Research in Language, 2024, vol. 22:4 (270-297) DOI: 10.18778/1731-7533.22.4.02

RECONSIDERING THE RELIANCE ON FUNCTIONAL LOAD: THE ROLE OF PHONETIC DISTANCE IN PREDICTING L2 SEGMENTAL SUBSTITUTIONS

KATE CHALLIS
Iowa State University
kchallis@iastate.edu.pl

ZOË ZAWADZKI
Iowa State University
zawadzki@iastate.edu

EWA KUSZ
University of Rzeszów
ekusz@ur.edu.pl

Abstract
Much research agrees that Functional Load (FL), i.e., the extent to which a phoneme pair distinguishes between different words in a language, is a useful feature to consider in prioritizing phoneme pairs for pronunciation instruction in the second language (L2) classroom. However, FL measures are not always easy to access and are often calculated according to different principles, whereas other more easily observable features exist, including Phonetic Distance (PD), or the degree of physiological similarity between phones in a phoneme pair. One way to evaluate features and their interrelatedness is to use them in a linear mixed effects regression (LMER) model to predict the rate of observed L2 substitutions that are actually made in speech. This study examines the relationship between two measures of FL (Brown, 1988; Gilner & Morales, 2010) and an estimate of PD we devised from 22 unique articulatory features of vowels and consonants in their ability to predict substitutions in the L2-ARCTIC dataset (Zhao et al. 2018) while accounting for other sources of variation. It was found that even when PD had a resolution of only 2 points, it was highly associated with variance in substitution rates, but that the best model included FL and PD measures together. This finding suggests that PD may also be an important consideration when deciding which phoneme pairs to prioritize in L2 instruction.

Keywords: Functional load, phonetic distance, L2 segmental substitutions, pronunciation errors

1. Introduction

It seems intuitive that segmental substitutions (i.e., when a single phoneme replaces another phoneme) should somehow be related both to how frequently that phoneme pair occurs in the language as well as how physiologically similar that phoneme pair is in the oral cavity. The former idea is captured somewhat by the concept of functional load, while the latter refers to phonetic distance, which are the primary features this study explores as predictors of second language segmental substitutions. This section begins by broadly defining these concepts in order to provide context for the literature review. Functional load (FL) refers to the significance of a specific phonemic distinction in a particular language. It quantifies how important one sound contrast is for distinguishing words and conveying meaning. Sounds with high FL, such as /p/ and /b/ in English, can easily change the meaning of a word if substituted (e.g., bin and pin). Although this concept was first mentioned about a century ago by the Prague School of Linguistics (i.e. Jakobson, 1931) and expanded upon by Brown (1988), the complex interactions between FL and phonetic distance (PD) to predict observed segmental substitutions made by L2 speakers of English remain largely ignored (Munro and Derwing 2006; Kang and Moran 2014; Suzukida and Saito 2019; Sewell 2021). Since L2 pronunciation is a complex dynamic system (Liu and Reed 2022), FL and PD cannot be artificially isolated from one another without losing the essence of the phenomena in question (Larsen-Freeman 2017). For example, the sounds in words are not independent of the words themselves, meaning that the frequency of a word (in both the language and a sample attempting to model the language, i.e., a corpus) will have an impact on segmental substitutions in second language speech.

Ahmed et al. (2021) proposed *Phonetic Edit Distance* (PED) as a method to measure the distance between two consonants or two vowels by quantifying their acoustic and/or articulatory properties. Both phonemic distance and phonetic distance affect L2 speaker comprehensibility and the usefulness of L2 pronunciation teaching materials (Schmitt and Dunham 1999; McCrostie 2007). However, rather than use the PED, which was a composite measure and therefore not easily reproducible, our study measured Phonetic Distance by means of comparison of articulatory properties only, as described in detail in sections 2.2.2 and 3.3.

A type of phonological change that refers to the replacement of a sound or phoneme is segmental substitution, which can be conceived as a deviation from normal pronunciation that might negatively impact comprehensibility (Munro and Derwing 2006; Isaacs and Trofimovich 2012; Crowther et al. 2015). Although it is not the subject of our study, it is worth noting that certain segmental substitutions in second language speech are associated with comprehensibility ratings of L2 learners with specific L1 backgrounds (Crowther et al. 2015; Suzukida and Saito 2019: 3). According to Crowther et al., (2015), the negative impact of segmental substitutions in L2 speech on comprehensibility depends on the speakers' L1. For instance, segmental substitutions in L2 speech made by Chinese learners of English were associated with lower

comprehensibility ratings by native speakers of English (Crowther et al. 2015; Suzukida and Saito 2019).

English language teaching (ELT) materials often suggest that certain phonemic contrasts should be prioritized over others without justifying the logic behind the specific prioritizations (Munro and Derwing 2006; Suzukida and Saito 2019; Sewell 2021). The FL of a phoneme pair could be used to justify its prioritization, however, generally ELT material creators lack access to the resources (i.e., large language corpora) and/or technical knowledge (i.e., corpus-based methodology) necessary to perform the required frequency calculations to obtain FL measurements. While prior research supports the idea that FL and PD can be useful to help prioritization in ELT (Munro and Derwing 2006; Levis 2018; Sewell 2021), there is a lack of empirical, corpus-based research connecting the effect on L2 substitutions observed in the real world.

The purpose of this study is to investigate the relationship that FL and PD have as predictors of phone-pair counts in observed L2-Substitutions. This information will help inform the prioritization of phonemic contrasts in the ESL/EFL classroom.

2. Literature Review

2.1. L2 segmental deviations in speech

Segmental substitutions in second language speech (L2-Substitutions) have been shown to hinder L2 communication, both by decreasing the speaker's *comprehensibility* and *intelligibility* to an interlocutor (Jułkowska and Cebrian 2015; Huensch and Nagle 2021) and by causing the speaker to fail to attend to important, meaning-carrying sound differences (Grant and Brinton 2014; Blasi et al. 2016). Despite its importance, most teachers either do not have sufficient training in teaching pronunciation (Burgess and Spencer 2000) or they lack self-confidence in pronunciation teaching (MacDonald 2002; Couper 2017). Additionally, the so-called 'foreign accent' carries a certain stigma which may lead to problems in social adaptation and navigating the labor market (Gluszek and Dovidio 2010).

According to Munro and Derwing (1998:160), *accentedness* is "the extent to which an L2 learner's speech is perceived to differ from native speaker norms." These deviations can be either segmental (i.e., at the level of individual sounds) or suprasegmental (i.e., at the word or sentence level). A lot of the current research on pronunciation, including accentedness, tends to focus on segmental perception rather than production (Gao and Weinberger 2018:136; Kim et al. 2018; Barrientos 2023). Prior research has shown that consonant substitutions are of crucial importance in terms of the correlation between L2-Substitutions and perceived accentedness (Slowiaczek and Hamburger 1992; Connine et al. 1994; Cutler et al. 2000). When examining specific features of consonants, voice onset time (VOT) duration has been shown to be associated with accentedness in both L2 English speech (Flege and Eefting 1987; McCullough 2013; Gao and Weinberger 2018:136), and liquid consonants L2-Substitutions (Riney and Ota 2000; Gao and Weinberger 2018). On the other hand, results from research on the association of

vowel quality with accentedness are not conclusive. There are several studies that show greater deviations in formant frequency from native speaker norms result in a higher accentedness ratings (Munro 1993; Wayland 1997; McCullough 2013). However, according to Chan et al. (2016), deviations of formant frequency are of less importance for accentedness than the location of a vowel within the vowel space, suggesting that PD could predict accentedness, which is an indirect predictor of L2-Substitutions. Gao and Weinberger (2018) also found that both L2 vowel and syllable structure deviations receive higher accentedness ratings by native speakers than L2 consonant substitutions; in other words, there may be subcategories of L2-Substitutions related to whether they are vowels or consonants.

2.2. Possible Predictors of L2 segmental substitutions

L2-Substitutions also appear to be related to L1 phonological categories ingrained in the L2 learner's mind since their very first months of life (Shi et al. 2006; McQueen, Tyler and Cutler 2012). There are universal linguistic tendencies that, as stated by Neri et al. (2006: 358), "manifest themselves in implicational orders of acquisition". Such models enable L2 learners to become more aware of similarities and differences between the L1 and the target language, and also may predict the extent to which the L2 may present more of a challenge to acquire for a speaker of a given L1 (Neri et al., 2006). These characteristics include an implicit understanding of phone-pair importance (which can be expressed as FL) as well as intuitive recognition and categorization of similar sounding phones as phonemes (which is directly related to PD) (Kissling 2013; Liu et al. 2023). The challenge for L2 speakers is that their L1 intuitions about differentiating phonemes are not always applicable in their L2, which is one source that can lead to L2-Substitutions.

