UCL

DOCTORAL THESIS

Measuring phonological distance between languages

Author:

S Elizabeth EDEN

Supervisor:

Prof. John HARRIS

A thesis submitted in fulfillment of the requirements

for the degree of Doctor of Philosophy

in the

Department of Linguistics

2018

Declaration of Authorship

I, Sarah Elizabeth Eden, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Three independent approaches to measuring cross-language phonological distance are pursued in this thesis: exploiting phonological typological parameters; measuring the cross-entropy of phonologically transcribed texts; and measuring the phonetic similarity of non-word nativisations by speakers from different language backgrounds.

Firstly, a set of freely accessible online tools are presented to aid in establishing parametric values for syllable structure and phoneme inventory in different languages. The tools allow researchers to make differing analytical and observational choices and compare the results. These tools are applied to 16 languages, and correspondence between the resulting parameter values is used as a measure of phonological distance.

Secondly, the computational technique of cross-entropy measurement is applied to texts from seven languages, transcribed in four different ways: a phonemic IPA transcription; with Elements; and with two sets of binary distinctive features in the SPE tradition. This technique results in consistently replicable rankings of phonological similarity for each transcription system. It is sensitive to differences in transcription systems. It can be used to probe the consequences for information transfer of the choices made in devising a representational system.

Thirdly, participants from different language backgrounds are presented with non-words covering the vowel space, and asked to nativise them. The accent distance metric ACCDIST is applied to the resulting words. A profile of how each speaker's productions cluster in the vowel space is produced, and ACCDIST measures the similarity of these profiles. Averaging across speakers with a shared native language produces a measure of similarity between language profiles.

Each of these three approaches delivers a quantitative measure of phonological similarity between individual languages. They are each sensitive to different analytical choices, and require different types and quantities of input data, and so can complement each other. This thesis provides a proof-of-concept for methods which are both internally consistent and falsifiable.

Acknowledgements

My thanks to John Harris, James White, and all the other staff and students in the UCL Linguistics department who have contributed so much both intellectually and personally, especially the phonology crew. Thanks also goes to all the researchers at other institutions who have given help and feedback over the years; in particular, many staff and students at SOAS. Thanks for translation, technical support, recruitment and running of the non-word experiments by Martin Rönsch at HHU Düsseldorf, Giorgos Markopoulos at Aristotle University of Thessaloniki and Faidra Faitaki. Finally, especial thanks to Dominic Hunt, Jonathan Oliver and Martin Eden for their help with maths and coding.

Contents

List of Figures

List of Tables

This thesis is dedicated to my husband, Martin. Thank you for putting your dreams on hold so that I could pursue mine.

Chapter 1

Introduction

In this thesis, I address the question: Is it possible to derive a meaningful quantitative measure of phonological similarity between individual languages?

Language similarity is a prominent aspect of any discussion of comparative phonology, but that similarity is usually based on qualitative, not quantitative judgements. I present three different approaches to calculating a metric of phonological language distance.

1.1 Background

Many of the most interesting questions in language differences are questions about rate of historical change. Do phonological systems evolve at the same rate in all isolated speech communities? Do they evolve at a faster rate in speech communities who have contact with speakers of other languages? Are all aspects of a phonological system equally prone to change? Do languages borrow phonological features at a constant rate? Do languages borrow lexemes at a constant rate (e.g. Lees, 1953)? Do creoles evolve at a different rate from other languages (e.g. Mufwene, 2001)?

'Rate of change' as an expression leaves one of the key variables implicit. We want to know how much something has changed per unit of time – but what is that something? What is it we are measuring that we can say has changed? For some of these questions, the answer is relatively straightforward – the rate of lexical borrowing is a measure of percentage of words in some defined vocabulary which change. For others, no clear system has yet been defined.

This is partly due to the vagueness of the term 'similarity'. A 'language' is more or less similar to other languages - but what does that mean? Is it the percentage of shared cognates which is

important (e.g. Lees, 1953), or the phonemic inventory (e.g. Bartelt, 1989, Bardel and Lindqvist, 2006)? This is often left unspecified (as discussed further in Section 2.2), but it is crucial for gaining complete answers to questions about similarity.

Similar questions may arise in the fields of second language acquisition and bilingualism: Does similarity affect the likelihood or amount of transfer from an individual's first language to their second? Does similarity affect the likelihood or amount of transfer from their second to first language? Does similarity affect which previous language is the source of transfer to their third language (e.g. Major, 2008)? To what extent does similarity between languages affect the magnitude of the cognitive effects of bilingualism (Section 2.1)? Does similarity to a first language affect second language production under the influence of alcohol (Nevins, pc.)?

Any scientific explanation of a phenomenon ought to be internally consistent and falsifiable. Individual subjective judgements of similarity, even by professional examiners of language, do not meet these requirements. (See Section 2.2 and Chapter 6 for further discussion.) Furthermore, any claim of similarity or rate of change should by definition relate to a measurable property, so a metric of language distance is required to make such claims.

1.2 Overview

In Chapter 2, I look at some of the potential areas of application for a phonological distance metric. I examine the current distance measurements in use in diachronic linguistics in Chapter 3. In Chapter 4, I present a typological database of phonotactic parameters, and evaluate the success of a parameter-based metric. In Chapter 5, I present a comparison of four phonological representation systems as the bases for a cross-entropy based metric. In Chapter 6, I examine an existing metric of accent distance used in speech recognition, and compare it to the results of Chapter 4 and Chapter 5. Chapter 7 compares the all three approaches.

Chapter 2

Applications of a quantitative measure of language distance

In this chapter, I briefly examine some of the applications of a quantitative measure of phonological language distance. In particular, I look at the fields of bilingualism, second language acquisition, mutual intelligibility and diachronic linguistics.

2.1 Bilingualism

Speaking more than one language has been shown to have cognitive effects in both linguistic and nonlinguistic domains. In the linguistic domain, being bilingual has advantages in, for example, learning new words (Kaushanskaya and Marian, 2009), but disadvantages in e.g. retrieving very-low-frequency words (Michael and Gollan, 2005) and vocabulary size (Bialystok, 2009). In the nonlinguistic domain, being bilingual gives benefits in many aspects of executive function, including inhibitory control (Bialystok, Martin and Viswanathan, 2005) and spatial working memory (Luo et al., 2013).

There are known cognitive differences between monolinguals and bilinguals. However, it is an open question how much of this is a matter of kind and how much a matter of scale. For example, Green, Crinion and Price (2007) examine neural markers of vocabulary knowledge in different speaker groups. They find that the markers which correspond to increased vocabulary in monolingual English speakers are even stronger for English-Italian bilinguals - a difference in scale. However, they also find that English-Chinese bilinguals show other markers which "may reflect additional resources required to process tonal distinctions" - a difference in kind.

If some differences are scalar, rather than binary, we would expect to also see in those cases differences between bilinguals whose languages are more or less similar to each other; e.g. differences between a Spanish-Catalan bilingual and a Spanish-Nahuatl bilingual. But studies attempting to examine the effect of greater or lesser similarity between the speaker's languages are hampered by the lack of an objective measure of linguistic distance. For example, in Bialystok, Luk and Kwan (2005), the authors wish to provide a detailed description of how "the extent to which children transfer their skill in one language to the other language depends on the similarity of the systems, phonological structure in one case and writing system in the other". Yet they lack a method for assessing the similarity of phonological structure, relying on language family as a proxy: "For Spanish–English bilinguals, the languages are similar (Indo-European) and both are written alphabetically in a Roman script; for Hebrew–English bilinguals, the languages are different (Indo-European vs. Semitic)". Assuming that languages from the same family are similar is not always warranted, as we shall see in Chapter 4. From a cursory inspection, we see that unlike English, neither Spanish nor Hebrew have a tense/lax contrast, nor a rounding contrast in their back vowels, nor a velar nasal. Hebrew, like English but unlike Spanish, does not have a palatal nasal, does have a palato-alveolar fricative, and has initial sC clusters (Bolozky, 2006). So it is not immediately and unquestionably apparent that Hebrew and English are more phonologically dissimilar than Spanish and English. A more systematic approach is required to establish phonological distance between these languages.

Furthermore, since there is no clear divide between dialects and languages (Fishman, 1977), there is no clear divide between bidialectalism and bilingualism. Claimed cognitive effects of bilingualism "may also be attenuated or aggravated by factors operating within monolinguals, such as using different dialectal varieties of a language. To date, little is known about the cognitive demands imposed by dialect use" (Kirk et al., 2014). That is, models of bilingualism which ignore dialectal variation assume that the difference is one of kind - and that speakers of multiple 'dialects' are one kind, and speakers of multiple 'languages' are another. A metric of phonological distance could establish a threshold for treating speakers as belonging to the same kind for the purposes of phonological comparison.

2.2 Second language acquisition

How linguistic similarity affects performance has been a topic of great interest in second language acquisition research (Major, 2008), particularly the effects of a first language (L1) on a second (L2), but also the effects of L2 on L3 (e.g. Flynn, Foley and Vinnitskaya, 2004, Rothman, 2011) and L2 on L1. However, similarity has frequently been poorly defined. In many studies, genetic similarity is assumed to be the same as typological similarity, which is assumed to be the same as consensus judgements on how easily speakers of one language acquire the other[1] (e.g. Ahukanna, Lund and Gentile, 1981, Selinker and Lakshmanan, 1992, Cenoz, 2001, Bardel and Lindqvist, 2006). An explicit statement of this position can be found in Corder, 1979:

> "There are of course technical and theoretical problems in establishing and measuring degrees of language distance, but the assessment of the learning task undoubtedly correlates with some notion of genetic relatedness as established by studies of language typology...
>
> I suggest... that the collective experience of a community of learning different foreign languages does lead to a reasonably realistic assessment of the relative magnitude of the learning task of acquiring any particular foreign language, and that this largely corresponds to the formal linguistic relatedness of the languages in question to the mother tongue."

There are some obvious criticisms of this assumption – languages may be typologically similar in some aspects whilst being completely unrelated historically; differences in writing system or cultural factors may impact ease of acquisition; ease of acquisition is not necessarily symmetrical between the two languages; and so on.

Even if genetic similarity alone is used, and can be established to be a relevant factor in SLA independent of the other types of similarity, it is still of limited use as a metric. Phylogenetic distance can only be measured relative to other languages within the same family, meaning acquisition of languages not in that limited set can only be treated uniformly.

[1] Examples of consensus judgements on how easily students learn different languages include hours of study required by English speakers to gain proficiency from the US Foreign Service Institute (*Interagency Language Roundtable* 2015), or different rates of 'language proficiency allowance' from the British Foreign Service depending on the difficulty of learning the language (Corder, 1979).

2.2.1 Psychotypology

An alternative measure of similarity, not used interchangeably with the others, is *psychotypology*, which is the individual learner's perception of how similar their languages are. This may have a much larger impact on their willingness to transfer words and concepts than the other types (Cenoz, Hufeisen and Jessner, 2001). However, there does not seem to have been a systematic study of it; Bardel and Lindqvist (2006) argue that psychotypology is unique to the individual, and a more global psychotypology therefore cannot be established. There is as yet no established correlation between any individual psychotypology and other types of language similarity. It is perhaps assumed that learners make the same assumptions discussed above, and base their judgements on some combination of typological similarity and ease of acquisition, and possibly any meta-linguistic knowledge they have about the languages' history.

2.2.2 Individual phenomena

Finally, a topic of SLA research is the effects of similarity between individual phenomena in L1 and L2 on production and on acquisition (e.g. pronunciation of interdentals, Lombardi, 2003, use of phrasal verbs, Laufer and Eliasson, 1993). It may be that all similarity-related SLA effects can be accounted for simply by combining the effects of these individual phenomena, and that global similarity does not have an independent effect. However, this does not negate the usefulness of a metric for measuring overall phonological similarity, as a factor which should be controlled for (Major, 2008, p. 83).

2.3 Mutual intelligibility in L2

The effects of language distance on second language phenomena are not limited to acquisition. There have been various studies on the effect of language background on mutual intelligibility, examining whether sharing an L1 with the speaker helps a listener to understand speech in an L2.

The results of these studies have been somewhat mixed. Some studies (e.g. Wijngaarden, 2001, Bent and Bradlow, 2003) found that language background has no bearing on intelligibility,

whilst others (e.g. Wang and Heuven, 2005, Stibbard and Lee, 2006) found that listeners find speakers of the same L1 easier to understand in the L2.

Since the studies do not all use the same set of languages, it is difficult to directly compare their results. 'Shared language background' may be more significant for some pairs of languages than others; a study that found no difference between Norwegian and Swedish speakers in a second language might not be particularly meaningful. Without a metric of phonological similarity of the L1s, it is impossible to control for this factor.

Pinet, Iverson and Huckvale (2011) measured the similarity of speakers' and listeners' accents in their mutual intelligibility study using a measure of accent distance called ACCDIST. This method measures the similarity of the acoustic features of vowels in individual recordings[2]. They found that there was a significant correlation between talker-listener accent similarity and mutual intelligibility. I have therefore decided to compare this semi-acoustic measurement to the phonological metrics which I have developed. For more details, see Chapter 6.

2.4 Diachronic linguistics

There have been a variety of metrics of linguistic distance proposed in the field of historical linguistics, which I examine in Chapter 3, to complement the comparative method which forms the basis of the discipline. An additional metric based on a different set of data can provide additional insights (Longobardi and Guardiano, 2009). Since the majority of metrics used in historical linguistics have been based on cognacy, they are not able to be extended to unrelated languages.

[2] See Section 6.2

Chapter 3

Existing metrics in diachronic linguistics

In this chapter, I will give a brief overview of the comparative method, the principal tool of diachronic linguistics, and approaches to language distance based on its results. I will then give brief overviews of two alternative approaches which can be applied to phonology.

3.1 The Comparative Method

The relationships between languages for which we have no historical (written) record are primarily established using the comparative method. This method is very successful, though not without its limitations.

Its basis is the Neogrammarian hypothesis "sound laws suffer no exceptions" (Brugmann and Osthoff, 1878, in Campbell, 1998, p. 18). That is, diachronic changes in sounds are phonologically regular: all[1] words containing the relevant sound or sound sequence are affected in unison.

Because of this regularity, *sound correspondences* can be established between dialects or languages whose vocabulary is drawn from the same source language. Words with similar meanings are compared to see which sounds in one language correspond to which sounds in the other. If these words are found to be of the same origin, they are called cognates.

