balling, L. W. and Baayen, H. R. (2008). Morphological effects in auditory word recognition: Evidence from danish. *Language and Cognitive Processes*, 23(7-8):1159–1190.

Balling, L. W. and Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, 125(1):80–106.

Baroni, M., B. and S. Zanchetta, E. (2009). The wacky wide web: A collection of very large linguistically processed web-crawled corpora. In *Language Resources and Corpora*, volume 43, pages 209–226.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, 113(2):1001–1024.

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

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Bybee, J. L. and Hopper, P. J. (2001). *Frequency and the emergence of linguistic structure*, volume 45. John Benjamins Publishing.

Canavan, A., Zipperlen, G., and Graff, D. (1997). Callhome. *University of Pennsylvania, Philadelphia, PA, USA*.

Chan, K. Y. and Vitevitch, M. S. (2009). The influence of the phonological neighborhood clustering coefficient on spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 35(6):1934.

Chen, Q. and Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological review*, 119(2):417.

Chomsky, N. (1957). *Syntactic structures*. Walter de Gruyter.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. ERIC.

Christiansen, M. H. and Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39.

Cohen Priva, U. (2017). Informativity and the actuation of lenition. *Language*, 93(3):569–597.

Cohen Priva, U. and 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., and Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language*, 32(2):193–210.

Coupé, C., Oh, Y. M., Dediu, D., and Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9):eaaw2594.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory 2nd edition*. Wiley-interscience.

Cutler, A., Weber, A., Smits, R., and Cooper, N. (2004). Patterns of english phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, 116(6):3668–3678.

Dahan, D. and Magnuson, J. S. (2006). Spoken word recognition. In *Handbook of Psycholinguistics (Second Edition)*, pages 249–283. Elsevier.

Dautriche, I. (2015). *Weaving an ambiguous lexicon*. PhD thesis, Université Sorbonne Paris Cité.

Dautriche, I., Mahowald, K., Gibson, E., and Piantadosi, S. T. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive science*, 41(8):2149–2169.

Davies, M. (2008). Coca. corpus of contemporary american english. Available at `https://corpus.byu.edu/coca/`.

de Saussure, F. (1916). *Course in general linguistics*. New York, NY: McGraw-Hil.

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

Dryer, M. S. and Haspelmath, M., editors (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ehrlich, S. F. and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Memory and Language*, 20(6):641.

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

Ferrer-i Cancho, R. (2016). Compression and the origins of zipf's law for word frequencies. *Complexity*, 21(S2):409–411.

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. and Elvevåg, B. (2010). Random texts do not exhibit the real zipf's law-like rank distribution. *PLoS One*, 5(3):e9411.

Ferrer-i Cancho, R. and Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791.

"Finnish Text Collection" (2005). Collection of finnish text documents from years 1990–2000. Available at `http://www.csc.fi/kielipankki/`.

Flemming, E. (2004). Contrast and perceptual distinctiveness. *Phonetically based phonology*, pages 232–276.

Forster, K. (1976). Accessing the mental lexicon. *New approaches to language mechanisms*, pages 257–287.

Frank, A. F. and Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

Frauenfelder, U. H., Baayen, R. H., and Hellwig, F. M. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language*, 32(6):781–804.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Galati, A. and Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, 62(1):35–51.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of experimental psychology: Human perception and performance*, 6(1):110.

Genzel, D. and Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 199–206. Association for Computational Linguistics.

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., and Levy, R. (2019). How efficiency shapes human language. *Trends in cognitive sciences*.

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., and Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological science*, 24(7):1079–1088.

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

Gippert, J. and Tandashvili, M. (2012). Structuring a diachronic corpus. In *Proceedings of the international conference on Historical Corpora*.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the 8th International Language Ressources and Evaluation (LREC'12)*.

Graff, P. (2012). *Communicative Efficiency in the Lexicon*. PhD thesis, Massachusetts Institute of Technology.

Grosjean, F. (1996). Gating. *Language and cognitive processes*, 11(6):597–604.

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(4):643–672.

Hall, K. C., Hume, E., Jaeger, F., and Wedel, A. (2016). The message shapes phonology. *Ms. University of British Columbia, University of Canterbury, University of Rochester & Arizona University*.

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, 32(2):190–222.

Harris, Z. S. (1970). Morpheme boundaries within words: Report on a com-

puter test. In *Papers in Structural and Transformational Linguistics*, pages 68–77. Springer.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press on Demand.

Hayes, B. and Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, 39(3):379–440.

Hilpert, M. (2008). New evidence against the modularity of grammar: Constructions, collocations, and speech perception. *Cognitive linguistics*, 19(3):491–511.

Hockett, C. F. (1967). The quantification of functional load. *Word*, 23(1-3):300–320.

Hoolihan, K. (1975). *The Role of Word Boundary in Phonological Processes*. PhD thesis, The University of Texas at Austin.

Hurskainen, A. (2004). Helsinki corpus of swahili. *Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC*.

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

Jaeger, T. F. and Tily, H. (2011). On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):323–335.