Many other variables affect L2-Substitutions, and these are often more directly observable than the interaction between a learner's L1 and L2 that occurs in the mind. The following section explores some of these variables in greater depth and to explore how they connect to phonetic features of speech at the segmental level.

2.2.1. Functional load

Although functional load (FL) has been diversely defined (Hockett 1955; King, 1967; Catford 1987), it is possible to find common ground within different definitions. FL is generally understood to refer to the importance of a phoneme is given in a specific language. Conversely, it can also be calculated as the number of words that no longer exist when that phoneme is removed (Surendran and Niyogi 2003; Gilner and Morales 2010:136). Although it is not possible to quantify the number of words in a language, words can be quantified in language corpora. Corpora are samples of language designed to represent a target language domain, and as such, are useful tools in being able to empirically illustrate generalizable principles that occur within that domain (Egbert et al. 2022) and are thus able to demonstrate how patterns in the domain can be quantified. Researchers in L2 pronunciation argue that teachers

should prioritize phones with high FL (Suzukida and Saito 2021), however, the practical application of FL to L2 pronunciation teaching and research remains largely ignored (Munro and Derwing 2006; Kang and Moran 2014; Suzukida and Saito 2019; Sewell 2021). This may be in part due to the complexity involved in calculating empirical estimates of the FL of sound contrasts. It may also be because FL can easily be conflated with the similarly-named construct of deep interest to researchers in L2 pronunciation known as cognitive load, which refers to a burden on human working memory (Sweller 2011).

There are a handful of studies which show the usefulness of FL in predicting speech substitutions. For example, Stokes and Surendran (2005:588) showed that FL was the best predictor of age of emergence of consonants in English among English-speaking children. After investigating the link between FL and speech production in L2 adult learners of English, Munro and Derwing (2006:522), revealed that "high functional load errors had greater impact on listeners' perceptions of the accentedness and comprehensibility of L2 speech than did low functional load errors."

One of the seminal studies on FL was done by Brown (1988), who measured FL based on statistics provided by Denes (1963), which in turn were based on a corpus of texts from "a considerable body of conversational material and narrative taken from 'Phonetic Readers'," i.e. written extracts from The Readers Digest. These statistics used the relative frequency of sounds in lexical words to build a frequency-based estimate of FL. In his calculation of FL, Brown (1988) focused on two main concepts: cumulative frequency and the probability of occurrence. Cumulative frequency refers to the total number of times a specific phoneme occurs in a set of recorded speech samples or a corpus. It is calculated by adding individual frequencies of the phonemes which make a specific phonemic pair. For instance, the cumulative frequency for the pair /e, æ/ is 11.05%, which is the sum of individual frequencies of 7.16% for /e/ and 3.89% for /æ/ (Brown, 1988: 597). As Brown (1988) states, pairs with a high cumulative frequency are of greater importance than those with a low cumulative frequency and should prioritized in L2 pronunciation teaching and ELT materials. However, cumulative frequency does not take into account the fact that individual frequency of phonemes in a particular phonemic pair is never evenly distributed. For instance, in the phonemic pair /i:, ɪ/, 21.02% of individual frequency is for /ɪ/, whereas only 4.55% for /i:/. To see the potential L2 segmental error caused by this disproportion in a specific phonemic pair, the probability of occurrence should be calculated. For example, Brown (1988: 597) does this by dividing the individual frequency of phoneme pairs by the cumulative frequency for the pair. For example, /ɪ/ = freq /ɪ/ ÷ (freq /ɪ/ + freq /i:/) = 21.02 ÷ 25.57 = .82.

Contrast-based estimates of FL are based on a change of entropy calculation after the merger of a segment pair in a phonetically annotated corpus (Surendran and Niyogi 2003). In other words, this is a measure of the change in the total number of unique types in the corpus after all instances of one phoneme are

replaced with another. Interestingly, both frequency- and contrast-based measures of FL achieve approximately the same results.

Another source of variation in how FL is calculated is the choice to include or exclude function words; Brown's (1988) frequency-based FL estimates are based only on lexical words, while Gilner and Morales' (2010) change in entropy FL estimates contain the first 10.000 most frequent words of English, including function words. In our study, we explore the relationship between our measure of PD to the two FL two measures presented in Brown's (1988) and Gilner and Morales' (2010) studies, and then explore the relationship PD has to them. One of the motivations for comparing and investigating both studies is that, although it is highly cited and influential, Brown's (1988) FL estimates have some flaws. First, Brown based his measures of FL on a small corpus of written texts; this is unlikely to be the ideal corpus to represent spoken English. Second, his methods for calculating FL are mostly opaque; we do not know how he defined and identified function words, nor do we know how phonemes were labelled and counted. Meanwhile, even though it is not as influential in pronunciation research, Gilner and Morales' (2010: 137) study uses a language sample of almost 10.000 running words, and, more importantly, the sample is spoken language, comprising spontaneous conversation and task-oriented speech, which we believe is more likely to represent spoken English. Moreover, the dataset was taken from Kilgarriff's (1995) list, which was based on the 10-million-word spoken subcorpus of the British National Corpus (BNC), in other words, the methods for calculating FL are mostly transparent.

2.2.2. Phonetic distance

Over the last few decades, researchers have measured phonetic distance (PD), i.e. the degree of difference between two sound contrasts contributing to encouraging results in subfields of linguistics such as dialectology (Gooskens and Heeringa 2004), diachronic linguistics (Covington 1998), and the diagnosis of articulation issues in Speech Language Pathology (Schaden 2006:2442). The base algorithm applied in computing the distance of a phonetic pair of sequences is consistently Levenshtein distance (Schaden 2006; Pucher et al. 2007), also known as *minimum editing distance*.

According to Pucher et al. (2007:1), there is a link between PD measures and word confusion. In their study, they revealed that the measures of phonetic distance can be used in order to evaluate "the quality of grammars and phonetic confusability of words/utterances or interpretations". The results of their study confirmed that phonetic distance measurements can be used for predicting word confusion, as most problems with communication appear with phonetically similar words (Pucher et al. 2007: 5).

According to the contrast and enhancement theory of phonological features, it is useful to distinguish between the physiological features of segmentals that correspond with their respective formant measures (Hall 2011). Formants are

concentrations of acoustic energy (measured in Hertz) around a particular frequency of sound. With vowels, the first formant (F1) is inversely related to the height of the tongue in the mouth, while the second formant (F2) is related to the degree of backness of the vowel (Ladefoged 2006). Thus, it is reasonable to measure phonetic similarity between segmentals based on physiological features such as vocalic quality (voiced/unvoiced), vowel height/length, and manner/place of articulation (Molemans et al. 2012; Wedel, Andrew, Kaplan, and Jackson 2013). Finally, although it is not currently known the degree to which the distance between physiological features is anatomically uniform, Hall (2011) suggests that the psycholinguistic element of human perception of sound is salient when differentiating between sound categories. Although measures of perceptual similarity may also yield salient results, these are much more challenging to measure empirically and do not necessarily attend directly to the research questions of this study, leading to the decision to measure phonetic distance via phonological features, which will be described in depth in section 3 (see Appendix 2 for features).

Seeing that linguistic theory suggests that both FL and PD may have important associations with L2-Substitutionstiwhether we generally consider as an imperfect proxy for pedagogical importance, we designed a study to compare their associations while holding constant several other research-grounded sources associated with increased L2-Substitutions, namely word frequency, phone location in a word, and whether or not the uttered word created a real minimal pair. This design was guided by the following research question: What is the relationship that phonetic distance, functional load measures, and other sources of phonetic variability have with L2 segmental substitution counts?

3. Methodology

This section is organized as follows: first, we provide a concise description of the data used in our study to provide the necessary context to understand the logic behind our chosen methodology.