From these sound correspondences, the *proto-sound* can be reconstructed. The more reflexes (descendent words) which have a given sound or feature in the specific dialects examined, the more likely that it was present in the ancestral word. There are also universal tendencies

[1] This does not exclude variation in pronunciation of individual lexical items, but those are exceptional.

which affect the likelihood of there having been a particular proto-sound. Firstly, certain inventories are more natural than others; for example, Jakobson (1962, p. 528) challenges the traditional reconstruction of Proto-Indo-European with voiceless, voiced and voiced aspirated stop series on the grounds that "no language adds to the pair /t/-/d/ a voiced aspirate /dʰ/ without a counterpart /tʰ/, whilst /t/,/d/, and /tʰ/ frequently occur without the comparatively rare /dʰ/". Such universal or near-universal implications must be considered. Secondly, some sound changes are more natural than others – assimilation of place or voicing is more likely than spontaneous change unrelated to the surrounding segments. Similarly, certain sound changes are more likely to occur in one direction than the other – a voiceless sound becoming voiced between vowels is more likely than devoicing between vowels, for example.

From proto-sounds and reflexes, the proto-language can be reconstructed. The validity of the comparative method has been proven by its successful application to many language families, including the Romance languages, whose proto-language can be compared to written records of Latin.

For example, let us examine the reflexes of Latin [k] (see Table 3.1).

The Italian sound [k] which begins ⟨capra⟩ *goat* corresponds to the Spanish sound [k] which begins ⟨cabra⟩ *goat*. This is not a coincidence, since the same correspondence holds across multiple lexical items, and across multiple languages. These sound correspondences imply that these words are cognate.

The French sound [k] also corresponds to the Italian/Spanish/Portugese [k] – in some words. In others, the French sound [ʃ] corresponds to their [k]. However, the appearance of this [ʃ] is predictable - it only appears where the Italian [k] precedes an [a].

We conclude that the proto-sound was a [k], and not an [ʃ], for the following reasons: the majority of languages examined have a [k]; the appearance of [ʃ] in French is conditional, whilst [k] in other languages appears throughout; since [ʃ] appears to have been conditioned by [a], the change to [ɛ] in ⟨chèvre⟩ occurred later, so the Italian/Spanish/Portuguese forms of that vowel are more conservative, and may be more conservative regarding [k] too.

Comparing these conclusions to the written evidence we have for Latin, we see that the Italian form is indeed closest to Latin, and the proto-sound was [k], both before [a] and before [o].

TABLE 3.1: Reflexes of Latin /k/

	French	Portuguese	Spanish	Italian	Latin
	ʃ	k	k	k	k (ka)
	/ʃɛvʀ/	/kabɾa/	/kabɾa/	/kapɾa/	
goat	chèvre	cabra	cabra	capra	capra
	/ʃj̃ɛ̃/	/kɐ̃w/		/kane/	
dog	chien	cão	(perro)	cane	canis
	/ʃɑto/	/kaʃtɛlu/	/kastiʎo/	/kastɛllo/	
castle	château	castelo	castillo	castello	castellum
	/ʃɑ̃te/	/kɐ̃tɑʀ/	/kantaɾ/	/kantare/	
sing, chant	chanter	cantar	cantar	cantare	canere
	/ʃɑ̃sɔ̃/	/kɐ̃sɐ̃w/	/kanθjon/	/kantsone/	
song	chanson	canção	canción	canzone	cantus
	k	k	k	k	k (ko, ku)
	/kɔʀ/	/koʀpu/	/kweɾpo/	/kɔrpo/	
body	corps	corpo	cuerpo	corpo	corpus
	/kuvʀiʀ/	/kobɾiʀ/	/kuβɾiɾ/	/koprire/	
cover	couvrir	cobrir	cubrir	coprire	cooperīre
	/ku/		/kweʎo/	/kollo/	
neck	cou	(pescoço)	cuello	collo	collus

Despite its successes, the comparative method is limited in its ability to recover dialectal or social variation (Campbell, 1998, p. 140), or data beyond a certain time depth. Therefore, several methods have been developed which use 'language similarity' to complement the comparative method. In Section 3.2, I examine methods based on cognacy, and in Subsection 3.3.1, an alternative based on synchronic parameters.

3.2 Cognate based similarity

There are several methods which take established cognates as a starting point for computing language distance. Their results may be used in historical linguistic inquiry, or applied to the synchronic problems already discussed.

3.2.1 Lexicostatistics

Lexicostatistics describes the similarity between languages as the percentage of basic cognates which they share. It is primarily used for grouping languages when there is a paucity of data

(Crowley and Bowern, 2010). Both Crowley and Bowern (2010) and Campbell (1998, p. 180) criticise the choice of items in most instances as not being particularly scientifically rigorous; it is difficult, if not impossible, to derive a universal 'basic vocabulary' which corresponds to cultures in both the Arctic and the tropics.

Lexicostatistics has previously been extended to measuring not just the degree of similarity, but the timespan since the separation of two languages, a method called 'glottochronology'. This has largely been discredited (Campbell, 1998), since it rests on the dubious assumption the average retention rate of core vocabulary is constant at around 80% per 1000 years. 'Core vocabulary' is a problematic concept, as we have said; beyond that, the borrowing of core vocabulary may not occur regularly, but in bursts (Crowley and Bowern, 2010); and the exact figure was derived from Lees's (1953) study of only 13 languages with a written history, hence all with a literary tradition, and all from the same language family.

Most problematic is the question of which 20% of the vocabulary changes (Crowley and Bowern, 2010). The same 20% each time, or a different one? After 3000 years, languages which started with identical core vocabularies could be anywhere between 40% and 80% similar, even assuming the constant rate theory is correct.

This criticism is not unique to lexicostatistics; it can be levelled at any method which groups languages based solely on synchronic similarity, such as those in Subsection 3.3.1 and Chapter 5. Such methods may however provide additional insights into the evolution of established language histories. For example, where lexical items have been borrowed, the source languages may be identified from similarities and differences in syntactic and phonological parameters, which do not necessarily exactly match the lexicon.

3.2.2 Cognate distance

Rather than comparing the percentage of cognates shared between two languages, the similarity of the cognates themselves can be measured.

Levenshtein distance, also called edit distance, is a measure of similarity between two sequences of characters, based on the number of insertions, deletions and substitutions necessary to transform one into the other. It can be used to measure the similarity between two cognates, by representing the cognate as a sequence of phonemes. It was applied to dialects of Irish Gaelic

by Kessler (Kessler, 1995), and to Dutch by Nerbonne and Heeringa (Nerbonne and Heeringa, 1997).

McMahon and McMahon (2005) criticise Nerbonne and Heeringa's work for being insufficiently phonetically motivated: treating all differences between segments equally, treating substitution as equal to insertion plus deletion, and providing no framework for matching segments in the event of, for example, metathesis. With Heggarty (2005), they propose a numerical method of measuring the 'phonetic' similarity of individual segments. The reflexes to be compared are aligned using certain features of the ancestor word, such as the order of consonants and vowels, or the presence of nasals. This allows them to compare corresponding segments even when insertions or deletions have taken place. Segment similarity is then measured using a closed set of articulatory and acoustic parameters, similar to SPE-style distinctive features. The core parameters for consonants are those of the IPA classification: location and degree of stricture, and voicing. The parameters are weighted by the number of different options which are cross-linguistically common. For example, two segments having identical voicing is given less weight than two segments having the same place of articulation, since most languages contrast only two types of voicing, but more locations.

The relative similarity of a set of dialects can be established by aggregating the similarity scores of a set of cognates, perhaps chosen from the most common words as established for a principal dialect. Unfortunately, since this method is completely dependent upon cognates, it cannot be extended to unrelated languages.

However, a parallel method can be used on the production of a given text by speakers of different language backgrounds, as I discuss in Chapter 6.

Alternatively, by examining the patterns of occurrence of such features in a text much longer than a single word, comparisons can be made without cognates, as I discuss in Subsection 3.3.2 and more fully in Chapter 5.

3.2.3 Phylogeny

There are several different approaches which use the results of cognate and sound-change identification to generate phylogenies.

Nakhleh, Ringe et al. (Ringe, Warnow and Taylor, 2002, Nakhleh, Ringe and Warnow, 2005, Nakhleh, Warnow et al., 2005) have used shared 'linguistic characters' to generate a 'perfect phylogeny' of Indo-European - that is, a tree with the minimal number of branches, and no duplication of innovations. Their characters are multi-state (not necessarily binary) parameters. One example is a particular merger, which is a binary parameter: did it or did it not occur in each language? Another is a particular meaning, which is a non-binary parameter: which one of several cognates is used for this meaning in each language? These characters are drawn from 'phonological, morphological and lexical evidence', from various criteria, but not aiming at a systematic and/or exhaustive exploration of any single domain. Their technique is quite successful at describing the evolution of a language family which has proceeded in a mainly tree-like fashion.

Gray and others (Gray, Greenhill and Ross, 2007) use the binary presence or absence of individual cognates in a language as their characters, and search for the most probable tree which accounts for the data, called a Bayesian Phylogeny.

Finally, there are programs such as NeighborNet, which simply calculate the number of shared characters between languages and plot the resulting distances as a network, rather than as a tree. Such characters may be drawn from any or all types of linguistic evidence, such as those which are used in generating a Perfect Phylogeny or Bayesian Phylogeny.

Whilst phylogenies are a useful visualisation of hierarchical or clustering structures, they are not intended to provide a numerical measurement of similarity, particularly of relative similarity of non-overlapping pairs of items. Neither of the investigations used solely phonological characters, and therefore neither makes any statement about the similarity of the phonology of the languages, as opposed to their inter-relatedness. However, this is due to the goals of the investigations, rather than any inherent restriction.

3.3 Alternative approaches to language distance

3.3.1 Parametric typology

All of the characters used in generating the phylogenies discussed above are derived from the results of the comparative method. It has generally been held that classifications based instead

on syntactic or phonological parameters have nothing to do with the lexical classification of languages – that is, with the history of languages as established by the comparative method.

Sound correspondences are so unlikely to occur by chance that they are valid evidence of a historical relationship (Ringe, 1992). But it is generally held that typological similarity, particularly phonetic or phonological similarity, being much more likely, does not provide such evidence. For example, the fact that Welsh and Zulu both have a voiceless lateral fricative is not evidence of a relationship between them.

However, more recent work has shown that syntactic typology may provide insights into historical relationships (Nichols, 1992). Longobardi, Guardiano et al. (2009, 2012, 2013) examine the values of 63 syntactic parameters drawn from the nominal domain across 23 languages (primarily Indo-European, some Semitic and some individual). From this typology, they calculate the Hamming Distance between language pairs – effectively the proportion of independent parametric settings which differ between them – and use this to construct phylogenetic trees. These trees are similar, but not identical, to those derived with the comparative method; the differences reflect, at least in part, known contact between people groups. This Parametric Comparison Method (PCM) is claimed to offer valid new insights, casting light on community contacts which are not visible in the lexicon, or on developments which were previously considered too far in the past.

It is commonly accepted that the phonology of a language changes more rapidly than its syntax (though without a consistent metric, this a somewhat empty statement). If true, applying the PCM with phonological parameters will not reveal older history than the comparative method. Nonetheless, it may offer a valid way of talking about language distance without making any claims about history. And since it does not rely on cognates, it offers an alternative avenue of exploration for those situations in which the comparative method cannot be applied, such as predicting the mutual intelligibility of L2 speakers with unrelated first languages.

The application of parameter-based measurements to phonology is discussed in Chapter 4.

3.3.2 Entropy

The typological approach requires a phonological analysis of the language as a whole to be performed. However, it would also be useful to have a metric which only requires a small quantity

of transcribed speech, and not necessarily a sample chosen to be representative of any particular property. For this reason, I am also looking at a metric based on *cross-entropy*.

Juola (1998) derives an Indo-European family tree from the similarity of translations of a written text[2]. The similarity of two strings of characters is calculated using their *cross-entropy*. Cross-entropy is a measure of how effectively a probabilistic model of one text can predict each subsequent orthographic letter of the other. The resulting tree closely aligns with the results of the comparative method.

Juola's experiments were limited to languages which share an orthography, but this technique can be expanded to any representation of speech as a series of discrete characters. The application of cross-entropy to phonology is discussed in Chapter 5.

[2]Translations of the Bodleian declaration, as gathered by the Oxford University librarians, in one experiment, and translations of samples from the book of Genesis in another.

Chapter 4

Nidaba : A segment distribution database for measuring language distance

4.1 Introduction

In this chapter, I investigate typological distance. The scope of the investigation is segmental phonology; in particular, syllable structure and its phonotactic consequences, as well as inventory structure.

In Section 4.2, I discuss mathematical approaches to measuring similarity in parametric or constraint-based systems. I have chosen 52 phonological parameters whose values can either be determined from lexical data or are prerequisites of such phonemic transcriptions.

In order to ensure consistency in the values assigned to parameters, and to provide tools for other researchers in this area, I have constructed a typological database of phonotactic distributions, called Nidaba (Section 4.3 on page 44). Section 4.4 on page 48 is a case study in which Nidaba is used to analyse Sylheti, an Indo-Aryan language. Section 4.5 on page 60 compares Nidaba to existing databases and computational tools. The data available through Nidaba are described in Section 4.6 and Section 4.7.

Section 4.8 on page 65 lists the 52 parameters and their values for 16 sample languages, and Section 4.9 on page 99 contains the resulting distances between language pairs.

4.2 Mathematical approaches

The phonological space which a language can exploit may be described using either constraints or parameters. These define the set of possible derivations, or in a derivation-free theory, the set of possible inputs and/or outputs (Odden, 1995). Optimality Theory is formulated in terms of constraints, for example, whereas Government Phonology and various typological studies such as Hayes (1995) are formulated in terms of principles and parameters.

In the following sections, I discuss the mathematical methods which can be used to measure similarity in constraint- or parameter-based systems.