Jurafsky, D. and Martin, J. H. (2014). *Speech and language processing*, volume 3. Pearson London:.

Kanwal, J., Smith, K., Culbertson, J., and 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.

Kessler, B. and Treiman, R. (1997). Syllable structure and the distribution of phonemes in english syllables. *Journal of Memory and language*, 37(3):295–311.

Khurshudian, V. and Daniel, M. (2009). Eastern armenian national corpus. *"Dialog'2009"*.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlỳ, P., and Suchomel, V. (2014). The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). Itri-04-08 the sketch engine. *Information Technology*, 105:116.

King, A. (2018). The lexicon and the noisy channel: Words are shaped to avoid confusion. *Glottometrics 43*, page 58.

King, A. and Wedel, A. (2018). Incremental word processing helps shape the lexicon. Presented at EvoLang XII.

King, A. and Wedel, A. (2020). Greater early disambiguating information for low probability words: the lexicon is shaped by incremental processing. *Open Mind*.

Kleinschmidt, D. F. and Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, 122(2):148.

Köhler, R. (1987). System theoretical linguistics. *Theoretical linguistics*, 14(2-3):241–258.

Köhler, R. (1993). Synergetic linguistics. In *Contributions to quantitative linguistics*, pages 41–51. Springer.

Krajčovič, R. (1988). *Vỳvin slovenského jazyka a dialektológia*. Slovenské pedagogické nakl.

Kurumada, C. and Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in japanese. *Journal of Memory and Language*, 83:152–178.

Landauer, T. K. and Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, 12(2):119–131.

Lee, Y. (2006). Sub-syllabic constituency in korean and english. *Unpublished doctoral dissertation, Northwestern University*.

Leung, M.-T. and Law, S.-P. (2001). Hkcac: the hong kong cantonese adult language corpus. *International journal of corpus linguistics*, 6(2):305–325.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Levy, R., Bicknell, K., Slattery, T., and Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50):21086–21090.

Levy, R. P. and Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*, pages 849–856.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the h&h theory. In *Speech production and speech modelling*, pages 403–439. Springer.

Linzen, T. (2009). Corpus of blog postings collected from the israblog website. Available at `http://tallinzen.net/frequency/`.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *LREC*. European Language Resources Association.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies*

for Teaching Natural Language Processing and Computational Linguistics - Volume 1, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Luce, P. A. and Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and hearing*, 19(1):1.

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., Dittmann, H., Lenci, A., and Pirrelli, V. (2014). The paisa'corpus of italian web texts. In *9th Web as Corpus Workshop (WaC-9)*, pages 36–43. EACL (European chapter of the Association for Computational Linguistics).

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., and 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., and Piantadosi, S. T. (2018). Word forms are structured for efficient use. *Cognitive science*, 42(8):3116–3134.

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

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Marslen-Wilson, W. and Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human perception and performance*, 15(3):576.

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

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

McClelland, J. L. and Elman, J. L. (1986). The trace model of speech perception. *Cognitive psychology*, 18(1):1–86.

McMurray, B., Tanenhaus, M. K., and 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.

Mendonca, A., Graff, D. A., and DiPersio, D. (2009). *Spanish gigaword second edition*. Linguistic Data Consortium.

Meylan, S. C. and 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.

Miller, G. A., Newman, E. B., and Friedman, E. A. (1958). Length-frequency statistics for written english. *Information and control*, 1(4):370–389.

Mollica, F. and Piantadosi, S. T. (2019). Humans store about 1.5 megabytes of information during language acquisition. *Royal Society open science*, 6(3):181393.

Mortensen, D. R., Dalmia, S., and Littell, P. (2018). Epitran: Precision G2P for many languages. In chair), N. C. C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Morton, J. and Jassem, W. (1965). Acoustic correlates of stress. *Language and Speech*, 8(3):159–181.

Moscoso del Prado Martin, F. (2013). The missing baselines in arguments for the optimal efficiency of languages. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Nelson, N. R. and Wedel, A. (2017). The phonetic specificity of competition:

Contrastive hyperarticulation of voice onset time in conversational english. *Journal of Phonetics*, 64:51–70.

New, B., Pallier, C., Ferrand, L., and Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: Lexique^{TM}//a lexical database for contemporary french: Lexique^{TM}. *L'année Psychologique*, 101(3):447–462.

Nooteboom, S. G. (1981). Lexical retrieval from fragments of spoken words: beginnings vs. endings. *Journal of Phonetics*, 9(4):407–424.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52(3):189–234.

Norris, D. and McQueen, J. M. (2008). Shortlist b: a bayesian model of continuous speech recognition. *Psychological review*, 115(2):357.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, 41(5):673–690.

Pate, J. K. and Goldwater, S. (2015). Talkers account for listener and channel characteristics to communicate efficiently. *Journal of Memory and Language*, 78:1–17.

Pellegrino, F., Coupé, C., and Marsico, E. (2011). Across-language perspective on speech information rate. *Language*, 87(3):539–558.