3.1. Data

The data in this study come from the L2-ARCTIC dataset which is a set of audio recordings containing 26,857 phonetically annotated utterances of speech samples from 24 L2 speakers of English from 6 different L1 backgrounds, namely Hindi, Korean, Mandarin, Spanish, Arabic and Vietnamese[1]. The dataset is comprised of two types of speech samples: read speech and elicited speech. The texts of the read speech task were approximately 1,000 sentences from 19th century

[1] The L2-ARCTIC data can be accessed at https://psi.engr.tamu.edu/l2-arctic-corpus/. All other data generated for our study can be accessed in the following GitHub Repository: https://github.com/kchallis/PD_and_FL_to_predict_L2-Substitutions

literature available on Project Gutenberg (gutenberg.org); these were chosen in order to represent the range of possible phoneme combinations in English for the original ARCTIC corpus, which was one of the earliest datasets used in text to speech (Kominek and Black 2004). For the read speech task, 24 participants (2 male, 2 female per language background) from each of the 6 L1 backgrounds read approximately 300 of these sentences, with intentional overlap between speakers. For the elicited speech task, 22 of the 24 original participants described the action in a series of drawn images known as the suitcase narrative (Derwing et al. 2009). The overall dataset contains 27.1 hours of speech with an average of 67.7 minutes of speech per speaker and utterances averaging 3.6 seconds in duration. Included in the dataset are orthographic transcriptions for all sentences (both read and elicited), TextGrid files with forced-aligned transcriptions using the Montreal Forced Aligner ("GitHub - MontrealCorpusTools/Montreal-Forced-Aligner: Command Line Utility for Forced Alignment Using Kaldi."), and phonetic transcriptions in ARPAbet The reason why the L2 Arctic Dataset uses ARPAbet instead of the International Phonetic Alphabet (IPA) for phonetic transcriptions is because this dataset was designed to replicate the original ARCTIC dataset, which was created during a time before UTF-8-character encodings were widespread; ARPAbet, which is a phonetic code developed in the 1970's using ASCII characters, to use. The ARPAbet symbols are available online (see https://docs.soapboxlabs.com/resources/linguistics/ARPAbet-to-ipThe TextGrids in the L2-ARCTIC dataset contains manual annotations for 19,667 total phoneme insertions (1,174), deletions (3,641), and substitutions (14,852).

For our study, we extracted all annotated substitutions from the L2-ARCTIC dataset using an R script (available on Github). We did not include insertions and deletions because our research questions focus on calculating the phonetic distance between phones, and it was unclear to us how we would calculate a phonetic distance between a phoneme and silence, which is always present in deletions and additions, but never present in substitutions. The R script also extracted the following information from the L2-ARCTIC Dataset for every instance of a phoneme substitution: the participant ID, the participant L1, the task type (read or elicited), the orthographic transcription of the intended word, the intended phoneme, the uttered phoneme, the combined phoneme pair (note that directionality was preserved, i.e. /tʃ, ʃ /was counted separately from /ʃ - tʃ/) and the phoneme substitution location in the word (1 = first phoneme in word, 2 = second phoneme in word, etc.).

Because the data extracted from L2-ARCTIC included an orthographic representation of the target word, we were able to gather data about whether the uttered word created a real English word, i.e. a minimal pair. This was done by using an R script to look up whether the word was present in a large dataset of English words and their ARPAbet phonetic transcriptions known as CMU Dict (The CMU Pronouncing Dictionary, n.d.). If the word was not present in CMU Dict, it was assigned a value of 0 for the field "inDict". Examining the significance of the inDict in our study reveals that its association with L2-Substitutions,

although minor, is statistically significant. This might suggest that L2 learners of English tend to become familiar with the typical phonetic characteristics of English words, or they may simply remember words they have already heard.

3.2 Target Variable

The target variable that we wanted to predict was the count of each occurrence of a phone-pair specific L2-Substitutions in the L2-ARCTIC dataset. To illustrate this concept, we present a small sample from our data in Table 1 with 4 total L2-Substitutions.

Table 1: Count of Phone-Pair Specific L2-subtitution Occurrences in the L2-ARCTIC Dataset

Unique Occurrence of a Phoneme Pair	Intended Word	Spoken Word	Phoneme Pair Count
L-R	full	Fur	3
L-R	full	Fur	3
L-R	peeled	Peered	3
M-N	them	Then	1

In this illustration, there are two instances of the word ‘full’ in which the L-R substitution was made. Every instance of a substitution counts as a separate unit of observation. This is because we hypothesize that there is a relationship between the PD and the count of each phone-pair L2-Substitution. Our predictor variable, i.e. the pair count, was calculated in our R script as follows:

pair_count = number of times that pair count occurred in the substitution data

3.3. Creating the Corpus: Methodology and Procedures

Sounds are parts of words, and words in a language are not normally distributed, but instead follow a Zipfian distribution (Zipf 1932). For this study, we were primarily interested in understanding how PD and FL effect counts of phone-pairs of observed L2-Substitutions; we were not measuring the effect of specific words or word categories (e.g. part of speech, function vs. lexical word) on the total phoneme pair error ratio, nor were we measuring which words or word categories were more likely to contain errors.

However, as noted in section 1, sounds are not independent of words. This means that the count of phone-pairs of observed L2-Substitutions for words that are relatively more frequent could appear to be more important simply due to a frequency bias. This concept is illustrated in Table 2.

Table 2: Frequency Bias in Count of Observed L2-Substitutions for Phone-Pairs

Target Word	Count of Target Word as L2-Substitutions	Total Count of Target Word in Corpus	Normed Frequency for Target Word
the	1050	12,239	85.79
brief	1	1	1000

Notice how in this example, while there are more than 1.000 times more instances of L2-Substitutions for 'the' than 'brief', only 1.050/12.239 (9%) of the instances of 'the' contain a substitution. Meanwhile, the word 'brief' only accounts for one of the total L2-Substitutions, but it was substituted 1/1 (100%) of the time. While the phoneme pair substitution ratio (observed substitution pairs/total substitutions) will always be directly connected to the frequency of the observed sounds (since words are built of sounds), this effect could be greatly lessened by normalizing the number of instances an intended word was mispronounced. In our study we are not directly normalizing the number of phoneme substitutions per word, but this is effectively done indirectly because there is only a limited number of substitutions that can occur in each word. Normed Frequency was calculated according to the following formula:

$$normed\ frequency\ = \frac{occurrences\ of\ intended\ word\ as\ any\ L2\ substitution}{total\ occurrences\ of\ intended\ word\ in\ corpus} \cdot (1000)$$

The final number in the equation is a norming number used for ease of interpretation. To count the number of instances of each word, we used a command line script to extract the transcription text for each utterance into a single, larger text file (the corpus). Although the same sentence was read by multiple participants, it was included in the corpus each time it was read. This text file was then uploaded to Sketch Engine (Kilgarriff et al. 2004), which is an online concordancer software that allows users to build and analyze their own corpora. The statistics page of the corpus then provided information about the total word count in the corpus. Additionally, we tested whether the normed word frequency was predictive of the pair counts by running a general linear regression model. The results of this model show that the normed frequency was weak (R2 = 0.02), therefore it was not considered to be collinear and could be included in the model. NormedFreq had a Zipfian distribution (1932), but log10(normedFreq) had a normal distribution, as illustrated in Figures 1 and 2. It did not make a difference to our model whether NormedFreq was logged or not, so we chose to use log10(normedFreq).

Figure 1: Histogram of Normed Frequency

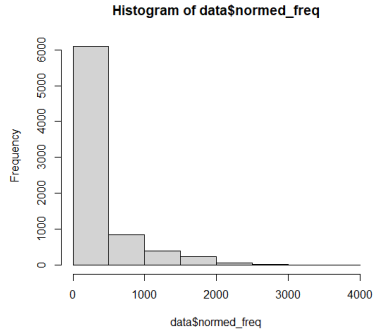
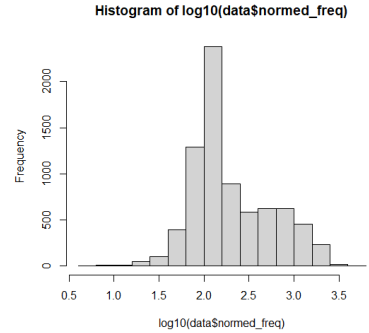


Figure 2: Histogram of Log 10 of Normed Frequency



3.4. Phonetic Distance Calculation Methodology

As described in section 2, phonetic distance is calculable by comparing the number of phonetic features that differ between two phones. While prior research, such as Ahmed et al. (2021), only compared vowels to vowels and consonants to consonants, to maximize the number of possible substitutions included in our analysis, we devised a way to compare all phones (except diphthongs) to one another. This was done in the following manner:

First, we gathered an attested list of phonetic features for vowels, namely: consonantal, sonorant, syllabic, continuant, voicing, labial, round, dorsal, high, low, back, tense, and reduced. This was done by consulting linguistics textbooks, phonetic descriptions based on Dobrovolsky and Czaykowska-Higgin's work (2001), in which features, and natural classes of vowels and consonants were carefully described and presented. Next, we gathered an attested list of phonetic features for consonants, namely consonantal, sonorant, syllabic, nasal, continuant, lateral, delayed release, voicing, closed glottis, labial, round, coronal, anterior, strident, high and back. This was done in a similar way, also by using the classification presented in Dobrovolsky and Czaykowska-Higgins (2001).

We then created a chart containing all the features and all the phones included in the L2-ARCTIC dataset and annotated each phoneme for the presence or absence of the phonetic feature, as illustrated in Figure 3. Cells shaded in green indicate the presence of a phonetic feature, while cells shaded in while indicate the absence of a phonetic feature. Cells shaded in grey indicate that the phonetic feature is not applicable to a certain sound and is also calculated as 0.