4.2.1 Constraints and correlation coefficients

Firstly, I examine methods for measuring similarity in constraint-based systems. There are two correlation coefficients that can be used to measure the agreement between sets of ranked items (such as phonological constraints drawn from a universal set). Kendall's tau is a coefficient of concordance: it measures the proportion of pairs of ranked items which appear in the same order (are concordant) in both sets. Spearman's rho can be viewed as a coefficient of weighted concordance. Items whose ranks are inverted contribute more to disorder if their ranks are more different. Both measurements are symmetrical about 0, and range from -1 to +1. However, they do not give the same values except when there is perfect order or perfect disarray. Although Spearman's coefficient is probably more widely known than Kendall's, Kendall's Tau has a more obvious interpretation for linguistic purposes: it directly examines which of a pair of constraints is more highly ranked, without reference to how many other constraints intervene.

Spearman's Rho is defined as:

$$r_s = 1 - \frac{6 \sum d^2}{(n^3 - n)}$$

where n the number of items in a set, d is the difference in rank between each pair of items.

Kendall's Tau for measuring agreement between sets including tied items is defined as:

$$\tau = \frac{S}{\sqrt{(\frac{1}{2}n(n-1) - U)}\sqrt{(\frac{1}{2}n(n-1) - V)}}$$

where

$$U = \frac{1}{2}\sum(u(u-1))V = \frac{1}{2}\sum v(v-1)$$

S is the total score of concordant (+1) and discordant (-1) pairs; u is the number of tied pairs from the first set, v the number of tied pairs from the second, and n the total number of pairs in a set.

Data requirements

If the metric is to be capable of distinguishing accurately between all known human languages, the probability of the parameters having identical values in both sets by chance should preferably be beneath the 5% threshold. To quote Ringe (1992): "resemblances between languages do not demonstrate a linguistic relationship of any kind unless it can be shown that they are probably not the result of chance."

There is a minimum number of constraints required for similarity in rankings to be significant. For Spearman's rho $r_s = 1$ (identically ranked constraints in both languages) to occur with a probability of less than 0.05, five constraints must be used. As the rankings become more disordered, more constraints are needed for r_s to be significant. For example, a moderately strong correlation of $r_s = 0.5$ requires at least 13 constraints to be considered more than a chance result.

4.2.2 Parameters and Hamming Distance

Having looked at approaches for constraint-based systems, we turn to measuring the similarity of parametric descriptions of languages. As we saw in Subsection 3.3.1, Longobardi and Guardiano (2009) do so with the Hamming Distance.

Hamming Distance

The Hamming Distance is the proportion of differently valued parameters:

$$H = \frac{d}{i+d}$$

where d is the number of differently-valued parameters, and i is the number of identically-valued parameters.

Data requirements

For similarity in Hamming Distance to be significant, it must be calculated from at least 15 independent binary parameters, which I derive as follows.

Out of n binary-valued parameters, the probability of k of them sharing values between two languages is:

$$\sum_{1}^{k} \frac{{}^{n}_{k}\mathrm{C}}{2^{n}}$$

Since the subset of parameters which share values is not predetermined, there are multiple different combinations of parameters which could give rise to the same outcome. The probability is therefore the sum of the number of ways of choosing $1...k$ from n (the cumulative binomial probability).

Assuming the number of known human languages to be approximately 7000 (Lewis and Gary, 2017), the 5% probability threshold for identifying individual languages is determined by:

$$\frac{{}^{n}_{k}\mathrm{C}}{2^{n}} < \frac{0.05}{7000} \approx 10^{-5}$$

and the threshold for a "borderline useful" result is:

$$\frac{{}^{n}_{k}\mathrm{C}}{2^{n}} < 10^{-4}$$

The binomial coefficient is symmetrical:

$${}^{n}_{k}\mathrm{C} = {}^{n}_{n-k}\mathrm{C}$$

so the probability of all the parameters having the same value is the same as that of none of them being the same; the probability of only one parameter being identically-valued is the same as only one of them being differently-valued, and so on.

For a simple binary test of whether two languages are the same or not - where a completely identical set of parameters implies that they are - at least 15 parameters are necessary, using these figures.

This minimal parameter set would obviously not be useful in comparing the *degree* to which languages are similar. The greater the proportion of parameters differing in value, the smaller the Hamming Distance, and the larger the size of the parameter set needed for the Hamming Distance to be significant. Longobardi and Guardiano (2009) used a set of 63 parameters, which allows for between 0 and 13 parameters differing in value whilst maintaining significance, assuming that the parameters are all independent. However, this is not necessarily a valid assumption: some parameters are made redundant (or set to a default value) by particular values of others. In fact, only 16 of the 63 parameters have no such dependencies. Longobardi and Guardiano handle this by only including them if they are currently independently set; only a third of the language pairs examined have probabilities low enough to be significant, but with over a hundred pairs, this is still a useful result.

Subsequent experiments using the PCM have used an updated parameter set - for example, Longobardi, Guardiano, Boattini et al. (2012) uses 56 parameters, of unrecorded dependencies. This allows for highly-related language pairs to have up to 10 parameters differing in value, whilst being at a significantly low probability.

4.2.3 Interchangeability of representations

In the abstract, parameters and constraints are logically intertranslatable: to say that three items are ranked A>B>C is the same as "Is A > B? Yes." "Is B > C? Yes." "Is A > C? Yes." [1]

Therefore, whilst one formulation or another may be preferable for explanatory reasons, if a metric of language distance can be produced for one, it will be applicable to both. Since most existing typological data is formulated in terms of parameters, rather than constraints,[2] my implementation in Chapter 4 is likewise based on parametric data. However, there is in principle nothing to prevent grammars based on constraint rankings from being compared using Spearman's Rho or Kendall's Tau, as outlined above.

[1] For more information on translating between constraints and binary parameters, see *comparison sort algorithms* in e.g. The Art of Computer Programming: Volume 3: Sorting and Searching (Knuth, 1973).

[2] c.f. Gordon's (2002) typology of stress

4.3 Nidaba overview

A set of typological parameters used as input to the Hamming Distance metric ideally has the following characteristics: Firstly, there is a reproducible methodology for deciding parameter values, which gives consistent results when applied by different researchers, and is extensible to new languages. Secondly, the parameter set is flexible, and can be adapted to different theoretical positions, so the consequences of those positions for Hamming Distance can be contrasted.

To aid in this, I have written a database and lexical analysis tool, called Nidaba. Its core functions are the search and comparison of segmental patterns in transcribed lexicons.

Nidaba contains wordlists drawn from a variety of sources, which have been transcribed phonemically, either by myself or the original authors. (The principles used in determining phonemic representation for a given analysis of each language are stored in Nidaba, and alternative mappings can be uploaded by other researchers if they prefer another analysis.) For each language where such data is available, the frequency of each lexical item is listed, principally drawn from film subtitle data (see Section 4.7). From this source data, consonant or vowel sequences can be extracted from different positions within the word (see Subsection 4.3.2). The syllabic parameter values can be derived from these sequences. The values of the vowel and consonant parameters are derived from the phonemic transcription chosen.

The values so derived have been manually checked against other sources where these exist, and any discrepancies noted.

4.3.1 Input data

To analyse a language with Nidaba, two sets of input data are required: firstly, a list of lexical items in some transcription system, together with any data the researcher would like to tag items with (e.g. English gloss, part of speech, origin of loan items, frequency in some corpus); secondly, a conversion to IPA transcription.

Initially, this conversion will be a simple phonetic mapping. This stage allows the researcher to confirm the phonetic inventory of their initial transcription, identifying any typographical errors (e.g. [c] in place of [k]). The mapping system can handle combinations of characters, using a longest-match-first approach. This allows for lexicons derived from semi-regular orthographic systems, containing digraphs or loan words which follow different pronunciation rules.

Once a lexicon has been uploaded, the researcher can compare the occurrence of different segments in different positions (word initial, medial and final), which can assist in identifying allophones. Once the researcher has completed a phonemic analysis, the list of lexical items can be retranscribed with a new, phonemic, mapping, for use in further analysis.

By combining word lists with transcription conversions, we derive a 'doculect', a particular documentation of a dialect, which is transcribed in the IPA (nidaba.co.uk/Contents/Doculect). Since a word list can be associated with multiple conversions, this allows a choice of analysis without any data loss; for example, I have chosen not to use the linking R of the DISC transcription in my IPA representation of English, but another researcher can include it in their own analysis by using a different conversion (nidaba.co.uk/Contents/TranscriptionConversion).

4.3.2 Pattern retrieval

Since IPA symbols have static values, they can be pre-assigned place and manner values, and sorted into vowels and consonants[3]. Using pre-constructed regular expressions[4], Nidaba automatically locates certain combinations of word edges, vowels and consonants.

For any given doculect, the researcher can view word initial, medial or final sequences of vowels or consonants. These sequences are displayed with the number of lexical items in which they are found, and a link to all known examples. This latter feature can help in discovering commonalities, such as all examples of a given sequence deriving from the same morpheme.

From this basic overview, more detailed searches can be conducted. The researcher can specify properties of sequences such as length, number of items, or sonority profile; place, manner and/or voicing features; and part of speech or other lexical tags, such as loan words of a particular origin.

Nidaba also generates composite properties for each sequence from the relevant lexical items. For example, if corpus frequency data is available, Nidaba will give the total frequency of a sequence summed over all items.

If lexical items have associated frequency data, Nidaba can produce total and average frequency statistics for any given pattern retrieved. If token frequency is not available - for example, in an unwritten and unbroadcast language - Nidaba also produces the number of items in which

[3] Mapping IPA symbols to a user's feature set of choice is an extension goal.

[4] i.e. search patterns to be located in a longer text

a given sequence occurs in the lexicon, and what those items are. Each sequence is linked to its list of source words, to verify the original context.

By filtering out sequences only found in relatively few lexical items, or with very low frequency, the researcher can exclude noise arising from errors in input data, loan words or regional variants (nidaba.co.uk/Tools/CompareSets). Because this data is not excluded automatically, users can compare marginal sequences - such as [sf] in *sphere* - with non-existent sequences.

Nidaba has a default set of binary features for every IPA segment known to the database, covering place, manner and voicing. These features are not hard-coded, and can be straightforwardly replaced with alternatives; I hope to make this functionality available through the web interface in a future version. These features are available to the pattern retrieval tool, simplifying the task of examining the contexts in which segments appear.

4.3.3 Comparison

The results of the detailed searches can be automatically compared, making it easy to see which sequences occur word-initially but not word-finally; in nouns but not in verbs; or in high frequency items but not in numerous ones (e.g. English [ð], which is the most frequent word-initial consonant, but only occurs in a couple of dozen items).

These comparisons are not limited to a single doculect or even language. As well as customisable sequence set comparisons, Nidaba also has two default comparison pages designed to give a quick overview of the similarities and differences between multiple doculects. The first presents sequences located by a set of default searches, including multi-consonant initial sequences, word final consonants and sonority violating sequences. The second automatically calculates parameter values from the results of such searches, and provides researchers with links to the relevant lexical items.

Finally, Nidaba has a tool for locating subsequences. For example, the researcher can divide all word-medial consonant sequences into sequences also found word-finally ('codas'), and any following consonants ('onsets')[5]. This data can then be fed into the set comparison tool mentioned above, and word-internal and word-edge 'onset' and 'coda' sequences compared.

[5]'Onset' and 'coda' here being terms of convenience for particular subsets of consonant sequences, not commitments to a particular syllabification.

4.3.4 Accessing Nidaba

The principal use case is through a web interface, with data stored centrally and potentially made accessible to other researchers. It is available at the URL nidaba.co.uk.

Data which has been uploaded unrestricted can be viewed by anyone. By contrast, to upload data, users need to register for an account. The uploader then maintains control over the accessibility of their data. They can choose to make their data available to all, or they can share it with only named collaborators.

The software is open source, and is available at bitbucket.org/selizabetheden/nidaba. Users can also download the source code to run a local copy of Nidaba, for use with very sensitive data or without a reliable internet connection. However, this removes access to inter-language comparisons, because the database itself is not downloadable. Uploaders are however encouraged to provide URLs to public domain lexicons elsewhere, which could then be imported into the local copy of Nidaba.

Finally, users may also wish to run a local copy to make custom modifications, but I would prefer to receive suggestions for any useful modifications so that they can be implemented in the main web app.

4.3.5 Further applications

The set of computational tools I have outlined here were primarily designed for collating and analysing data to provide syllable structure parameter values. However, by making every step explicit and configurable, Nidaba has several secondary uses, including in experimental set-up and field work. For example, it can be used to:

- create a set of experimental stimuli from a set of constraints, such as 'words with a minimum frequency of x with branching onsets'
- locate possible errors in transcriptions via unique distributional patterns
- locate cognates using shared glosses or phonemic features
- generate minimal pairs
- collaborate with other researchers during data collection, editing and analysis

- compare the effects of different phonemicisations
- find data, sources and collaborators for new languages

Nidaba contains functions for set comparison. Rather than manually comparing the results of searches, a user can specify two separate sets of criteria, and receive a list of segments which match either or both. The most basic use of this tool is in locating or verifying positional allophones. However, criteria for comparison can also include type or token frequency, part of speech, or other factors which may contribute to variation. Comparisons can also be performed with other doculects.

The segmental properties of a language are not interesting only in isolation, but in how they relate to languages of the same family, or with which they exchange lexical items. Nidaba contains multiple tools to aid in the investigation of cognacy and loanword adaptation.

Using the custom tagging system, lexical items can be glossed in multiple languages. Properties of these glosses can be used in filtering results to establish correspondences between putatively cognate items. Nidaba also contains a "word comparison" tool for comparing lexical items across multiple doculects, based on whole or partial overlap in transcription or gloss.

Nidaba contains a tool for generating minimal pairs. This tool provides examples of all instances in which transcriptions differ by only a single segment. Examples are grouped by contrasting segments, illustrating not just minimal pairs, but minimal triplets or larger sets.

4.4 Case study: Sylheti

In this section[6], I shall demonstrate how Nidaba can be used to analyse a language. I look at Sylheti, an Eastern Indo-Aryan language spoken in Bangladesh, as well as in London and other diaspora communities.

4.4.1 Input data

The input data consists of a lexicon compiled by the SOAS Sylheti Project up to November 2016 (SOAS Sylheti Project, 2015). The lexicon was imported into Nidaba from Fieldworks Language Explorer with minimal editing (e.g. column labelling). Each complete entry contained a Sylheti

[6] Parts of Section 4.4 originally appeared in Eden, in press

transcription, part of speech data, English gloss, and additional tags such as a Bangla ('Standard Bengali') gloss.