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

Piantadosi, S. T., Tily, H. J., and Gibson, E. (2009). The communicative lexicon hypothesis. In *The 31st annual meeting of the Cognitive Science Society (CogSci09)*, pages 2582–2587.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Radeau, M. and Morais, J. (1990). The uniqueness point effect in the shadowing of spoken words. *Speech Communication*, 9(2):155–164.

Radeau, M., Mousty, P., and Bertelson, P. (1989). The effect of the uniqueness point in spoken-word recognition. *Psychological Research*, 51(3):123–128.

Rayner, K. and Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, 3(4):504–509.

Sak, H., Güngör, T., and Saraçlar, M. (2008). Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In

International Conference on Natural Language Processing, pages 417–427. Springer.

Salasoo, A. and Pisoni, D. B. (1985). Interaction of knowledge sources in spoken word identification. *Journal of memory and language*, 24(2):210.

Schiller, N. O., Meyer, A. S., Baayen, R. H., and Levelt, W. J. (1996). A comparison of lexeme and speech syllables in dutch. *Journal of quantitative linguistics*, 3(1):8–28.

Schriefers, H., Zwitserlood, P., and Roelofs, A. (1991). The identification of morphologically complex spoken words: Continuous processing or decomposition? *Journal of Memory and Language*, 30(1):26–47.

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

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

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379 – 423.

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21.

Siew, C. S. (2013). Community structure in the phonological network. *Frontiers in psychology*, 4:553.

Siew, C. S. and Vitevitch, M. S. (2016). Spoken word recognition and serial recall of words from components in the phonological network. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(3):394.

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Smits, R., Warner, N., McQueen, J. M., and Cutler, A. (2003). Unfolding of phonetic information over time: A database of dutch diphone perception. *The Journal of the Acoustical Society of America*, 113(1):563–574.

Sóskuthy, M. and Hay, J. (2017). Changing word usage predicts changing word durations in new zealand english. *Cognition*, 166:298–313.

Steriade, D. (1994). Positional neutralization and the expression of contrast. *ms., UCLA*.

Stojanović, D. (2013). *Cross-linguistic comparison of rhythmic and phonotactic similarity*. PhD thesis, Doctoral dissertation, University of Hawaii at Manoa.

Strauss, T. and Magnuson, J. S. (2008). Beyond monosyllables: Word length and spoken word recognition. In *Proceedings of the 30th annual conference of the cognitive science society*, pages 1306–1311.

Tang, K. and Bennett, R. (2018). Contextual predictability influences word and morpheme duration in a morphologically complex language (kaqchikel mayan). *The Journal of the Acoustical Society of America*, 144(2):997–1017.

Tomaschek, F., Tucker, B. V., Fasiolo, M., and Baayen, R. H. (2018). Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguistics Vanguard*, 4(s2).

Toscano, J. C., Anderson, N. D., and McMurray, B. (2013). Reconsidering the role of temporal order in spoken word recognition. *Psychonomic bulletin & review*, 20(5):981–987.

Treiman, R. and Kessler, B. (1995). In defense of an onset-rime syllable structure for english. *Language and speech*, 38(2):127–142.

Turnbull, R. (2015). *Assessing the listener-oriented account of predictability-based phonetic reduction*. PhD thesis, The Ohio State University.

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

Uther, M., Knoll, M. A., and Burnham, D. (2007). Do you speak e-ng-li-sh? a comparison of foreigner-and infant-directed speech. *Speech communication*, 49(1):2–7.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from

erps and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3):443.

Van Berkum, J. J., Hagoort, P., and Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the n400. *Journal of cognitive neuroscience*, 11(6):657–671.

Van Berkum, J. J., Zwitserlood, P., Hagoort, P., and Brown, C. M. (2003). When and how do listeners relate a sentence to the wider discourse? evidence from the n400 effect. *Cognitive brain research*, 17(3):701–718.

Van Rossum, G. and Drake, F. L. (2011). *The python language reference manual*. Network Theory Ltd.

van Son, R. and Pols, L. C. (2003). How efficient is speech. In *Proceedings of the institute of phonetic sciences*, volume 25, pages 171–184.

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

Váradi, T. (2002). The hungarian national corpus. In *LREC*.

Vaux, B. (1998). *The phonology of Armenian*. Oxford University Press.

Vitevitch, M. S., Armbrüster, J., and Chu, S. (2004). Sublexical and lexical representations in speech production: effects of phonotactic probability

and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2):514.

Vitevitch, M. S. and Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, 40(3):374–408.

Vitevitch, M. S., Stamer, M. K., and Sereno, J. A. (2008). Word length and lexical competition: Longer is the same as shorter. *Language and Speech*, 51(4):361–383.

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

Wedel, A., Ussishkin, A., and King, A. (2019). Incremental word processing influences the evolution of phonotactic patterns. *Folia Linguistica*, 40(1):231–248.

Weide, R. (2005). The carnegie mellon pronouncing dictionary [cmudict. 0.6].

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

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

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