Figure 3: Phonetic Feature Annotation Chart for Phones in the L2-ARCTIC Dataset

	vowels												consonants																										
	i	ɪ	e	ɛ	æ	ə	ʌ	u	ʊ	o	ɔ	ɑ/ a	p	b	t	d	k	g	f	v	θ	ð	s	z	ʃ	ʒ	ʧ	ʤ	m	n	ŋ	l	ɹ	j	w	ʍ	h	ʔ
consonantal	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
sonorant	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0
syllabic	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
continuant	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1	1	1	0
voice	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	1	1	0	0	0
labial	0	0	0	0	0	0	0	1	1	1	1	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0
round	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
dorsal	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0
high	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0
low	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
back	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0
tense	1	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
reduced	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
nasal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0
lateral	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
delayed release	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
CG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
SG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
coronal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	0	0	0
anterior	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	1	0	1	1	0	0	0	0	0
strident	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
dorsal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0

After we created this chart, we then created a second matrix comparing the count of phonetic feature differences between phones (which were transcribed to ARPAbet) as illustrated in Figure 4. In this illustration, the closer the PD, the darker the shade of blue.

Figure 4: Phonetic Feature Differences Matrix: ARPAbet Transcriptions of Phones

	IY	IH	EY	EH	AE	AX	AH	UW	UH	OW	AO	AA	P	B	T	D	K	G	F	V	TH	DH	S	Z	SH	ZH	CH	JH	M	N	NG	L	R	Y	W	WH	H	Q
IY	0																																					
IH	1	0																																				
EY	1	2	0																																			
EH	2	1	1	0																																		
AE	3	2	2	1	0																																	
AX	4	3	3	2	3	0																																
AH	3	2	2	1	2	1	0																															
UW	3	4	4	5	6	5	4	0																														
UH	4	3	5	4	5	4	3	1	0																													
OW	4	5	3	4	5	4	3	1	2	0																												
AO	5	4	4	3	4	3	2	2	1	1	0																											
AA	3	4	2	3	2	3	2	4	5	3	4	0																										
P	8	7	7	6	7	8	7	9	8	8	7	9	0																									
B	7	6	6	5	6	7	6	8	7	7	6	8	1	0																								
T	9	8	8	7	8	9	8	12	11	11	10	10	3	4	0																							
D	8	7	7	6	7	8	7	11	10	10	9	9	4	3	1	0																						
K	7	6	8	7	8	7	6	8	7	9	8	8	5	6	6	7	0																					
G	7	6	8	7	8	7	6	6	5	7	6	8	5	4	8	7	2	0																				
F	7	6	6	5	6	7	6	8	7	7	6	8	1	2	4	5	6	6	0																			
V	5	4	4	3	4	5	4	8	7	7	6	6	3	2	4	3	6	6	2	0																		
TH	8	7	7	6	7	8	7	11	10	10	9	9	4	5	1	2	7	9	3	3	0																	
DH	7	6	6	5	6	7	6	10	9	9	8	8	5	4	2	1	8	8	4	2	1	0																
S	9	8	8	7	8	9	8	12	11	11	10	10	5	6	2	3	8	10	4	4	1	2	0															
Z	8	7	7	6	7	8	7	11	10	10	9	9	6	5	3	2	9	9	5	3	2	1	1	0														
SH	8	7	7	6	7	8	7	11	10	10	9	9	4	5	3	4	7	9	3	3	2	3	1	2	0													
ZH	7	6	6	5	6	7	6	10	9	9	8	8	5	4	4	3	8	8	4	2	3	2	2	1	1	0												
CH	10	9	9	8	9	10	9	13	12	12	11	11	4	5	3	4	7	9	5	5	4	5	3	4	2	3	0											
JH	9	8	8	7	8	9	8	12	11	11	10	10	5	4	4	3	8	8	6	4	5	4	4	3	3	2	1	0										
M	7	6	6	5	6	7	6	8	7	7	6	8	3	2	6	5	8	6	4	4	7	6	8	7	7	6	7	6	0									
N	8	7	7	6	7	8	7	11	10	10	9	9	6	5	3	2	9	9	7	5	4	3	5	4	6	5	6	5	3	0								
NG	6	5	7	6	7	6	5	7	6	8	7	7	8	7	9	8	3	3	9	7	10	9	11	10	10	9	10	9	5	6	0							
L	7	6	6	5	6	7	6	10	9	9	8	8	7	6	4	3	10	10	6	4	3	2	4	3	5	4	7	6	6	3	9	0						
R	6	5	5	4	5	6	5	9	8	8	7	7	6	5	3	2	9	9	5	3	2	1	3	2	4	3	6	5	5	2	8	1	0					
Y	3	2	4	3	4	5	4	6	5	7	6	6	7	6	8	7	4	4	6	4	7	6	8	7	7	6	9	8	6	7	3	6	5	0				
W	6	5	7	6	7	6	5	3	2	4	3	7	8	7	11	10	5	3	7	7	10	9	11	10	10	9	12	11	7	10	4	9	8	3	0			
WH	7	6	8	7	8	7	6	4	3	5	4	8	7	8	10	11	4	4	6	8	9	10	10	11	9	10	11	12	8	11	5	10	9	4	1	0		
H	7	6	6	5	6	7	6	10	9	9	8	8	3	4	4	5	6	8	2	2	3	4	4	5	3	4	5	6	6	7	9	6	5	6	9	8	0	
Q	8	7	7	6	7	8	7	11	10	10	9	9	2	3	3	4	5	7	3	3	4	5	5	6	4	5	4	5	5	6	8	7	6	7	10	9	3	0

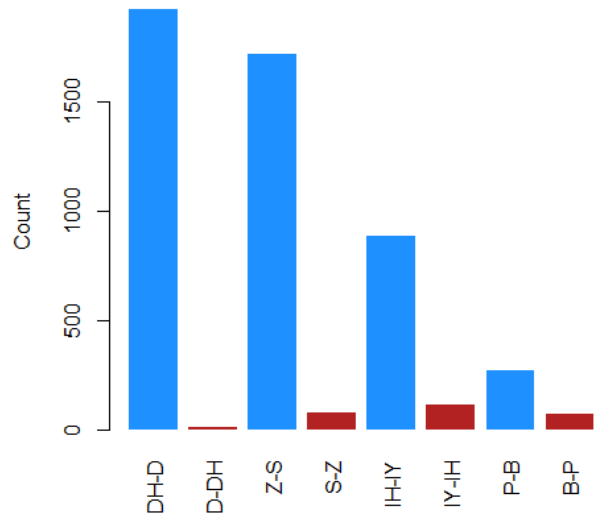
Although we did not include all possible phonetic features, the 22 attested and documented phonetic features chosen created a PD scale with a range from 1 to 13.

3.5. Functional load calculations

As discussed in the introduction, although FL is often discussed in the literature, the methodology for calculating it is variable and often not fully transparent. In fact, the original Brown (1988) study included a chart in the appendix with FL estimates but did not describe the underlying corpus from which the FL estimates were derived, nor the precise methodology for calculating the FL estimates. For this reason, we used the Brown measures from the appendix of the original article, which we then transformed to ARPAbet.

As explained in section 1, Gilner and Morales approached FL estimates by accounting for the most frequent 10.000 words in the corpus, which included function words. Due to the Zipfian distribution of words in a corpus, this meant that the Gilner and Morales FL estimates were quite different from the Brown estimates, and for that reason we felt it necessary to compare both. This was done by using the Gilner and Morales measures discussed in this article, which we then transformed to ARPAbet (see Appendix 1 for full alignment) Each FL estimate for both Brown and Gilner and Morales was bidirectional, or in other words, /s/-/z/ and /z/-/s/ were not calculated separately. However, for our analysis, we preserved the directionality of the phoneme pairs because this is a feature of the observed substitutions. Figure 5 illustrates this feature; if phoneme pair directionality did not matter, we would expect for the count of each of the bar graphs to be the same; they are not.

Figure 5: Phoneme Pair Directionality in a Sample of Observed Substitutions



3.6. Linear Mixed Effects Regression Models

We used linear mixed effects regression models (LMERs) to determine whether there was a difference between how phonetic distance, two different functional load estimates, and all three of these at once would effect the L2-Substitutionstitution counts, while still taking into account other factors such as word frequency, phone location, and if the word that was uttered was actually 'real' (i.e. in the dictionary).

Four regression models with L2-Substitution counts per participant as the dependent variable were run in R (lmer function), as well as a fifth control model. As shown in Table 3, fixed factors for models 1-4 were the phone location within the word and whether word uttered was in the dictionary. Random effects were included for the target word's frequency (the log base ten of the normalized frequency, see section 3.3), participant, and participant L1. We included the additional fixed factors of phonetic distance, functional load estimates based on Brown (1988), and functional load estimates based on Gilner and Morales (2010), where Models 1, 2, and 3 had one factor each, and Model 4 included all three together, as shown in Table 3.