Sylheti is for the most part unwritten, with speakers writing in Bangla, the medium of education; token frequencies are therefore not readily available, and not analysed here.

4.4.2 Phonemic analysis of consonants

The following consonants were present in the lexicon, once any typographical errors had been eliminated as discussed above:

TABLE 4.1: Full set of consonants used in Sylheti phonetic transcription

p	b			t̪	d̪	ʈ	ɖ	ʧ	ʤ	k	g			
f				s	z	ʂ		ʃ		x				h
		m			n		ɳ				ŋ			
					l	ɾ			ɽ					

Using Nidaba's pattern retrieval tool, I identified the subset of these consonants found as singletons, not neighbouring any other consonants (Table 4.2). Those consonants not found in all positions (initial, medial and final) are in parentheses; consonants not found as singletons in any position are replaced with a dash.

TABLE 4.2: Sylheti singleton consonants

(p)	b			t̪	d̪	ʈ	ɖ	ʧ	(ʤ)	k	g			
f				s	z	-		ʃ		x				(h)
		m			n		-				(ŋ)			
					l	ɾ		(ɽ)						

Nasals

One example of a positionally-dependent consonant is the retroflex nasal [ɳ], which is only found preceding retroflex stops. Given the relative incidence of homorganic nasal-stop sequences to heterorganic sequences for other nasals, and the complete absence of any alveolar nasal-retroflex stop sequences, I conclude that [ɳ] is an allophone of /n/.

The velar nasal [ŋ] is not found word-initially, and like the other nasals, is most commonly found in homorganic sequences. Whilst found in many fewer items than the labial or alveolar nasal – comparing only instances in medial or final position – I do not conclude that it is an

allophone of /n/. A large proportion of word-medial sequences containing [ŋ] are heterorganic, and the majority of word-final occurrences are in isolation. It is found contrasting with both /m/ and /n/: [gam] *sweat*; [gan] *song* and [gaŋ] *river*.

Retroflex flap and stop

Like the velar nasal, the retroflex flap [ɽ] is also not found word-initially.

By contrast, the voiced retroflex stop [ɖ] is only found word-finally in two items, [blɛɪɖ] *blade* (*of grass*) and [bɛɾɛɖ] *bread*. These are almost certainly borrowed: both items have synonyms with Bangla cognates, and English alveolar stops are borrowed as retroflexes in most Indo-Aryan languages. Nidaba includes a word comparison tool, which locates all items in the selected lexicons which share a (partial) gloss, transcription or orthography.

These two consonants are not quite in complementary distribution in word-medial position. [ɖ] is found word-medially between two vowels in 11 items, whereas [ɽ] is found in 112. [ɖ] is also found following [ɳ] and as a geminate; and in [maɽɖal] *to strain* and in [ɖalɖa] *Dalda*, a brand name. [ɽ] is found preceding [b], [d], [n], [t], [ɖ], [k], and [ʃ]; following [m]; and in [fifɽa] *ant*, [laxɽi] *wood*, [zɔgɽa] *argument* and [lɛŋgɽa] *lame*.

The distribution of these two sounds in Sylheti appears to be similar to that in other Indo-Aryan languages, such as Bangla and Hindi, including the apparent contrast found in loan words (Dasgupta, 2003, Masica, 1991, p. 91 & p. 97, Śa', 2001).

Both of these sounds are found contrasting with the voiceless retroflex stop [ʈ]. For example, [aʈ] *eight* versus [aɽ] (*third*) *month* and [ɖali] *solider* versus [ʈali] *pan*.

Affricates

The postalveolar affricate [d͡ʒ] is not found intervocalically in Sylheti; the Sylheti cognates of Bangla words containing [d͡ʒ] are realised with [z] (Ferdous p.c.). This is the same development found in Assamese and neighbouring Bengali dialects (Masica, 1991, pp. 95–95). With the development of fricative [z] from the voiced stop [ɟ] (via [d͡ʒ]), Sylheti now has a voicing opposition in its fricatives, unlike most Indo-Aryan languages. For example, [sal] *ash* versus [zal] *net*.

Using Nidaba's transcription search, I find that [d͡ʒ] is only present in the contexts [nd͡ʒ] and [d͡ʒd͡ʒ]. Appearances in other contexts are as a variant of [z], possibly Bangla: [xɔɪld͡ʒa] (a variant

of [xɔɪlza] *liver*); [ɾad͡ʒniti] (a variant of [ɾazniti] *politics*); [ʈɔɾd͡ʒɔni] (a variant of [ʈɔɾzɔni] *ring finger*); and as an English loan [sad͡ʒɛʂʈ-xɔɾ] *to suggest*.

[ʧ] is found individually predominantly in loan items: [ʧɛɾi] *cherry*, [ʧɔkɔlɛʈ] *chocolate*, [biʧ] *shore* (*beach*), and [pɾoʧuɾ] *enough*. Like [d͡ʒ], [ʧ] is found in the contexts [nʧ] and [ʧʧ]. Otherwise, it is found only in [lalʧɛ] *reason*, [ʧup] *quiet* and [ʧɔk] *bright*. The vast majority of [ʧ]-initial Bangla glosses in the lexicon correspond to [s]-initial Sylheti items. Nidaba allows filtering of results based on custom tags, returning only items with e.g. [ʧ]-initial Bangla glosses.

The majority of nasal-affricate sequences correspond to Bangla nasal (vowel) - affricate sequences. It appears that post-nasal position is enough to protect the affricate from lenition, which accords with the cross-linguistic phenomenon of post-nasal fortition.

Based solely on the distribution of these two affricates in native Sylheti words, I would conclude that they behave, and should be treated, identically. However, native speakers produce loan items differently in the two cases: [d͡ʒ] is pronounced as [z], but [ʧ] is retained. It may be that Camden Sylheti is transitioning or has already transitioned to treating [ʧ] as a phoneme in its own right.

Other fricatives

The retroflex fricative [ʂ] is only found before the retroflex stop [ʈ]; it is an allophone of either /s/ or /ʃ/, both of which occur independently.

The glottal fricative [h] is not found in consonant sequences (except for the single item [bɾahmi] *type of plant*). It is found word-initially but not finally, and contrasts with the other fricative phonemes, e.g. [xasi] *knife*, [xaʃi] *cough*, and [xahi] *bowl*. [h] predominantly corresponds to Bangla [ʃ], with 61 [ʃ]-initial and 20 [h]-initial Bangla translations of Sylheti [h]-initial words. Unlike in Assamese, [h] is not an allophone of [x]: [hɔɾ] *to move* contrasts with [xɔɾ] *to do*.

Instead, [x] and [k] are allophones. [k] is found preceding or following a high vowel, as a geminate, and in a few loan items, with [x] found elsewhere. Given the existence of a number of loan items with [k] where [x] would usually be expected (e.g. [nɛklɛs], [kampuʈɔɾ]), it is possible that the allophony rule has become fossilised. For example, the borrowed word [ɾɪʃka], rickshaw, has had metathesis applied, but [k] is retained as though still in the environment of a high vowel.

We may see a split into two separate phonemes over the next few decades, particularly if there is an influx of English loanwords into Camden Sylheti.

Labials

The voiceless labial stop [p] is found only infrequently, and predominantly in two environments: following a labial nasal, and word-initially in the sequence [pɾ]. Items which are cognates with, or loans of, English items that contain [p] usually have [f] instead. Several items in the lexicon are recorded with both pronunciations (e.g. [ɪʂʈɛmp] / [ɪʂʈɛmf], [sappanno] / [saffanno]). I therefore conclude that [p] is an allophone of /f/. In terms of the development of this allophony, the fricative /f/ may be pronounced as [ɸ] or [f]; it may be that exposure to English labiodental [f] in Camden Sylheti is having an effect.

Phonemic consonant inventory

TABLE 4.3: Sylheti phonemic consonant inventory

b	t̪	d̪	ʈ	ʧ	k g	
f	s	z		ʃ		h
m		n			ŋ	
	l	ɾ	ɽ			

4.4.3 Syllable structure

Once a phonemic mapping has been established, Nidaba can be used to answer other segmental distribution questions.

The properties of items in the lexicon do not necessarily correspond to the properties of phonological words. For example, the Sylheti lexicon contains both stems and bound morphemes. Results can be restricted to free morphemes, using the custom filtering, since bound morphemes may contain final sequences that never surface. The filtering can be done directly, if bound morphemes are tagged as such, or using a combination of other tags such as part of speech data.

4.4.4 Word final consonants

In this section, I examine sequences found in word-final position in the lexicon. Since the lexicon contains both stems and bound morphemes, it contains final sequences such as [fn] belonging to bound morphemes [afn-] which do not appear as free morphemes, but only with a following vowel. The discussion below refers only to free morphemes, and hence consonants which surface in word-final position.

Nearly 45% of items in the Sylheti lexicon end in a consonant. The following consonants and clusters were found finally, in order of decreasing frequency: [ɾ], [l], [n], [ʃ], [t], [m], [ʈ], [x], [s], [ɽ], [k], [z], [d], [f] (>1% of items); [b], [g], [ŋ], [nd], [nd͡ʒ], [ɳʈ] (>0.1% of items). Voiced obstruents were not permitted in Sanskrit codas (Kessler 1994); this may account for the low frequency of [b] [d] and [g] relative to their voiceless counterparts.

Word-final consonant sequences

Setting aside sequences found in only one item – and those mostly loan items (e.g. [ɛbaɾɪsʈ], Everest) – we find the following multi-segment sequences: [nd], [nd͡ʒ], [ɳʈ], [ɾʈ] and [ɾd].

[nd] is found in verbal stems, and in nouns (see Table 4.4). These are mostly cognate with Bangla nouns which have a nasal vowel, instead of a stop-nasal cluster. NC (nasal-consonant) clusters were present in the protolanguage of Assamese-Bengali (see Table 4.5), though many were subsequently lost through a variety of processes (Pattanayak, 1966). Final clusters are not allowed in modern Bangla, but are in Assamese (Masica, 1991, p. 126). More investigation is needed to determine whether Sylheti retained the NC clusters like Oriya, or redeveloped them more recently from a nasalised vowel system like Bangla's.

Table 4.4: Examples of word-final [nd] in nouns

Sylheti	English	Bangla		Sanskrit	
[tɔbɔnd]	knot	বাঁধন	[bãdʰana]	बन्ध	⟨bandha⟩
[sand]	moon	চাঁদ	[ʧãd]	चन्द्र	⟨candra⟩
[xand]	shoulder	কাঁধ	[kãdʰa]	स्कन्ध	⟨skandha⟩
[damand]	son-in-law	জামাতা	[d͡ʒamata]	जमातृ	⟨jamātṛ⟩
[fand]	trap	ফাঁদ	[pʰãda]		
[ŋgland]	England	ইংল্যান্ড			

Table 4.5: Correspondences involving NC clusters

Sylheti	Bengali	Assamese	English	Reconstructed form
ɾɔŋ	ɾɔŋ	ɾɒŋ	colour	*ɾɔŋg
ɾaŋga	-	ɾɒŋa	red	*ɾɔŋg
aʃ	hãs	pati hãh	duck	-
sand	ʧʃãd	sɒndɾɒ	moon *ʧʃand	

[nd͡ʒ] is found in a single morphological item, [gɔmd͡ʒ] গঞ্জ *district*, and in place names derived from it: [hɔbigɔnd͡ʒ] *Habiganj*, [xɔɾimgɔnd͡ʒ] *Karimganj*, [sunamgɔnd͡ʒ] *Sunamganj*.

[ɾd] is found in the nouns [mɔɾd] *man* মরদ [mɔɾɔd], and [dɔɾd] *pain*. The status of these items is not clear; [bɛʈa] is the common term for man, and [biʃ]/[bɛdna] *pain* are listed in the lexicon both in isolation and, unlike [dɔɾd], in related compounds such as [bukut bɛdna], chest pain.

[ɳʈ] is found in four items which appear to be loan items from English: [kʊɾɛɳʈ] *electricity* (*current*), [ɾɛʂʈuɾɛɳʈ] *restaurant*, [fɛɳʈ] *trousers* (*pants*), and [happɛɳʈ] *shorts* (*half pants*). Likewise, [ɾʈ] is found only in [ɛɾfɔɾʈ] *airport* and [ʃaɾʈ] *shirt*.

Sylheti is more tolerant of syllable structure violations than segment quality violations. There are no cases where [p] is retained but a complex onset or coda is repaired. By contrast, in [happɛɳʈ], not only is [p] retained, but [f] is adapted to match it. We have seen that [ʧʃ] and [d͡ʒ] are protected from spirantization in geminates. Sylheti does not allow differing allophones within a sequence, and has a preference for stops over fricatives in geminates, resulting in these 'non-native' geminates in all three cases. Regarding the other segment quality adaptations, we have seen that English alveolar stops are borrowed as retroflexes. Nasals and fricatives are normally borrowed as dental / alveolar (e.g. [bɾɪʈan], [pɾofɛsaɾ]), but undergo place assimilation to retroflex as in native items. [ɛɾfɔɾʈ] and [ʃaɾʈ] have been borrowed from a rhotic variety of English (cf. Masica, 1991, pp. 75–76). In both onset and coda position, [ɾ] is borrowed as dental / alveolar, and does not undergo place assimilation. [ʈ] is an allophone of /ɖ/, and the sequence */ɖʈ/ would be ungrammatical; Sylheti does not have any homorganic stop sequences which differ in voicing. This results in the unusual sequence [ɾʈ], otherwise found only in the loan item [xaɾʈɔn] *curtain* and the pronouns [aɾʈa] *next* and [amaɾʈa] *mine*.

4.4.5 Word initial consonants

The initial consonants of Sylheti, in decreasing order of frequency, are the singletons [b], [f], [x], [s], [m], [ʃ], [h], [g], [d], [t], [z], [k], [n], [l], [ɾ], [ʈ], [ɖ] (found in >1% of items) and the sequences [bɾ], [pɾ], [fɾ], [kl], [st] and [gɾ]. There are other sequences, but each is found in only one lexical item, such as Hindi and Arabic greetings. The infrequent sequences appear to represent borrowings or re-borrowings from English and Sanskrit. Almost all are nouns, the most frequently borrowed class of lexical items (Campbell, 1993).

Table 4.6: A selection of low-frequency lexical items with otherwise ungrammatical initial clusters

Sylheti	English	Bangla		Sanskrit	
[bɾɪʈan]	Britain				
[bru]	brow	ভুরু	[bʰuɾu]	भ्रू	⟨bhrū⟩
[bɾiʂʈi]	rain	বৃষ্টি	[bɾiʃʈi]	वृष्टि	⟨vṛṣṭi⟩
[klas]	class				
[klantɔ]	tired	ক্লান্ত	[klantɔ]	क्लान्त	⟨klānta⟩
[gɾam]	village	গ্রাম	[gɾam]	ग्राम	⟨grāma⟩
[gɾiʃʃo]	'hot season'	গ্রীষ্ম	[gɾiʃmɔ]	ग्रीष्म	⟨grīṣma⟩
[pɾaʃnɔ]	question	প্রশ্ন	[pɾɔʃnɔ]	प्रश्न	⟨prazna⟩
[pɾotizʊgita]	competition	প্রতিদ্বন্দ্বিতা	[pɾɔtidbɔndbita]	प्रतियोगिता	⟨pratiyogitā⟩
[pɾofɛsaɾ]	professor				
[stiɾi]	wife	স্ত্রী	[stɾi]	स्त्री	⟨strī⟩
[stɔn]	breast	স্তন	[ʃtɔn]	स्तन	⟨stana⟩
[zɔlfɾɔfat]	waterfall	জলপ্রপাত	[jalaprapāta]	प्रपात	⟨prapāta⟩

Repair strategies

Metathesis A repair strategy which maximises retention of the original sounds is metathesis. Syllable structure requirements are met by transposing vowels and consonants, in this case to convert CCV.CV sequences to CVC.CV sequences. I have not located any examples of this strategy being applied to English borrowings; metathesis may no longer be an active repair strategy in modern Sylheti.

Table 4.7: Metathesis between Sanskrit (Old Indo-Aryan) and modern Sylheti

प्रति	⟨prati⟩	→	[fɔɾti]	every
प्रोष	⟨proṣa⟩	→	[fɔɾsa]	light

Anaptyxis Syllables with a pre-existing coda cannot have their onsets repaired by metathesis, given Sylheti's ban on complex codas, since this would simply replaced CCVC sequences with CVCC sequences. Instead, they are repaired with anaptyxis, the insertion of a vowel.

Table 4.8: Anaptyxsis

[bɛɾɛɖ]	'bread'
[fɛlɛɪt]	'plate'
[ɖɛɾɛm]	'drain'
[ʈɛɾɛm]	'train'
[gɔllas]	'glass'

Singha and Ahmed (2016) record three different vowels used in epenthesis: [i], [e] and [o]. Given limited examples in both corpora, there is not yet conclusive evidence for whether vowel choice is determined by vowel harmony (a feature of Bangla and Assamese, e.g. Mahanta, 2008) or by consonant quality. If the former, [i] requires [i], [e] requires [e], and [a] requires [o]; we have no examples with the other two vowels as triggers. If the latter, [i] is used with [k], preventing its adaption to [x]; [e] is used following labials and retroflexes (non-back consonants); and [o] is used following velars (back consonants).

Singha and Ahmed (2016) contains the example /silip^{h}/ *slip*, which supports their assertion that Sylheti has vowel harmony; however, this example contains both /p/ and contrastive aspiration, so I am reluctant to include it as reflective of Camden Sylheti.

Table 4.9: Prothesis

[ɪspid]	'speed'
[ɪʂʈɔf]	'stop'
[ɪʂʈɛmf]	'stamp'
[ɪʂʈiʃɔn]	'station'

Prothesis Loan words with an initial sT (s + stop) cluster are repaired through prothesis, the insertion of a vowel preceding the sequence. This holds for both sCVC(C) words, which cannot undergo metathesis, and for sCVCV(C) words, which could. This result is consistent with Goswami (2013)'s findings for North Tripura Sylheti. All examples of prothesis use [ɪ], regardless of the vowel quality of the following syllable, so again epenthetic vowel quality could be determined by the (here empty) onset. Alternatively, the intervention of a coda between the epenthetic

vowel and the following one might also play a role in blocking harmony, as in Assamese (Mahanta, 2008). The treatment of sC clusters as coda + onset, with repair being through prothesis instead of anaptyxis, is cross-linguistically common (Goad, 2012). The location of the boundary between the two strategies varies. For example, Hindi treats sT- and sm- clusters with prothesis, and sn-, sl-, s+r and s+glide sequences with anaptyxis. The single example of this in Sylheti is the repair via anaptyxis of [sɛlɛʈ] *slate*. It is not clear from this limited data if sn- sequences would be adapted with anaptyxis or with prothesis.

4.4.6 Word-internal consonant sequences

In this section, I examine evidence for word-internal codas and for complex onsets in Sylheti, derived from the application of Nidaba's subsequence and set comparison tools.

Using [ʃɔnda] as an example: it contains the word-internal consonant sequence [nd]. [d] appears word-initially in the lexicon, but [nd] does not. The longest possible internal 'onset' sequence in [ʃɔnda] is therefore [d], leaving [n] as the preceding coda. The set of word-final consonants can be compared to the set of internal codas calculated this way. Such a comparison shows that all word-final consonants that occur singly can also occur as word-internal codas.

Repeating the comparison for word-initial consonants, I find that all word-initial singletons also appear in word-internal onset position, as well as the previously mentioned retroflex allophones and geminates.

Word-internal complex codas

There are only a few items transcribed with CCC word-internal sequences.

Firstly, there are two bimorphemic items, [dɔkknɔɾ] southern and [ʊttɾɛ] northern. They appear to be formed by suffixation plus deletion from [dɔkkin] south and [ʊttɔɾ] north. In Bangla, there is a preference for disyllabic trochees, which Nagarajan (2014) proposes has been the case since at least the 17th century. This may account for the deletion. However, there is limited other evidence of this preference in the Sylheti lexicon, since the creation of disyllables through epenthesis (see Subsection 4.4.5) is more easily explained as a side-effect of syllable structure

repair. Furthermore, [kn] is not otherwise valid as either an onset or a coda sequence in Sylheti. More detailed studies are required into geminate behaviour under adjective and adverb formation.

Secondly, there is the bimorphemic item [zɔl-frɔfat] *water-cascade, waterfall.* As discussed in Subsection 4.4.4, there has been segment quality adaptation of the [p] of प्रपात ⟨prapāta⟩, but no apparent syllable structure repair. Being both bi-morphemic and potentially a re-borrowing, this is not a good candidate for a word-internal complex onset.

The remaining -CCC- items are of the form [ŋgC], and mostly [ŋgL]. [gl-] is not found as a word-initial cluster, and [gɾ-] only in a few loan items, as discussed in Subsection 4.4.5. Nor is [-ŋg] is found as a word-final sequence. There are no minimal pairs contrasting [ŋg] and [ŋ]. The loan item 'English' is pronounced variously with and without the [g], and the Bangla cognates also lack it. A more detailed phonetic study of these items and their variability is required to determine the phonological status of the [g], but the initial distributional data points towards it being excrescent, not phonemic.

Table 4.10: ŋ(g)C sequences

Sylheti	English	Bangla	
[hɪŋggi]	type of eel		
[ʈɛŋgɾa]	type of catfish	টেংরা	
[xaŋgla]	type of fish	ফলি	
[hamɔkbaŋgɾa]	'snail shell' stork	শামুকখোল	
[baŋgladɛʃ]	Bangladesh		
[baŋgla]	Bangla	বাংলা	[baŋla]
[ɪŋgland]	England		
[lɛŋgɽa]	lame	লেংড়া	[leŋɽa]
[baŋglagɔɾ]	room		
[ɪŋgɾɛzi]/[ɪŋgɾɛz]	English		
[ŋlɪʃ]	English		
[fiŋla]	pink		
[siŋla]	bamboo switch		

Syllable contact

Of the sequences of two word-medial consonants in the Sylheti lexicon, nearly 50% have falling sonority; 20% are identical consonants; 5% are non-geminates with level sonority; and 25% have rising sonority. Some of the rising sequences are loan items from languages with complex onsets

(e.g. [madrasa]), whereas others have been retained from Sanskrit. Whilst the Syllable Contact principle holds that sonority should drop across syllable boundaries, it is "often [overridden by] the prohibition of complex syllable onsets" (Clements, 2009). The incidence of word-medial rising sonority sequences in Sylheti therefore does not rule out a prohibition on complex onsets both initially and medially.

4.4.7 Vowels

The vowels of Sylheti, in descending frequency of occurrence as single vowels in the lexicon, are [a, ɔ/o, i/ɪ, ɛ/e, u/ʊ].

[o], [e] and [ɪ] are almost certainly allophonic variants of /ɔ/, /ɛ/ and /i/ respectively, if not transcription variants. There are only a few items transcribed with these segments. There are no minimal pairs which distinguish between [e] and any other segment; no minimal pairs distinguishing between [i] and [ɪ]; and no minimal pairs distinguishing between [o] and [ɔ]. [u] and [ʊ] are fairly evenly distributed in initial, medial and final position. However, the only minimal pair for these items is [-u] (emphatic morpheme) vs [-ʊ] (first person morpheme). Sylheti has multiple homophonous single vowel morphemes, such as [-ɔ]: locative / second person for type I verbs / third person for type II verbs. There is therefore no compelling evidence for a contrast in the absence of native speaker clarification. [u] and [ʊ] are almost entirely predictably distributed when in sequence with another vowel, so in the discussion that follows, I treat Sylheti as a five vowel system.

The distribution of vowel combinations in Sylheti implies that the majority of VV sequences are diphthongs. In descending order of frequency, the observed sequences are: [ai, ɔi, ia, ʊa, ɔʊ, ʊi, aɔ, aʊ, ɛi, aɛ, ɔa]; [eɔ, iɔ] at around 1% of VV occurences; and least frequently [ɛu, iʊ, ɛa, ʊo, ɔɛ]. [iɛ] and [ʊɛ] are missing altogether. Given the relative frequency of the vowels in isolation, VV sequences with i or ʊ as the second member are overrepresented (with the exceptions of [ʊi] and [iʊ]), as are the sequences ia and ʊa.

As a first approximation, Sylheti allows diphthongs and short open syllables both word-internally and finally.

Like all contemporary Indo-Aryan languages (Masica, 1991, p. 128), Sylheti has syllable initial vowels (e.g. [afne] *you*), and allows morpheme-internal vowel hiatus (e.g. [gaɪɔx] *male singer*).

The maximum number of morpheme-internal vowel qualities in a sequence is three. These triple vocoid sequences likewise do not show free combination of vowel qualities: the majority of them can be sequenced as Vi and Vʊ diphthongs with following vowels. However, there are only 54 morphemes containing such sequences in the lexicon, so more detailed conclusions cannot be drawn from the limited data available.

By combining vowel-final verb stems with vowel-initial suffixes, Sylheti can have sequences of up to five vowels (three syllables), like Assamese. For example, [xaʊa] *to cough* may be inflected [amɪ xaʊaɪaɾ] *I'm coughing*.

A fuller analysis of vowel phonotactics in Sylheti will require a detailed examination of the status of diphthongs and their potential interactions with tone.

4.4.8 Conclusion

I have illustrated the use of Nidaba in examining the inventory and syllable structure of a language, and its relationship to neighbouring languages. More information can be found at nidaba.co.uk.

4.5 Similar databases and tools

In this section, I discuss eight existing databases and computational tools which are similar in function to Nidaba, and what makes Nidaba unique.

4.5.1 AusPhon-Lexicon

The AusPhon-lexicon project (Round, 2017) is a 'data warehouse' currently containing normalised lexicons for 166 Australian language varieties, with data querying tools including an extended regular expression language.

Nidaba is effectively an application of this idea, trading depth of analysis for universality: Nidaba users are required to scrub their own data and produce their own normalisations, but are not restricted to a given language family.

4.5.2 World Phonotactics Database

The World Phonotactics Database has broadly similar aims of providing a typology of parameters (termed 'features') which describe syllable structure. However, it does not have any parameters dealing with sonority, which forms the basis for many phonotactic formulations (e.g. Blevins, 1995). The raw data is not available to verify how parameter value choices were made, which also limits flexibility in adding extra parameters, or making alternative choices using different cues.

Nidaba, by contrast, is primarily concerned with distributional data, including place and manner information. It aims to provide the tools necessary for users to replicate my results. It is also intended to be sufficiently flexible that users can make different assumptions about valid input data, phonemic representation, sonority, or syllable structure, or add new parameters.

4.5.3 P-base

P-base (Mielke, 2008) "is a database of several thousand sound patterns in 500+ languages". However, these are not distributional patterns but processes such as nasalisation or devoicing. Again, the data on which these patterns are based is not available to the user.

Nidaba can be used to duplicate some of the functionality of P-base, by inputting a narrowly transcribed wordlist, and searching for particular combinations of properties. In this way, the results of P-base can be verified, and specific examples of its sound patterns found in a lexicon. However, the primary purpose of Nidaba is to look at more static distributional patterns.

4.5.4 TalkBank

The TalkBank project (MacWhinney, 2000) comprises CHILDES (Child Language Data Exchange System) and other corpora. Each corpus contains audio and/or video recordings and a transcription of the data in CHAT format. This is the input format for the accompanying analysis program CLAN, which performs various kinds of discourse analysis. Among the analyses is token frequency, which is a useful input to Nidaba. You can also get PHONFREQ, which performs similar functions to Nidaba's segment search, but with much less powerful search tools.

Another accompanying analysis program is Phon (Rose et al., 2006). Phon contains tools for searching by features, like Nidaba; but its use case is analysing a spoken corpus, not a lexicon, and it does not contain tools for comparison between different phonemic analyses or languages.

4.5.5 Phonology Assistant

Phonology Assistant (SIL, 2008) provides tools for inventory analysis, given a corpus of transcription data. Whilst Nidaba provides a basic inventory tool, its main focus is instead on distributional data.

4.5.6 Phoible

PHOIBLE (Moran, McCloy and Wright, 2014) "is a repository of cross-linguistic phonological inventory data". Its two guiding principles have also been applied to Nidaba, namely that all data should be encoded in Unicode IPA, and that data from multiple doculects should be faithfully included. Nidaba also includes much information beyond inventory data, e.g. it cross-references all phonemes with lexical items, to aid in the treatment of marginal items.