Table 3. Comparison of Fixed and Random Effects across all models

		Control	Model 1	Model 2	Model 3	Model 4
Random Effects	Participant*	✓	✓	✓	✓	✓
	Participant L1*	✓	✓	✓	✓	✓
Fixed Effects, Not of Interest	Word Frequency		✓	✓	✓	✓
	Phone Location		✓	✓	✓	✓
	Uttered Word is in Dictionary		✓	✓	✓	✓
Fixed Effects, of Interest	Phonetic Distance		✓			
	Functional Load estimate (Brown 1988)			✓		✓
	Functional load estimate (Gilner & Morales 2010)				✓	✓

Note. Random effects marked with * were removed during the final phase of model evaluation.

LMERs rely on several underlying assumptions: 1) linearity, 2) bivariate normality, 3) homoscedasticity, 4) independence, and 5) non-collinearity. For the first assumption, SLA theory (see section 2.1) indicates that L2-Substitutions are likely to be related to PD and FL, satisfying the first assumption. Like most corpus data, many of the variables in our data fail to satisfy assumptions of normality, homoscedasticity, and independence. However, LMER models are relatively flexible and robust, meaning they can still provide informative results even when these assumptions are violated. Additionally, the large sample size of our data helps to mitigate some of these violations, as we will now explain.

First, the predictor variable in our model (i.e. phoneme pair counts) is numeric, and not factorized, or in other words, we did not dichotomize this feature into

values of low vs. high; LMERs are able to handle numeric features (lengths, frequencies, counts) implicitly, without needing to perform a repeated-measures analysis of variance (ANOVA). Second, LMERs are "better at handling unbalanced designs (i.e. designs in which not all the experimental situations are equally frequent)" (Gries 2013). As described in section 1, sounds are pieces of words, and words are not normally distributed in natural language. The other variables we included in our model were similarly not normally distributed. Third, LMERs are able to simultaneously account for predictors of interest (i.e. fixed effects) and "the fact that data points are related because they were provided by the same subject or for the same item." (Gries 2013). Another way to say this is that LMERs can handle data that violate the assumption of independence that occurs in hierarchical data structures such as ours.

Finally, we satisfied the underlying assumption of non-collinearity by examining variable correlation between our numeric values via a correlation matrix in R, as shown in Table 4.

Table 4: Correlation Matrix between Numeric Variables

L2-Substitution counts						
InDict	0.37					
Phone Location	-0.07	-0.36				
Normed Word Frequency	-0.07	-0.34	0.49			
FL_Brown	-0.28	-0.29	0.28	0.18		
FL_GM	-0.51	-0.31	0.13	0.04	0.60	
Phonetic Distance	-0.37	-0.18	0.03	-0.06	0.03	-0.51
	L2-Substitution counts	InDict	Phone Location	Normed Word Frequency	FL_Brown	FL_GM

Values close to 1 and –1 indicate a likely existence of collinearity. The highest correlation is 0.6, and is between the two separate measures of FL. Since these were calculated according to very different principles, and since the correlation is still relatively low, we considered this to be an acceptable level to meet the assumptions. For further details about feature selection for our four LMER models, including those features which were deemed insignificant and could thus be dropped, see Appendix 4.

3.7. Training and Testing for Model Evaluation

Each model was run 1) on all the data together, and 2) on the data after partitioning the data at random into subsets for training (75% of the original data) and testing (25% of the original data) using a base R function (R Core Team 2021). This second step allows the model to be trained separately and then run again on unseen data, which helps measure the reliability of the model, or in other words,

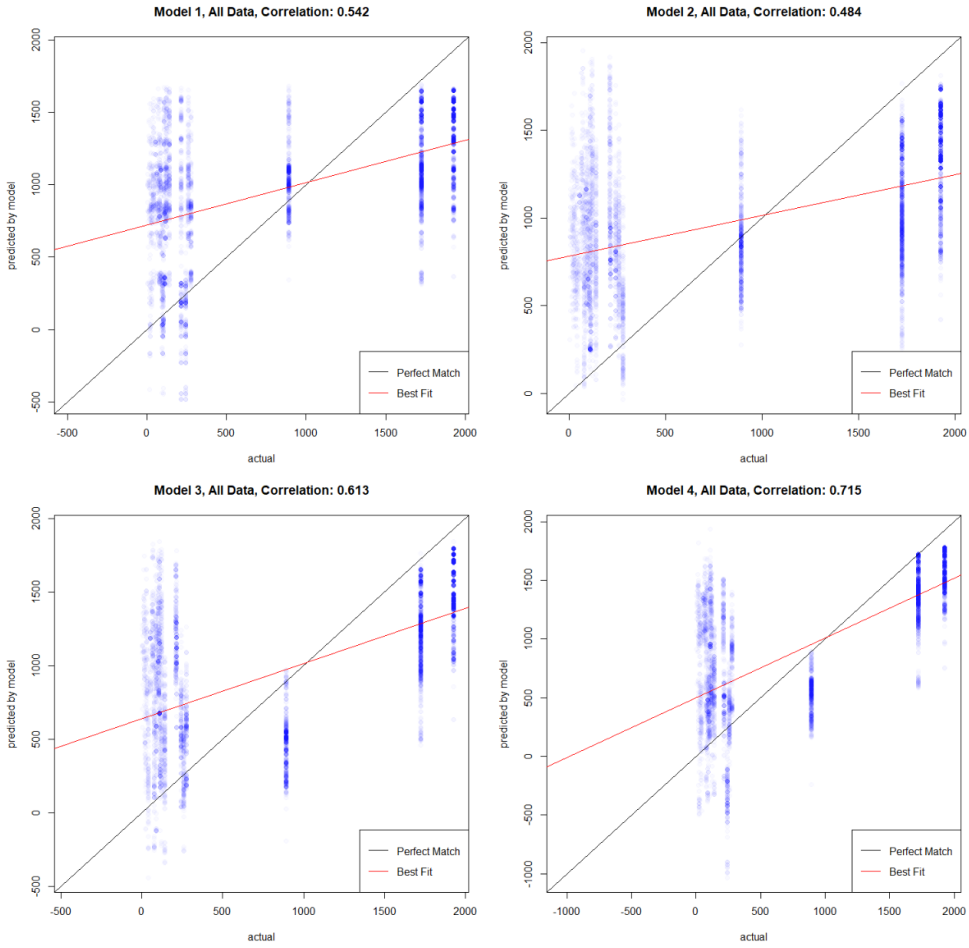
the extent to which the model is overfitting. Evaluating overfitting is not an exact science, and largely depends on the research questions and data at hand. In general, if a model performs significantly better on the training data than on the testing data, it is typically an indication of overfitting.

4. Results

4.1. Results and model evaluation based on all data

When run on all the data, Model 1, which primarily focused on phonetic distance as a fixed effect, the total explanatory power was moderate and had conditional R2 of 0.30, while the part related to the fixed effects alone was 0.21. In other words, the fixed effects were the primary contributors to this model, rather than the random effects of participant and participant L1 background. In this model, the fixed effects with the largest t-values were phonetic distance (t = -30.79, beta = -76.99, 95% CI [-85.88, -68.11]) and the uttered word being in the dictionary (t = 26.45, beta = 0.60, 95% CI [0.56, 0.64]). Neither word frequency nor phone location were significant variables in this model. This can be interpreted as meaning that L2-Substitutions are moderately associated with closer phonetic distances (i.e. similar sounding target and actual phones) and similar sounding words, suggesting a pattern in which L2 speakers may be making substitutions based on something connected to it sounding plausible; perhaps they have heard the word in another context, or perhaps the L2 speaker is applying L2 phonology rules they have intuited. There is moderate correlation (R = 0.54) between the predicted values of model 1 and the actual values from the data, as illustrated for all models in Figure 6.

Figure 6. Model predictions plotted against the actual value



Models 2 and 3, which focused respectively on functional load estimates according to Brown (1988, FL_Brown) and Gilner and Morales (2010, FL_GM), on face value, each had moderate explanatory power. Model 2 had an R2 of 0.23 (marginal R2 = 0.15), and the effect of FL_Brown was significant and negative (beta = -76.99, 95% CI [-85.88, -68.11]), and was the second most important contributor to the model, with a t-value of -16.98. The variable with the biggest effect on the model was actually the uttered word being in the dictionary, which was significant and positive, with a t-value of 28.67 (beta = 536.61, 95% CI [499.92, 573.30]). Although they contributed less to the model, both phone location (t = 5.212, beta = 29.42, 95% CI [18.35, 40.48]) and word frequency (t = 4.28, beta = 89.74, 95% CI [48.66, 130.81]) were also significant and positive. At face value, this model seems to be indicating that words sound contrasts with lower functional load are associated with L2-Substitutions.