4.5.7 CLTS

CLTS (List, 2017) is "a cross-linguistic database of phonetic notation systems". When complete, this will be a useful source for generating or verifying transcription conversions for Nidaba, which is currently a manual process for individual researchers.

4.5.8 ILSP PsychoLinguistic Resource

ILSP PsychoLinguistic Resource (Protopapas et al., 2012, located at speech.ilsp.gr/iplr/) provides computational tools for in depth search and analysis of Greek, based on two printed text corpora. Many of tools are similar in function to Nidaba tools: returning subsets of a corpus based on length, frequency, and syllable structure. The available data for Greek is more extensive than that in Nidaba, including orthographic / phonological 'neighbours' of lexical items, as measured by Levenshtein distance; stress; and 'orthographic transparency' (predictability of grapheme/-phoneme correspondence); but it is limited to Greek only.

4.5.9 SYLLABARIUM

SYLLABARIUM (Duñabeitia et al., 2010) is a web tool for examining syllables in Spanish and Basque. It provides similar functions to Nidaba in locating type and token frequency of different syllables, but is limited to orthographic data, and only in those two languages.

4.6 Languages

Nidaba contains phonemically transcribed[7] word lists for the following languages:

- Ambel, an Austronesian language (fieldwork of Laura Arnold)
- Cheke Holo, an Oceanic language (White, Kokhonigita and Pulomana, 1988)
- Dutch (CELEX: Baayen, Piepenbrock and Rijn, 1993)
- English (CELEX: Baayen, Piepenbrock and Rijn, 1993)
- French (Lexique3: New, Pallier et al., 2001)
- German (CELEX: Baayen, Piepenbrock and Rijn, 1993)
- Greek (GreekLex: Ktori, Heuven and Pitchford, 2008)
- Hrusso Aka, a Tibeto-Burman language (fieldwork of Vijay D'Souza: D'Souza, 2015)
- Lithuanian (Tang and Harris, In prep(a))
- Matbat, an Austronesian language (Remijsen, 2015)
- Portuguese (PorLex: Gomes and Castro, 2003)
- Polish (Tang and Harris, In prep(b), Howell et al., 2017)
- Romanian (Tang and Harris, In prep(c), Howell et al., 2017)
- Spanish (EsPal: Duchon et al., 2013)
- Sylheti, an Indo-Aryan language (SOAS Sylheti Project, 2015)
- Welsh (Ellis et al., 2001)

[7] The exact type of transcription varies widely between projects. Many have been derived by applying pronunciation rules to orthography, with resulting oddities, including Greek, Lithuanian, Polish, Romanian and Spanish.

4.6.1 Phonemic inventories

For the following languages, the source (or at least reference) of the phonemic transcription is separate from the source of the lexicon: Cheke Holo (Corretta, pc.); Dutch (CELEX: Burnage, 1990); English (CELEX: Burnage, 1990); German (CELEX: Burnage, 1990); Sylheti (Eden, in press); and Welsh (Pronunciation data from Williams, Jones and Uemlianin, 2006, converted into transcription by Florian Breit). In the case of English, I adapted the DISC transcription system to remove nasal vowels: æ̃ː → ɒ; ɑ̃ː → ɒ; æ̃ → ɑː; ɒ̃ː → æ.

4.7 Frequency data

For every parameter and diagnostic discussed in Section 4.8, it is necessary to consider how to treat loan words and other marginal examples. Neither dismissing them completely nor treating them as contributors to phonotactics equivalent to the core vocabulary of the language adequately captures the facts.

For this reason, Nidaba contains corpus frequency information on the lexicons of the languages, where it exists. This allows parameter values to be set for a minimum frequency threshold, or number of distinct lexical items in the input. For the parameters below, this threshold has been set at one hundred occurrences per million tokens, or five Zipf[8] (Heuven et al., 2014).

Since film subtitle corpora have been shown to be superior sources of frequency norms than traditional written corpora (New, Brysbaert et al., 2007, Brysbaert and New, 2009), I have where possible combined phonemic word lists with frequencies in subtitles via the written forms common to both sources. Lexique3 (French), EsPal (Spanish) and the Lithuanian, Polish and Romanian corpora contain frequency counts drawn from subtitle data (New, Pallier et al., 2001, Duchon et al., 2013, Mandera et al., 2014). Dutch frequency data was drawn from Keuleers, Brysbaert and New, 2010, British English from Heuven et al., 2014, and German from Brysbaert, Buchmeier et al., 2011. European Portuguese frequency data was approximated using Brazilian Portuguese subtitle data (Tang, 2012).

For Greek and Welsh, such subtitle corpora have not yet been compiled. The GreekLex database contains frequency counts drawn from the Hellenic National Corpus, a collection of written Modern Greek texts (Ktori, Heuven and Pitchford, 2008). Welsh frequency data was taken from

[8] The Zipf scale is a logarithmic scale, related to frequency per million words by the formula: $\text{fpmw} = 10^{\text{Zipf}-3}$

the Cronfa Electroneg o Gymraeg, based on a million words of written Welsh prose (Ellis et al., 2001).

For languages such as Sylheti, which lack a written form distinct from the majority language, no token frequency data has as yet been provided; only the number of distinct lexical items can be derived. This is also the case for Ambel, Cheke Holo, Hrusso Aka, and Matbat.

4.8 Parameters

Having developed a program to aid in establishing parameter values, in this section, I describe an example set of parameters, and their values for 16 languages. These will be used to calculate Hamming Distance in Section 4.9.

The contents of this section are as follows: Choosing parameters; Diagnostics; Syllable structure parameters (*CV syllable, Consonant cluster analyses, Syllabic consonant parameters, Sonority reversal parameters, Sonority distance parameters*); Vowel inventory parameters; Consonant inventory parameters (*Laryngeal parameters, Obstruent place parameters, Nasal place parameters, Fricative place parameters, Manner parameters*).

Subsection 4.8.3 describes syllable structure parameters, Subsection 4.8.4 on page 83 describes vowel inventory parameters, and Subsection 4.8.5 on page 87 describes consonant inventory parameters. Tables summarising the values are found at page 97.

4.8.1 Choosing parameters

We saw in Subsection 3.2.3 that historical relationships could be modelled using parameters chosen to reflect known innovations in Indo-European. To instead model phonological similarities between languages, the parameters must reflect typological observations. There are of course many different strategies which could be employed to do so. My intention is that the functionality provided by Nidaba will allow other researchers to adopt different strategies for different purposes.

For the set of parameters below, I am following the principle that they are to be chosen independently of the expected result. That is, I am attempting to include or exclude no parameters on the basis of existing knowledge of a language relationship, or of ease of acquisition, or any other similarities between languages. For this reason, I have limited the parameters to two

particular areas, and attempted to exhaustively cover those areas. This should prevent cherry picking of 'relevant' values.

I have 27 syllable structure parameters, and 29 inventory parameters. The syllable structure parameters have been chosen to provide, as far possible, a typology of syllable and sonority types, as explained below.

The vowel and consonant parameters have been chosen to reflect those choices which characterise the greatest number of languages. The ideal parameter for this purpose would be one which equally partitions known languages, and so is true for 50% of languages and false for the other 50%, though the majority of the parameters have more unequal distributions. In the ideal case, two languages sharing a true value for a given parameter and two languages sharing a false value are equally likely scenarios, both of which would count equally towards the metric. For further discussion, see Subsection 4.9.3. The inventory parameters also generally reflect the options described by most systems of distinctive features.

4.8.2 Diagnostics

Due to the nature of the source data (i.e. lexical databases containing phonemic representations), I will be limiting my diagnostics for syllable structure to distributional information. I will not be using diagnostics which are based upon acoustic data, or experimental results such as the propensity of speakers to insert additional vowels when prompted, or of listeners to misperceive clusters found only in loanwords. Similarly, the inventory parameters are mostly focussed on contrasts, or on very broad place and manner categories which do not require detailed acoustic experiments.

Where possible, I have cited additional sources beyond Nidaba to verify its accuracy.

4.8.3 Syllable structure parameters

The syllable structure parameters which I am examining fall into four sets: those relating to deviations from a CV syllable; those relating to syllabic consonants; those relating to sonority profiles; and those relating to sonority distance.

Those parameters which might be expected for reasons of symmetry, but which are missing, are those which have been found to be uniformly valued in all languages. Any parameter

for which one of its values is the empty set can be restated as an unconditional universal (Greenberg, 1966). Universals are by definition irrelevant to a measurement of difference, and so will not contribute to the metric. However, it is perfectly possible to verify these universals using the data present in Nidaba.

It is not my intention to take a position on the mental representation of the syllable, or even whether the syllable is more than a convenient fiction. Nonetheless, I hope that the typological observations below may be relevant to a broad set of theoretical positions, and that the data in Nidaba will allow readers who disagree to create additional or replacement parameters of their own. The use of terms such as 'coda' is therefore purely conventional, and the following parameters and resultant distance measurements are a proof of concept, not a finished product. One of the major applications of this study is in comparing the consequences of different theoretical positions for language distance, as is explored in Chapter 5.

CV syllable

I begin with parameters relating to the segmental positions in the syllable. The most common syllable structure cross-linguistically is consonant-vowel, or CV: the words of many languages can be divided into alternating CV sequences, whilst there are no, or very few, languages which contain only VC alternating patterns (Hyman, 2008); that is, which forbid word-initial consonants, or which require word-final consonants (Dam, 2004).

Marked syllable structures consist, firstly, of syllables with one segmental change: a missing onset; an additional initial consonant (branching onset); an additional nuclear position (complex nucleus) or a final consonant (coda). I shall assume that a syllable minimally consists of a nucleus.

In some cases, the sequence is doubly distinct from a CV sequence, and we observe three or more initial consonants; two extra nuclear positions[9], as Remijsen and Gilley (2008) argue for in Dinka, a Nilo-Saharan language; or two or more final consonants.

I will not be including parameters examining whether a language has the unmarked structure, since in almost every case such a structure is arguably universal, and hence not useful for measuring similarity.

[9]Since my sample of languages does not however contain any which contrast three nuclei lengths, I will not be including a parameter examining this

Finally, the marked structures can be combined. The combination of the nuclear and coda structures (the rime) may be restricted, where the interaction between onset and nucleus or onset and coda is not (J. Harris, 1994, p. 47).

The traditional domain to examine for syllable structure is the word, and I have included parameters for the presence of these marked structures at word edges. However, codas may appear word-internally but not word-finally, or vice versa (Kaye, 1990), so these parameter values cannot be straightforwardly generalised to statements about syllable structure. My parameters referring to word-edge phenomena have been named with 'onset' or 'coda' for brevity and memorability, rather than as theoretical statements.

In any given language, the phonotactics of morphemes may pattern with word edges, or with the internal structure of morphemes, or be divided between the two types (J. Harris, 1994). Nidaba lacks morphological marking in its requirements for lexicons, and so parameters addressing phonotactic behaviour at morpheme boundaries are absent from the current parameter set. However, the code has been designed to be easily extensible to cover this data in future, as it has with other non-segmental properties such as tone.

For the parameters which follow, I summarise the question to be answered, discuss the diagnostics required to answer it, and note where the cross-linguistic pattern differs from the generally unmarked syllable structure.

Consonant cluster analyses

There are multiple parameters for which distinguishing between affricates and consonant clusters or between diphthongs or vowel hiatus is required. The relevant diagnostics are set out below.

Affricates To set the value of parameters 4.8.3.2, 4.8.3.5 and 4.8.3.6 below, it is often necessary to decide whether a sequence of two consonants forms an affricate or a cluster.

Using only distributional information, that means deciding if the sequence is distributed like singleton consonants or like clusters. It will be labelled an affricate if it can occur in final position where only otherwise only single segments occur; if it can occur in initial position where otherwise only single segments occur; and if it can occur in initial position in combination with another consonant, where three-segment sequences do not otherwise occur. Alternatively, the

potential affricate may contrast with same quality consonant-consonant sequences, as in Polish (Rubach, 1994).

Glides and diphthongs A sequence of two vocoids may constitute either two vowels in separate syllables, with hiatus; a diphthong (with an on-glide or off-glide) constituting a single nucleus; or a VC or CV sequence. The diagnostics must therefore distinguish between these three categories, for parameters 4.8.3.2, 4.8.3.3 and 4.8.3.4, below.

Diphthongs should pattern with (long) monophthongal nuclei. They should be found preceded by all possible onsets, and followed by all possible codas, with exceptions conforming to monophthongal phonotactics. Nidaba's corpus frequency and word count tools permit 'accidental' gaps to be spotted, i.e. where the expected frequency of certain phonotactic patterns, given the observed frequencies of their constituent segments or parallel patterns, is so low that they have failed to appear in a non-exhaustive lexicon, and no conclusions can be drawn from their absence. If a potential diphthong is never found following a branching onset in a language which has them, then it should be analysed as a CV sequence. Likewise, if it is never found preceding a coda, then it should be a VC. If a glide is found preceding (or following) a long vowel or diphthong, then it is consonantal, and forms part of the onset (or coda). If the only observed diphthongs are word-initial, with on-glides, then these are better analysed as CV sequences than VV, particularly if there are consonantal glides observed elsewhere, or these are the only vowel-initial sequences in the language. The same applies to word-final sequences with off-glides.

If there are no restrictions on which vowels can occur together, then the vocoid sequences are not diphthongs (J. Harris, 1994).

(1) **Obligatory onset parameter**

Does the language have vowel-initial words?

Since CV is the unmarked syllable, vowel-initial syllables are marked (Itô, 1989). Whilst most languages do have vowel-initial words, these will all have consonant-initial words, whereas the presence of consonant-initial words does not imply vowel-initial words.

The diagnostics used are: Are there words which always begin with a vowel? An example would be Spanish, which contains words which are always pronounced without initial constriction (Rakow and Lleó, 2011, p. 215). Are there words which sometimes begin with a vowel? An example would be English, which usually inserts glottal stops post-pausally with otherwise vowel-initial words (Cruttenden, 2014). Are there words which, whilst phonetically not beginning with a vowel, are pronounced with a default consonant which plays no other role in the phonology of the language? An example of such a language is German, which uses a glottal stop only word-initially in otherwise vowel-initial words (Benware, 1986, p. 28). In both of the latter cases, the glottal stop does not have a phonemic role. For all three cases, the language is categorised as having phonemically vowel initial words.

All the languages in my sample have vowel-initial words.

(2) **Double onset parameter**

Does the language have two consonants word-initially?

This is so if there are any words with two consonants word-initially, and these are true clusters and not affricates (see Subsection 4.8.3). If the second consonant in all such examples is a glide, it should belong to the onset and not the nucleus (see Subsection 4.8.3).

Ambel, Cheke Holo, Dutch, English, French, German, Greek, Hrusso Aka, Lithuanian, Polish, Portuguese, Romanian, Spanish and Welsh have two initial consonants. Matbat has two consonants sequences word-initially; whilst these sequences do occur internally, there is only one example in the lexicon which occurs following a word-internal coda. Sylheti does not have branching onsets, except morpheme-initially in loan items from Sanskrit, English or other branching languages, many of which are nativised with vowel epenthesis (Eden, in press).

(3) **Complex nucleus parameter**

Does the language have syllables with complex nuclei?

Firstly, does the language contrast long and short vowels of the same quality?

This diagnostic is thus phrased to simplify the classification of systems such as German, where the two classes of sounds are alternatively analysed as short vs long (with vowel quality a phonetic effect) or tense vs lax (with vowel length a phonetic effect)(Benware, 1986, p. 50).

Secondly, does the language contain diphthongs?

See Subsection 4.8.3 for diagnostics.

Of the languages in my sample, only the Dutch, German and Welsh lexicons contain long vowels which contrast with short vowels of the same quality. Whilst most German vowels differ in quality as well as length, [ɛ] and [a] have both been transcribed in this lexicon with length contrasts (Burnage, 1990). Likewise for the Dutch lexicon, in which words of French origin give rise to a length contrast in [ɛ]. Welsh has a full set of vowel contrasts, with every vowel quality having long and short counterparts (Jones, 1984).

All three of the languages above also contain diphthongs, as do English, French, Lithuanian, Portuguese (Mateus and d'Andrade, 2000) , Romanian (Chitoran, 2002), Sylheti and Spanish (Harris and Kaisse, 1999).

Ambel, Cheke Holo, Greek, Hrusso Aka, Matbat, and Polish do not have diphthongs. There are no restrictions on Cheke Holo vocoid sequences - all combinations of the five vowels are found word-medially - so I take these to be V.V sequences, not diphthongs. Hrusso Aka has consonantal glides, but no diphthongs, following the criteria above (D'Souza, 2015).

(4) **Coda parameter**

Does the language have word-final consonants?

See Subsection 4.8.3 for determining if final glides are consonantal or not. See 4.8.3.17 for languages with a limited set of word-final consonants.

Cheke Holo does not have word final consonants. The remaining languages in my sample do.

(5) **Triple onset parameter**

Does the language have three consonants word-initially?

This is so if there are any words with three or more consonants word-initially, which are true clusters (rather than an affricate combined with another consonant, see Subsection 4.8.3). None of the consonants must form a syllable peak (see 4.8.3.7).

Dutch, German, Greek, English, French, Lithuanian, Romanian and Welsh have word-initial sequences of three segments where the first is a sibilant[10]. Polish and Portuguese have other three-segment initial sequences (see 4.8.3.14). Ambel, Cheke Holo, Hrusso Aka, Matbat, Sylheti and Spanish do not.

(6) **Double coda parameter**

Does the language have multiple consonant segments word-finally?

This is so if there are any words with two or more consonants word-finally, which are true clusters (rather than affricates, see Subsection 4.8.3). None of the consonants must form a syllable peak (see 4.8.3.7).

Dutch, English, French, German, Lithuanian, Polish, Romanian and Welsh have multiple consonants word-finally.

Greek, Hrusso Aka, Matbat, Portuguese, Spanish and Sylheti do not, bar the exceptions listed in 4.8.3.17.

This parameter is not applicable to Cheke Holo.

(7) **Superheavy rime parameter**

Does the language have word-final superheavy rimes?

Are there any words which end in a complex nucleus followed by a final consonant? (As diagnosed in the Complex Nucleus and Word-Final Consonant parameters.)

Welsh has word-final consonants following diphthongs and long vowels[11]. Dutch, English, French, German, Lithuanian, Portuguese, Romanian, Sylheti and Spanish have word-final consonants following diphthongs.

[10] I am discounting the French word 'croissant', found in both English and Dutch lexicons, and the prefix 'pseudo', because the pronunciations listed are inaccurate (see e.g. Cambridge Dictionary, 2015); native speakers do not use [krw-] or [ps-]. English and Dutch are therefore listed only with sibilant-initial triples.

[11] Although only in monosyllables.

Syllabic consonant parameters

Segments in a syllable tend to be organised according to the sonority scale. From least to most sonorous, the scale is usually (e.g. Clements, 1990) given as:

obstruents - nasals - liquids - glides - vowels

Whilst the existence of such an organising principle is widely recognised, the exact phonetic basis of the scale, if any, is disputed (J. Harris, 2006). Options include intensity (Parker, 2002) resonance (Clements, 2009), or "universal markedness restrictions" (Berent, Harder and Lennertz, 2011). Since there is disagreement in the motivation of the scale, there is also disagreement about the details. Some versions of the scale are more fine-grained (e.g. Blevins, 1995, Baertsch, 2002), dividing obstruents into stops and fricatives, dividing liquids into laterals and rhotics, or dividing categories by voicing or place of articulation. Since Nidaba is configurable, it is possible to define an alternative sonority ranking to be applied to the lexicons, and thereby produce an alternative version of the parameter values below. For more radical departures from the sonority scale, the detailed information and tools provided by Nidaba will aid in the exploration of other principles of syllabic organisation. Nidaba is also designed to be extensible, so such an alternative based on e.g. perceptual distance between adjacent segments (J. Harris, 2006) could be implemented without requiring alteration of the existing codebase.

The sonority sequencing principle (SSP) states that syllables are organised with sonority minima at syllable edges, and a monotonic increase in sonority towards the centre (e.g. Kiparsky, 1979, Clements, 1990, Zec, 1995). All languages have syllable peaks which are vowels, but some languages also permit other segments.

In order to determine whether a consonant is the highest sonority segment of a syllable, it is necessary to decide what the syllables of the word actually are. All four sonority types discussed below rely on the same distributional diagnostics: Can the consonant occur as the highest sonority segment in a prosodic word? If so, then it must constitute a syllable peak; there is at least one syllable in that word which does not have a vocalic nucleus. Do syllabic consonants pattern with vowel nuclei? If they are true syllable peaks, they should occur in a position which is preceded by an onset and/or followed by a coda. Finally, does a syllabic C contrast with CV or VC? This last diagnostic distinguishes surface and underlying syllabic consonants, consistent with the methodology used for the World Phonotactics Database (Dawson and Donohue, pc.),

but contrary to Bell (1978). According to this diagnostic, Cantonese does allow syllabic nasals, where English does not, since English syllabic nasals can always alternate with [əN]. This should be represented in the phonemic transcription of lexical items in the database.

I outline four syllabic consonant parameters below. All four are false for all sixteen languages in my sample, so will not contribute to relative language distance.

(8) **Syllabic liquid parameter**

Can a liquid be a syllable peak?

For example, Sanskrit had syllabic liquids as its only syllabic consonants (Donohue et al., 2013).

(9) **Syllabic nasal parameter**

Can a nasal be a syllable peak?

Despite their lower position on the sonority scale, nasals are more common as syllable peaks than liquids (Bell, 1978). That a language has syllablic nasals does not imply that it has syllabic liquids - for example, Swahili (Donohue et al., 2013).

(10) **Syllabic fricative parameter**

Can a fricative be a syllable peak?

Syllabic fricatives are claimed to exist in Liangshang Yi, which does not have syllabic liquids or nasals (Ladefoged and Maddieson, 1990). Whilst other Chinese languages debatably also have syllabic fricatives, under some analyses these are allophones of vowels. Ultimately, the output of Nidaba is dependent on the phonemic transcriptions (or retranscriptions) of the input data. By allowing a variety of analyses for the same narrowly transcribed or orthographic input data, Nidaba allows users to choose the analysis they feel is most appropriate, and in doing so compare the results of using different analyses.

(11) **Syllabic stop parameter**

Can a stop be a syllable peak?

It is claimed that any segment may be a syllable peak in Tashlhiyt Berber (Ridouane, 2008), among other languages. In each of these languages, fricatives may also be syllable peaks, but the sample is too small to conclude that there is a implicational universal, particularly when no other type of syllabic peak implies the presence of another type.

Sonority reversal parameters

The Sonority Sequencing Principle is not obeyed in the clusters of certain languages. The violations are frequently initial fricative + stop (usually [s] + stop, hereafter sC) clusters, giving a dip in sonority. Explanations for the behaviour of sC clusters include describing [s] as extrasyllabic (Green, 2003), and describing it as a rime with an empty nucleus (Kaye, 1992). One of the additional types of evidence on which these hypotheses are based are apparent syllable structure violations, such as s-initial three-segment 'onsets' in English, which otherwise only permits two-segment onsets. I have referred to these three-segment sequences as 'triple onsets' below, but this is purely conventional.

My parameters therefore cover whether SSP and other syllable structure violations are permissible generally (as in Russian, e.g. Davidson and Roon, 2008) or are limited to a small set of segments.

These parameters can be derived automatically from the phonemic representations in Nidaba (see Subsection 4.3.2).

(12) **Word-initial sonority sequencing principle violation parameter**

Does the language contain word-initial sequences which are not monotonically increasing?

Dutch, English, French, German, Greek, Lithuanian, Polish, Romanian and Welsh contain word-initial sequences which are not monotonically increasing.

Ambel, Cheke Holo, Hrusso Aka, Matbat, Portuguese, Spanish, and Sylheti do not.

(13) **Initial sonority violations set parameter**

Are initial violations of the Sonority Sequencing Principle limited to a fixed subset of permissible onset segments?

The set members are identified by working from the sonority peak outwards. In a fricative-stop sequence, the fricative would be part of the set, and the stop would not.

The following (Indo-European) languages only permit s-initial onsets to violate the SSP: Dutch, English, French and Welsh. German and Lithuanian permit [ʃ] as well as [s]; this is limited to only a few lexical items in Lithuanian, only one of which – ⟨štai⟩ *here* – is high frequency. Romanian permits [z] as well as [s] and [ʃ] in voiced contexts. Greek permits [s], [f] and [x] in low frequency items, but in high-frequency items, only [s].

Whilst not without some combinatorial restrictions, word-initial violations of the sonority sequencing principle in Polish are not limited to a fixed subset of onset segments (Gussmann, 2007).

This parameter is not applicable to Ambel, Cheke Holo, Hrusso Aka, Matbat, Portuguese, Spanish, or Sylheti.

(14) **Onset structure violations set parameter**

Is there a set of segments which participate in violations of the normal onset structure of the language? (Hereafter the 'Onset structure violations set'.)

Examples of segmental exceptions are two consonants word-initially in a language which otherwise only permits one, or three consonants word-initially in a language which otherwise only permits two. The members of the Onset structure violations set may or may not also be participants in normal onset structure. For example, in English, the Onset structure violations set is {[s]}. This is the only segment which can begin a sequence of three consonants word-initially.

German has triple onsets with [s] or [ʃ] including all frequency and all incidence sequences; in high frequency or high incidence sequences, just [ʃ].

Dutch, English, French, Greek, Lithuanian and Welsh have triple onsets beginning with [s]. Romanian has triple onsets beginning with [s], and also, in a single low frequency sequence [zdr-], [z].

Polish allows multiple consecutive branching onsets, but this is not restricted to a particular set of segments (Gussmann, 2007). European Portuguese has vowel elision which results in quite permissive sequences of three or more consonant segments initially (Mateus and D'Andrade, 1998), but this is not reflected in the Porlex lexicon (Gomes and Castro, 2003) in Nidaba. However, the lexicon does contain examples of obstruent-liquid-glide sequences preceding diphthongs (e.g. ⟨frieira⟩ *chilblain*), which are not limited to specific obstruents, liquids or glides[12].

Ambel, Cheke Holo, Hrusso Aka, Matbat, Spanish and Sylheti have no triple onsets.

In Lithuanian, [ʃ] is found in the initial SSP-violating set, but not triple onsets.

(15) **Word-final sonority sequencing principle violation parameter**

Does the language contain word-final sequences which are not monotonically decreasing?

Dutch, English, French, German, Greek, Polish, Romanian and Welsh contain such sequences. Spanish has final ⟨-ts⟩ as a plural of items originating in English and French (e.g. robot, complot); all other sequences are below the frequency threshold.

Hrusso Aka does not contain word-final sequences; such sequences in the lexicon are only found in a few particular items, such as English words (e.g. ⟨Oxford⟩, ⟨dialect⟩). Lithuanian does not have any such sequences with a frequency above 100 per million items. Portuguese has final [ks] (⟨-x⟩) in eight loan words, and a few other items with stop-[s] sequences, all with very low token frequency. There are no such sequences in Ambel, Matbat or Sylheti.

This parameter is not applicable to Cheke Holo.

(16) **Final sonority violations set parameter**

Are violations of the Sonority Sequencing Principle limited to a fixed subset of permissible coda segments?

As for the initial SSP violations set, the members of the final SSP violations set are identified by working from the syllable peak outwards.

[12] Mateus and d'Andrade (2000) describe the third segment as behaving phonetically as a glide, and not patterning as part of the following rime, but they conclude that it should be treated as the nucleus of its own syllable, not as part of the onset. However, I am not changing the parameter value for Portuguese on this basis. Instead, I am adhering to the distributional information in Nidaba for consistency between languages.

Dutch, German, Greek and Welsh permit word-final sequences which increase in sonority to [s]. English also permits [z], in ⟨*wards⟩ e.g. towards, backwards.

In Romanian, the set of final sonority-violating segments is {[s], [m]}, found in the sequences [ks], [sm] and [tm].

In French, final sonority-violating sequences can be divided into three groups: sequences ending in liquids {[ʁ], [l]} which are found word-initially; sequences starting with [ʁ][13]; and [ks]. There also exist low frequency sequences ending in [m]. These groups do not form a single fixed subset of permissible coda segments.

In Polish, all segments that appear word-finally also appear as the endpoint of rising sonority sequences, except [ʃ] and [ʒ], and minimally sonorous stops and affricates. This parameter is therefore false for Polish.

This parameter is not applicable to Ambel, Cheke Holo, Hrusso Aka, Lithuanian, Matbat, Portuguese, Spanish or Sylheti.