Model 3 had a larger R2 of 0.37 (marginal R2 = 0.29) and the effect of FL_GM was significant and negative (beta = -1173.65 [-1224.11, -1123.18]), but by contrast FL_GM was the single most important contributor to the model, with a t-value of -45.59. The next highest t-value was being in the dictionary (t = 21.632, beta = 374.11, 95% CI = [340.21, 408.01]), which seems to show the same pattern as Model 2; lower functional loads and real minimal pairs seem to be associated with L2-Substitutions. This model also had phone location as a significant positive contributor to the model, but the t-value was much lower (t = 4.63, beta = 0.05, 95% CI [0.03, 0.07]).

To facilitate the interpretation of these results, we now compare them to a baseline control model which included none of the variables of interest, but only those others which were added based on theoretical grounding. The fixed effects in the Control Model were word frequency (normalized and logged), phone location, and being in the dictionary, and the random effects were participant and participant L1 (see Table x above). The Control Model had an R2 of 0.21 (marginal R2 = 0.12), which is quite comparable to Model 2, and only moderately worse than Model 3. In the control model, all the fixed effects were significant and positive, with being in the dictionary as by far most important contributor (t= 31.984, beta = 598.04, 95% CI [561.39, 634.70]), while word frequency and phone location had beta values very close to 0. In other words, the Control Model suggests that L2-Substitutions are associated with creating real minimal pairs, and maybe also slightly with phone locations towards the end of the word and higher frequency words. Adding functional load estimates to the baseline Control Model accounts for more variance in L2-Substitutionstitution rates, though the FL_GM estimates have much more power than the FL-Brown estimates, and interestingly seem to have a completely opposite polarity.

This can be interpreted to mean that as the FL_GM increases (e.g. words which contain phoneme pairs with higher FL), the likelihood that the L2 speaker will produce a segmental substitution decreases. This finding was surprising, and we think it might be related to qualities of frequent vs. infrequent words. For example, low frequency words may have more predictable morphology and/or orthography; further research is needed to understand this effect. The opposing polarity is evidence that these two FL measures do not measure precisely the same feature. Seeing that FL measures are not exactly collinear, we created a fourth model including both FL estimates and phonetic distance. Model 4 had moderate explanatory power, with an R2 of 0.51 (marginal R2 = 0.43). The most important contributor to the model was FL_GM, with a significant positive effect (t = -54.48, beta = -1530.82 [-1585.91, -1475.74]), followed closely by phonetic distance (t = -44.98, beta = -976.44 [-1018.99, -933.88]). Table 5 illustrates all the t-values and betas for the significant variables in Model 4; phone location was not significant.

Table 5: Comparison of T-values of Fixed Effects for Model 4

	t-value	**Beta and 95% confidence interval**
FL_GM	-54.48	beta = -1530.83, 95% CI [-1585.91, -1475.74]
Phonetic Distance	-44.98	beta = -976.44, 95% CI [-1018.99, -933.88]
Uttered word in the dictionary	13.225	beta = 208.53, 95% CI [177.62, 239.44]
FL_Brown	4.44	beta = 63.35, 95% CI [54.65, 72.04]
Word frequency	-4.28	beta = -72.47, 95% CI [-105.64, -39.30]

In summary, the model that explains the most variance in L2-Substitutions when run across the entire dataset was Model 4, and it was also the model in which the fixed effects alone, meaning variables unrelated to participants, explained most of the variance.

4.2. Results and Model Evaluation based on partitioned data

As previously stated, we partitioned our data into training (75%) and testing (25%) subsets to evaluate the degree to which models 1-4 could generalize to unseen data. Considering that training models are usually used to predict a variable that typically comes from 'the wild', we decided to include two steps in this evaluation. First, we ran the same exact models as in section 4.1, and discovered that none of the models suffered from overfitting, based on stable R2 and marginal R2 values.

Next, we modified the models so that they only included features which would be easily accessible 'in the wild'. For example, the random effects of participant and participant L1 rely on access to corpus metadata. In the real world, for example in a classroom scenario, these data are very unlikely to be available. These results are presented in Table 6.

Table 6: Model Comparison Summary with Results from Test Data (including random effects)

		Testing Data		Testing Data	
		Includes random effects		Excludes random effects	
	Main Variable of Interest	R2	Correlation	R2	Correlation
Control Model	n/a	0.19	0.44	0.14	0.44
Model 1	PD	0.29	0.54	0.23	0.48
Model 2	FL_Brown	0.23	0.48	0.18	0.42
Model 3	FL_GM	0.37	0.61	0.32	0.57
Model 4	PD FL_Brown FL_GM	0.51	0.72	0.45	0.67

* models were trained and tested on 75% and 25% respectively randomly selected subsets of the original data.

Although none of the models run on the test data experienced large changes in R2 (total variance explained by the model) or marginal R2 (total variance explained by fixed effects alone), the Pearson correlations between the predicted and actual values for all the models decreased. Although this change might be in part due to the test data containing dramatically smaller sample size, it is also way to compare model performance on equal terms. Model 4, which contained PD, Brown FL, and Gilner and Morales FL measures as fixed effects, explained the most variance in counts of phoneme pairs of L2-Substitutionsand also retained the most correlation when run on untrained data that lacked participant information.

5. Discussion and Conclusion

To the best of our knowledge, this is the first empirical study to compare the effect of functional load estimates to the effect of phonetic distance on L2-Substitutionstitition counts, while also considering other factors such as word frequency, phone location, and whether the word that was uttered created a minimal pair. The main finding of this study is that L2-Substitutions seem to associate with phonetic distance and two different kinds of FL estimates.

The implications of this finding are that even though the existing literature emphasizes FL measures in connection to segmental pronunciation teaching, PD is also an important feature to consider because it is nearly as important a variable in explaining variance in L2-Substitutionstitution rates. Another reason why PD might be of value for pronunciation researchers to consider is that it is far easier to understand than functional load. For example, “how similar is the shape of your mouth and tongue when you produce these two phones” is easier to wrap one’s mind around than “how important is one phoneme compared to another for

distinguishing words in a language?" The former question takes an etic approach in which objective criteria, such as the patterns in physiological aspects of phones, are applied universally, while the latter is implicitly emic, since "importance" can be defined subjectively based on the theoretical positions of the researcher. For example, the choice of Gilner and Morales to include function words in their FL estimates, as opposed to Brown's choice to omit them, is an implicitly subjective evaluation of which kinds of words are important in a language.

Additionally, this study shows that Brown FL and GM FL estimates differ in their ability to explain variance in L2-Substitutions. While both FL measures are theoretically grounded, we believe that for the sake of replicability, it would be necessary for future research to recalculate Brown-like FL estimates (aka those without function words) to include 1) transparent reporting of underlying corpora, including a discussion of corpus representativeness, 2) definitions of function words and lexical words, and 3) explicit reporting of all calculations. Still, considering that the original Brown FL estimates were established at a time when corpus-based methods were not possible due to the lack of access to high powered computers and large representative corpora, it is a remarkable tribute to Brown's diligent research efforts and forward-thinking brilliance that this measure can still be used nearly 40 years later as a significant predictor of variation in L2-Substitutions.

A surprising finding of this study is that the functional load in Gilner and Morales' (2010) study usually had a negative effect on predicting L2-Substitutions. This seems to indicate that "unimportant" (i.e. low functional load) sound contrasts have a higher substitution rate, possibly because they are not prioritized by learners since they may be less likely to cause problems with being understood. Future research could investigate the relationship between comprehensibility ratings, PD, and FL measures, and would likely have broader pedagogical importance than merely focusing on substitution rates alone.

While the aim of this study was to explore the relationship between PD and FL, one of the major findings is that consistent predictors of the variance in phoneme pair counts of L2-Substitutionsare whether the phoneme made a real minimal pair and where the phoneme is located in the word. Additionally, much of the variance remains unexplained by the features available to us. We suspect that word length, both in terms of graphemes and phonemes, might affect its proclivity to have L2-Substitutions. We also suspect that the pronunciation of a phoneme is heavily influenced by the phones occurring immediately before and after, which could be another feature to examine in future research. It appears that the directionality of the phoneme pair substitutions (e.g. /s/-/z/ vs./z/-/s/) is not random, and could also be a feature of interest for predicting L2-Substitutions. Finally, one of the most obvious potential sources of variance in L2-Sub that was unaddressed in this study is the degree of orthographic depth of the word, which is certainly a factor for sentences from the 19th century with unfamiliar words with opaque spellings, such as 'debutante', 'maelstrom', and 'physique.' Additionally, while L1 background of participant was treated as a random effect in this study, it is possible that more patterns would be revealed if the models were not run

across all the data at once; in other words, perhaps the story is getting muddled by the variety of patterns in L2-Substitutions made by speakers of different L1 backgrounds. Another possibility would be to consider counts of L2-Substitution phone-pairs by participant, rather than globally.

While this research is a step towards deeper understanding of which phone-pairs are important for L2 learners, it did not address the issue of comprehensibility. The fact is that many – possibly even most - of the substitutions included in this study are unlikely to impede comprehensibility, even though they occur frequently. Furthermore, the CMU Dictionary might have too many examples, that is, it might not be a reflection of the words which English speakers are likely to have in their lexicon.

When considering how to prioritize phone-pairs in pronunciation teaching, researchers must remember that sounds in language always occur within larger contexts of clauses, sentences, and discourse; high frequency of a phone-pair of L2-Substitutions is not necessarily an indication of its importance in comprehensibility. However, we believe that those L2-Substitutions which occur exponentially more frequently also have exponentially more opportunities to impede comprehensibility in diverse ways; future research can examine this hypothesis empirically by providing comprehensibility ratings of the L2-ARCTIC utterances.

In conclusion, this study showed that phonetic distance consistently contributes to L2-Substitutionstitution counts. This may be because L2 learners may struggle to perceive phonemic differences that are allophonic in their L1. This issue can be addressed in a classroom setting through targeted discrimination practice. Additionally, phonetic distance may also be important for teachers to consider because it is likely connected to physiological challenges L2 learners may have when trying to produce certain phones, perhaps lacking muscle familiarity to form phones and phoneme clusters that are less common or prohibited in their L1. This can also be addressed in a classroom setting through awareness raising tasks that involve explicit instruction and emphasis on tongue positioning. What all these potential classroom interventions share in common, however, is the need for corpus-informed data about precisely which contexts are likely to produce what patterns in productions by which speakers, or in other words, studies which focus on producing word lists which can help prioritize pronunciation instruction.

References

Ahmed, Tafsser, Muhammad Suffian, Muhammad Yaseen Khan and Alessandro Boglio. 2021. Discovering similarity using articulatory feature-based phonetic edit distance. *IEEE Access, 10*, 1533-1544. DOI: 10.1109/ACCESS.2021.3137905

Barrientos, Fernanda. 2023. On segmental representations in second language phonology: A perceptual account. *Second Language Research*, *39*(1), 259-285. https://doi.org/10.1177/02676583211030637

Blasi, Damián. E., Søren Wichmann, Harald Hammarström, Peter F. Stadler and Motern H. Christiansen. 2016. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39), 10818-10823.

Brown, Adam. 1988. Functional load and teaching of pronunciation. *TESOL Quarterly* 22(4), 593-606.

Burgess, John and Sheila Spencer. 2000. Phonology and pronunciation in integrated language teaching and teacher education. *System* 28(2), 191-215.

Catford, John C. 1987. Phonetics and the teaching of pronunciation: a systemic description of English phonology. In Morley, J. (Ed.), Current Perspectives on Pronunciation: Practices Anchored in Theory. *TESOL*, Washington, DC, 87-100.

Chan, Kit Y., Michael D. Hall and Ashley A. Assgari. 2016. The role of vowel formant frequencies and duration in the perception of foreign accent. *Journal of Cognitive Psychology, 29*(1), 1–12.

Connine, Cynthia M., Dawn G. Blasko and Jian Wang. 1994. Vertical similarity in spoken word recognition: Multiple lexical activation, individual differences, and the role of sentence context. *Perception Psychophysics, 56*, 624-636.

Couper, Graeme. 2017. Teacher cognition of pronunciation teaching: teachers' concerns and issues. *TESOL Quarterly* 51(4), 820–43.

Covington, Martin V. 1998. *The Will to Learn: A Guide for Motivating Young People*. Cambridge University Press.

Crowther, Dustin, Pavel Trofimovich, Kazuya Saito and Talia Isaacs. 2015. Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly, 49*(4), 814-837.

Cutler, Anne, Nuria Sebastián-Gallés, Olga Soler-Vilageliu and Brit Van Ooijen. 2000. Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & Cognition, 28*, 746-755.

Denes, Peter B. 1963. On the statistics of spoken English. *Journal of the Acoustical Society of America, 35*(6), 892-904. doi: http://dx.doi.org/10.1121/1.1918622.

Derwing, Tracey, Murray Munro, Ronald Thomson, and Marian Rossiter. 2009. The relationship between l1 fluency and l2 fluency development. *Studies in Second Language Acquisition, 31*(4), 533-557. doi:10.1017/S0272263109990015.

Czaykowska- Higgins, Ewa and Dobrovolsky, Michael (2010). Phonology: the function and patterning of sounds. In W. O' Grady, J. Archibald, M. Aronoff & J. Rees-Miller (Eds.), *Contemporary linguistics an introduction*, 59-113. Bedford / St. Martin's.

Egbert, Jesse, Douglas Biber, and Bethany Gray. 2022. *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*. Cambridge: Cambridge University Press. doi:10.1017/9781316584880

Flege, James and Wieke Eefting. 1987. Production and perception of English stops by native Spanish speakers. *Journal of Phonetics*, *15*(1), 67–83.

Gao, Zhiyan and Steven Weinberger. 2018. Which phonetic features should pronunciation instructions focus on? An evaluation on the accentedness of segmental/syllable substitutions in L2 speech. *Research in Langauge*, *16*(2), 135-154. doi:10.2478/rela-2018-0012.

Gilner, Leah and Franc Morales. 2010. Functional load: Transcription and analysis of the 10,000 most frequent words in spoken English. *The Buckingham Journal of Language and Linguistics*, 3, 135-162.

Gluszek, Agata and John Dovidio. 2010. Speaking with a nonnative accent: Perceptions of bias, communication difficulties, and belonging in the United States. *Journal of Language and Social Psychology*, *29*(2), 224–234.

Gooskens, Charlotte, Wilbert Heeringa. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change*, *16*, 189-207. doi: 10.1017/S0954394594163023.

Grant, Linda and Donna Brinton. 2014. *Pronunciation Myths: Applying Second Language Research to Classroom Teaching*. University of Michigan Press.

Gries, Stefan Th.. *Statistics for Linguistics with R: A Practical Introduction*, Berlin, Boston: De Gruyter Mouton, 2013. https://doi.org/10.1515/9783110307474

Hall, Daniel C. 2011. Phonological contrast and its phonetic enhancement: Dispersedness without dispersion. *Phonology*, 28(1), 1-54. doi:10.1017/S0952675711000029.

Hockett, Charles F. 1955. *A Manual of Phonology: Memoir 11*. Baltimore: Waverly Press, Inc.

Huensch, Amanda and Charlie Nagle. 2021. The Effect of Speaker Proficiency on Intelligibility, Comprehensibility, and Accentedness in L2 Spanish: A Conceptual Replication and Extension of Isaacs, Talia and Pavel Trofimovich. Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition, 34*(3), 475-505.

Jakobson, Roman. 1931. Prinzipien der historischen Phonologie, *4. Prague: Travaux du cercle linguistique de Prague*, 246–267.

Jułkowska, Izabela and Juli Cebrian. 2015. Effects of listener factors and stimulus properties on the intelligibility, comprehensibility and accentedness of L2 speech. *Journal of Second Language Pronunciation*, 1(2), 211-237. doi: 10.1075/jslp.1.2.04jul.

Kang, Okim, and Meghan Moran. 2014. Functional Loads of Pronunciation Features in Nonnative Speakers' Oral Assessment. *TESOL Quarterly*, 48, 176–187. doi:10.1002/ tesq.152.

Kilgarriff, Adam. 1995. BNC Database and Word Frequency Lists. Available from http://www.kilgarriff.co.uk/bnc-readme.html.

Kim, Donghyun, Meghan Clayards and Heather Goad. 2018. A longitudinal study of individual differences in the acquisition of new vowel contrasts. *Journal of Phonetics*, 67: 1-20. https://doi.org/10.1016/j.wocn.2017.11.003.

King, Robert. 1967. Functional load and sound change. *Language*, 43(4), 831.852.

Kissling, Elizabeth M. 2013. Teaching Pronunciation: Is Explicit Phonetics Instruction Beneficial for FL Learners? *The Modern Language Journal* 97, no. 3 (2013): 720–44. http://www.jstor.org/stable/43651702.

Kominek, John and Alan W. Black. 2004. The CMU Arctic speech databases. In *Fifth ISCA workshop on speech synthesis (SSW 5)*, 223-224. http://www.festvox.org/cmu_arctic.

Ladefoged, Peter. 2006. *A Course in Phonetics*. California, Thomson Wadsworth Corporation.

Lancaster University (n.d.). UCREL CLAWS5 Tagset. Retrieved April 17, 2023, from https://ucrel.lancs.ac.uk/claws5tags.html

Larsen-Freeman, Diane. 2018. Second language acquisition, WE, and language as a complex adaptive system (CAS). *World Englishes*, 37(1), 80-92.

Levis, John. 2018. Setting Priorities: What Teachers and Researchers Say. In *Intelligibility, Oral Communication, and the Teaching of Pronunciation* (Cambridge Applied Linguistics, pp. 33-58). Cambridge: Cambridge University Press. doi:10.1017/9781108241564.005

Liu, Di and Marnie Reed. 2022 [Manuscript in publication]. From technology-enhanced to technology-based language teaching: A complexity theory approach to pronunciation teaching. In S. McCrocklin (Ed.), *Technological resources for second language pronunciation learning and teaching*. [Publisher unknown].

Liu, Di, Tamara Jones and Marnie Reed. 2023. *Phonetics in Language Teaching (Elements in Phonetics).* Cambridge: Cambridge University Press. doi:10.1017/9781108992015

McCrostie, James. 2007. Investigating the accuracy of teachers' word frequency intuitions. *RELC journal*, 38(1), 53-66.

McCullough, Elizabeth. 2013. *Acoustic correlates of perceived foreign accent in non-native English*. PhD Dissertation. The Ohio State University, Ohio: Columbus.

Macdonald, Shem. 2002. Pronunciation - views and practices of reluctant teachers. *Prospect*, 17, 3-18.

McQueen, James M., Michael D. Tyler and Anne Cutler. 2012. Lexical retuning of children's speech perception: Evidence for knowledge about words' component sounds. *Language Learning and Development*, 8(4), 317-339.

Molemans, Inge, Renate van den Berg, Lieve van Severen and Steven Gillis. 2012. How to measure the onset of babbling reliably? *Journal of Child Language*, 39(3), 523-552.

MontrealCorpusTools. "GitHub - MontrealCorpusTools/Montreal-Forced-Aligner: Command Line Utility for Forced Alignment Using Kaldi." GitHub, n.d. https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner.

Munro, Murray. 1993. Productions of English Vowels by Native Speakers of Arabic: Acoustic Measurements and Accentedness Ratings. *Language and Speech*, 36 (1), 39–66.

Munro, Murray and Tracey Derwing. 1998. The Effects of Speaking Rate he Effects of Speaking Rate on Listener Evaluations of Native and Foreign-Accented Speech. *Language Learning*, 48 (2), 159–182.

Munro, Murray and Tracey Derwing. 2006. The Functional Load Principle in ESL Pronunciation Instruction: An Exploratory Study. *System* 34 (4), 520–531. doi: 10.1016/j.system.2006.09.004.

Neri, Ambra, Catia Cucchiarini, Helmer Strik. 2006. Selecting segmental substitutions in non-native Dutch for optimal pronunciation training, *International Review of Applied Linguistics in Language Teaching IRAL 44*(4), 357-404. doi: https://doi.org/10.1515/IRAL.2006.016.

Pucher, Michael, Andreas Turk, Jitendra Ajmera and Natalie Fecher. 2007. Phonetic distance measures for speech recognition vocabulary and grammar optimization. *3rd Congress of the Alps Adria Acoustics Association.*

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Riney, Timothy, Mari Takada and Mitsuhiko Ota. 2000. Segmentals and global foreign accent: the Japanese flap in EFL. *TESOL Quarterly 34*(4), 711-737.

Schaden, Stefan. 2006. Evaluation of automatically generated transcriptions of non-native pronunciations using a phonetic distance measure. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (LREC'06), Genoa, Italy. European Language Resources Association (ELRA).

Sewell, Andrew. 2021. Functional load and the teaching-learning relationship in L2 pronunciation. *Frontiers in Communication*, *6*, 1–6. doi: https://doi.org/10.3389/fcomm.2021.627378.

Schmitt, Norbert and Bruce Dunham. 1999. Exploring native and non-native intuitions of word frequency. *Second Language Research*, *15*(4), 389-411.

Shi, Rushen, Janet F. Werker and Anne Cutler. 2006. Recognition and representation of function words in English-learning infants. *Infancy*, *10*(2), 187-198.

Slowiaczek, L. M., and Hamburger, M. (1992). Prelexical facilitation and lexical interference in auditory word recognition. Journal of Experimental Psychology: Learning, Memory, Cognition, 18, 1239-1250

Stokes, Stephanie and Dinoj Surendran. 2005. Articulatory complexity, ambient frequency, and functional load as predictors of consonant development in children. *Journal of Speech, Language, and Hearing Research*, 48, 577-591.

Surendran, Dinoj and Partha Niyogi. 2003. Measuring the functional load of phonological contrasts. In: *Tech. Rep. No. TR-2003-12*, Chicago.

Suzukida, Yui and Kazuya Saito. 2019. Which Segmental Features Matter for Successful L2 Comprehensibility? Revisiting and Generalizing the Pedagogical Value of the Functional Load Principle. *Language Teaching Research*, 1–20. doi:10.1177/1362168819858246.

Suzukida, Yui and Kazuya Saito. 2021. Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the functional load principle. *Language Teaching Research*, *25*(3), 431-450.

Sweller, John. 2011. Cognitive load theory. *Psychology of learning and motivation*, 55, 37-76.
Wayland, Ratree. 1997. Non-native Production of Thai: Acoustic Measurements and Accentedness Ratings. *Applied Linguistics*, *18*(3), 345–373.
Wedel, Andrew, Scott Jackson and Abby Kaplan. 2013a. Functional Load and the Lexicon: Evidence that Syntactic Category and Frequency Relationships in Minimal Lemma Pairs Predict the Loss of Phoneme Contrasts in Language Change. *Language and Speech 56*(3), 395-417. doi: doi:10.1177/0023830913489096.
Wedel, Andrew, Abby Kaplan and Scott Jackson. 2013b. High Functional Load Inhibits Phonological Contrast Loss: A Corpus Study. *Cognition*, *128*(2), 179–186 doi:10.1016/j.cognition.2013.03.002.
Zipf, George K. 1932. *Selected studies of the principle of relative frequency in language.* Harvard University Press. doi: https://doi.org/10.4159/harvard.9780674434929.
Zhao, Guanlong, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis and Ricardo Gutierrez-Osuna. 2018. L2-ARCTIC: A non-native English speech corpus, 2783-2787. doi: 2783-2787. 10.21437/Interspeech.2018-1110.

Appendix 1

ARPAbet to IPA Symbols

ARPAbet	IPA	ARPAbet	IPA
AA	ɑ	K	k
AE	æ	L	l
AH	ʌ	M	m
AO	ɔ	N	n
AW	aʊ	NG	ŋ
AX	ə	OW	oʊ
AY	aɪ	OY	ɔɪ
B	b	P	p
CH	ʧ	R	ɹ
D	d	S	s
DH	ð	SH	ʃ
EH	ɛ	T	t
ER	ɝ	TH	ɵ
EY	eɪ	UH	ʊ
F	f	UW	u
G	g	V	v
HH	h	W	w
IH	ɪ	Y	j
IY	i	Z	z
JH	ʤ	ZH	ʒ

Appendix 2

Classification features of consonants and vowels

Consonant	Vowel
Consonantal	consonantal
Sonorant	sonorant
Syllabic	syllabic
Nasal	continuant
Continuant	voice
Lateral	labial
Delayed release	round
Voice	dorsal
Closed glottis	high
Spread glottis	low
Labial	back
Round	tense
coronal	reduced
Anterior	
Strident	
Dorsal	
High	
back	

Appendix 3

Procedure: Feature Selection

Feature selection for our four LMER models (described in section 3) was done by starting with a "saturated" model containing as many potential fixed effects and random effects as possible, and then removing those which did not contribute to the overall variability. This model included fixed effects that were of primary interest and/or were not theoretically explainable by random chance. Random effects were features which were primarily related to the hierarchical structure of the data, but which still contributed to the overall variability. Fixed effects and random effects, as well as their theoretical justifications, are illustrated in Table 7:

Table 7: Potential Features for Predicting Phoneme Pair Counts in L2-Substitutions

Variable	**Type**	**Description**
gm_fl	fixed	Functional load estimate with function words (Gilner & Morales, 2010)
brown_fl	fixed	Functional load estimate without function words (Brown, 1988)
distance	fixed	Phonetic distance we designed based on a chart of phonetic features
phone_location	fixed	The location of the phoneme in a word
inDict	fixed	Whether the uttered word makes a real minimal pair
NormedFreq10*	random	
L1	random	Language background of the participant
Participant	random	Individual participant
wordIndex	Not significant (dropped)	Location of the word in the sentence
modality	Not significant (dropped)	Whether the utterance was read or elicited

*Note that NormedFreq10 represents the log 10 of Normed Frequency.

Next, we examined variance component estimates to assess the contribution of different random effects to the overall variance, and dropped word Index and modality because they did not contribute significantly to the models.