(17) **Coda structure violations set parameter**

Is there a set of segments which participate in violations of the normal coda structure of the language? (Hereafter the 'Coda structure violations set'.)

For example, a language may have only single consonant codas, except in the case of [s], which can attach to the end of any syllable, creating final sequences. In this instance, the Coda structure violations set is {[s]}.

The Coda structure violations set should be determined on the basis of monomorphemes where possible; it should be an observation of phonological behaviour, not simply a list of possible affixes.

Dutch permits two consonants word finally, except for the set {[s], [t]}, which can create three segment sequences.

English permits two consonants word finally in monomorphemes, with the majority of three-segment sequences containing the past tense or plural affixes. However, the coda structure is

[13]The sonority of rhotics is a contentious topic; the French rhotic in particular varies in quality between a fricative and an approximant, and phonologically behaves as a sonorant (Wiese, 2001). Under this analysis, final sequences starting with [ʁ] do not violate the sonority sequencing principle. Since this does not make any material difference to the final parameter value, however, it is not treated in further detail here.

also violated in ⟨next, text⟩ and ⟨against⟩, as well as lower frequency items ⟨*tempt⟩ (e.g attempt) and ⟨glimpse⟩. Adhering to the minimum frequency limit of 5 Zipf, the coda set is therefore {[t]}.

French permits three consonants word finally. These sequences all take the form of a valid word-final consonant, followed by a sequence otherwise found word-initially (as outlined for two-segment sonority violating sequences in 4.8.3.13). These are described in Dell (1995) as a single coda followed by a complex onset; the overlap between these and word-initial branching onsets can be observed using Nidaba's set comparison tools.

German permits a single consonant following long vowels, or two following short vowels (Wiese, 2000). The exceptions to this are alveolar obstruents [s] and [t] (e.g. ⟨links, sanft⟩), and the sequence [st] (e.g. ⟨selbst, ernst⟩).

Greek permits a single consonant word-finally, except for the sonority violating sequences with [s] ([ks], [ts]). There are other exceptions at lower frequencies (< 5 Zipf) in loan items, such as [st].

Despite the well-documented use of long final consonant sequences in Polish (e.g. Gussmann, 2007), only one three-segment sequence is found more than one hundred times per million tokens: [rtv], in ⟨martw⟩. This is insufficient data to posit a set of segments.

Portuguese permits a limited set of single consonants word finally, with the exception of [s], following [k].

Spanish permits single consonants word-finally in the native stratum (J. W. Harris, 1983), with only a few two-segment sequences occurring more than one hundred times per million tokens. These sequences tend to occur in foreign items (e.g. York, Budapest), though not exclusively (e.g. récord, zinc). The finite list of exceptions do not form a coherent set; these sequences appear to be frequent solely because of the prevalence of certain non-Spanish names.

Cheke Holo does not have word-final consonants, with no exceptions. Hrusso Aka and Matbat permit single consonants word-finally, with no set of exceptions. Sylheti permits single consonants word-finally, with the exception of [nd]. Whilst this sequence is found in multiple lexical items, [d] does not otherwise participate in coda structure violations. Ambel permits sequences of two consonants word-finally; the first is always a glide, with no set of exceptions.

Lithuanian and Welsh permit sequences of two consonants word finally, with no set of exceptions. Romanian has a few lower frequency word-final consonant sequences of three segments: [nkt], [kst], and several that appear only in single lexical items (e.g. [astm] *asthma*). The final consonant in these sequences seems limited to [t], [s] or [m]. With a minimum frequency of 5 Zipf, there are no word-final sequences with three or more consonants.

Sonority distance parameters

Per Clements (1990), the parameters describing the first part of the syllable (onset) are independent of those describing the second part (rime): there is no parameter to describe the interaction of the two.

Not only do most languages require that onset clusters obey the Sonority Sequencing Principle, they may also require a minimum sonority difference between segments. There are two different models of this behaviour, neither of which fully accounts for all the observed types.

According to the Minimal Sonority Distance model (Steriade, 1982), each language has a minimum difference between segments in an onset, be that three steps (obstruent to glide) or zero, a sonority plateau (e.g. liquid-liquid). There is no opposing pressure to minimise sonority, so the default case is a stop to glide cluster, since that the largest sonority distance possible.s

According to the Sonority Dispersion Principle (Clements, 1990), the maximisation of sonority distance extends beyond the onset to the nucleus. In the default case, an onset cluster should be obstruent-liquid, since the liquid is maximally dispersed from both obstruent and vowel.

Some languages permit only glides as the second member of an onset cluster, as the MSD model would predict; some languages permit only liquids, as the SDP would predict (Parker, 2012). Others, like English, have a minimum sonority distance but no fixed requirement for the second consonant. To capture these differences, I have included Onset Gap parameters. These parameters are not applicable to Sylheti, which does not have onset clusters.

(18) **Onset Gap of** 0

Can an initial cluster contain a sonority step of length zero?

This is a cluster with a sonority plateau: two oral stops, two fricatives, two nasals, two liquids, or two glides.

Dutch, French, German, Greek, Hrusso Aka, Lithuanian, Matbat[14], Polish and Romanian have word-initial consonant sequences with sonority plateaus. All except Matbat also have all possible greater sonority steps.

Ambel, Cheke Holo, English, Portuguese, Spanish and Welsh do not have initial sonority plateaus.

(19) **Onset Gap of 1**

Can an initial cluster contain a sonority step of length one? (From obstruent to nasal, nasal to liquid, or liquid to glide.)

Ambel, Cheke Holo, English, Spanish, and Welsh have word-initial consonant sequences with a sonority step of length one. They all also have all possible larger sonority steps.

Matbat does not have any such sequences.

(20) **Onset Gap of 2**

Can an initial cluster contain a sonority step of length two? (From obstruent to liquid, or nasal to glide.)

Matbat and Portuguese have sequences with a sonority step of length two, as do all other languages in my sample except Sylheti, to which this parameter does not apply. Hrusso Aka has [ɾ] only in recent loanwords (D'Souza, 2015). With frequency data to impose a minimum threshold, this parameter might be false for Hrusso Aka.

(21) **Onset Gap of 3**

Can an initial cluster contain a sonority step of length three? (From obstruent to glide.)

[14] The Matbat lexicon has [mn-] sequences. Whilst Remijsen, 2010 states that Matbat syllable structure is (C)V(C), the paper contains the counterexample "hi^{21}p mni^{12}k" *rub oil*. In Ambel, another Raja Ampat language, [mC-] roots are realised with a vowel-final prefix, so such sequences never surface as word-initial (Arnold, pc.), but this does not appear to be the case for Magey Matbat.

All the languages in my sample have sequences with a sonority step of length three, except Sylheti, to which this parameter does not apply.

(22) **Obligatory Glide parameter**

Must the second consonant of an initial cluster be a glide?

Parker (2012) discusses two restrictions which languages may impose in addition to minimum sonority distance. The first restriction, which this parameter captures, is that the second consonant must always be a glide. The most unmarked structure for these languages is stop-glide, as in the minimum sonority distance model in general. This parameter requires the disambiguation of branching onsets with glides from diphthongs with an initial vowel, as discussed in the Branching Onset parameter.

All of the languages in my sample with two initial consonants allow for non-glides as the second consonant.

(23) **Obligatory Liquid parameter**

Must the second consonant of an initial cluster be a liquid?

The second potential restriction, mutually exclusive with an obligatory glide, is that the second consonant is obligatorily a liquid. The most unmarked structure in this case would be stop-liquid, as predicted by the Sonority Dispersion Principle. However, Parker found languages which also allowed nasal-liquid clusters, but not the obstruent-nasal which the Sonority Dispersion Principle predicts should be less marked. Therefore the differences between languages which have only liquid final clusters can be described perfectly adequately by combining the obligatory liquid parameter with the Onset Gap parameters.

All of the languages in my sample with two initial consonants allow for non-liquids as the second consonant.

4.8.4 Vowel inventory parameters

The vowel inventory parameters capture not fine or even broad phonetic detail, given the inherent difficulties of categorising vowels that way (Lass, 1984), but rather the presence or absence of phonological contrasts. They cover vowel height, ATR, backness and rounding (Rice, 2002); nasality, and phonation types.

(1) **Height parameter**

Does the vowel system have more than one contrast in height?

"Every phonological system contrasts at least two degrees of aperture" and therefore has at least one contrast in height (Hyman, 2008). The majority of languages have more than two heights (Maddieson, 1984).

All the languages in my sample had contrasts between (at least) three heights.

(2) **ATR contrast**

Is there at least one ATR or tense/lax contrast?

A language with a tense/lax contrast has an additional contrast in its front or back vowels on top of two existing height contrasts. For the purposes of this parameter, any vowel contrast which includes a quality difference is counted, regardless of whether there is also a corresponding length difference. This parameter depends on Parameter 4.8.4.1.

Crothers (1978) categorises [ɛ], [a] and [ɔ] as not (necessarily) contrasting in height, and hence most languages in his typology have no more than three distinct categories. However, I shall follow the common practice of categorising seven-vowel systems such as those of Italian and Yoruba as having a tense/lax or ATR contrast, rather than having a rounding contrast in the back vowels (e.g. Calabrese, 1998, Pulleyblank, 1996).

English, Dutch, German, and Lithuanian have at least four distinct categories. Even if the Dutch tense/lax contrast is instead analysed as a length contrast, French loan items give rise to a four-way contrast. The German /ɛ/ vs /e/ distinction is debated; I am here following Wiese (2000) and Baayen, Piepenbrock and Rijn (1993) in treating them as separate. French, Matbat

and Portuguese have a four-way contrast assuming that [ɛ] and [ɔ] are categorised as differing in height from [a].

The Welsh lexicon used in Nidaba evinces no tense/lax contrast, so this is the analysis I am following. However, there is disagreement on whether a certain category of contrast is more properly described as a length contrast or vowel quality contrast (Hannahs, 2013), with variation between speakers / dialects (Iosad, 2017).

Ambel, Cheke Holo, Greek, Spanish, and Sylheti all have five-vowel systems, with no ATR contrast. Hrusso Aka, Polish and Romanian have three contrasting vowel heights.

(3) **Multiple ATR contrasts**

Are there two or more ATR or tense/lax contrasts?

Such a language may also be described as having a five-way contrast in vowel height (Crothers, 1978, Lass, 1984). This parameter implies that 4.8.4.2 is true.

Dutch, English, German and Lithuanian have a tense/lax contrast in both high and mid vowels.

French and Matbat only have a single tense/lax contrast, between low-mid and low vowels. Portuguese does too, assuming that [ɐ], unlike [a], is not contrastive in height with [ɛ] or [ɔ].

This parameter is not applicable to Ambel, Cheke Holo, Greek, Hrusso Aka, Polish, Romanian, Spanish, Sylheti, or Welsh, which lack any tense/lax contrast.

(4) **Back parameter**

Does the vowel system have contrastive roundness or contrastive frontness?

This parameter captures the difference between vertical vowel systems, such as Kabardian, which only realise frontness or roundness on consonants or morphemes, and the more typical language with such a contrast inherent to vowels (Hyman, 2008).

All the languages in my sample contrast front unrounded vowels with back rounded vowels.

(5) **Front rounded parameter**

Is there a rounding contrast in the front vowels?

If so, 4.8.4.4 is true; the language has at least contrastive rounding. To avoid ambiguity in setting this parameter, the language must have at least one back or central vowel at the same height as the contrast, such that the front rounded vowel cannot be alternatively analysed as a central or back rounded vowel.

Dutch, French, and German have a rounding contrast in the front vowels.

Ambel, Cheke Holo, English, Greek, Hrusso Aka, Lithuanian, Matbat, Polish, Portuguese, Romanian, Spanish, Sylheti and Welsh do not.

(6) **Back unrounded parameter**

Is there a rounding contrast in the non-front vowels?

If so, 4.8.4.4 is true; the language has at least contrastive rounding. A vowel system may be described as having back rounded and back unrounded vowels, or back rounded and central unrounded vowels (e.g. Turkish, Rice, 2002); either of these contrasts sets this parameter as true. As in 4.8.4.5, there must be at least one front unrounded vowel at the same height as this contrast.

Polish and Portuguese have a contrast between high central unrounded and high back rounded vowels. Romanian and Welsh have a contrast in both high and mid vowels. Hrusso Aka has contrast in the high vowels, and a marginal contrast in the mid vowels. German has a contrast between mid central unrounded and mid back rounded vowels, though prosodically conditioned[15]. Lithuanian also has a contrast between [ʌ] and [o:], with a concomitant length distinction. English and Dutch have a contrast between central and back mid vowels, and rounded and unrounded low vowels.

Ambel, Cheke Holo, French, Greek, Matbat, Spanish and Sylheti do not have a rounding contrast in the non-front vowels.

(7) **Nasality parameter**

Does the vowel system have an oral / nasal contrast?

[15]Taking schwa to be a contrastive segment in German, following Féry (1991).

This parameter captures the difference between languages with no or allophonic nasal vowels (e.g. English) and languages which use vowel nasality contrastively (e.g. French). A language with nasal vowels will always have oral vowels too, giving an oral/nasal contrast.

French and Portuguese have nasal vowels. Polish is variously analysed with and without nasal vowels; I have chosen to categorise it as having an oral/nasal contrast in the vowel system, but the lexicon in Nidaba is transcribed with a nasal archiphoneme, allowing for alternative interpretations to be applied to the data. Hrusso Aka contains vowel nasalisation only marginally (see also D'Souza, 2015), with nasalisation present in only seven lexical items out of over 3200; but without token frequency data, I am not conclusively excluding it.

The English, Dutch and German lexicons from the CELEX database contained items transcribed with nasal vowels (i.e. French loanwords). These items are both small in number and infrequent. Furthermore, the loanwords are not (consistently) produced with nasal vowels, regardless of their transcription in CELEX (see e.g. Cambridge Dictionary, 2015).

Ambel, Cheke Holo, Greek, Lithuanian, Matbat, Romanian, Spanish, and Sylheti do not have a contrast between oral and nasal vowels.

(8) **Breathiness parameter**

Does the vowel system have a modal / breathy contrast?

Of the different phonation types, all languages have modal voicing in vowels, so any language with phonemic breathy vowels will have a contrast between modal and breathy phonation.

None of the languages in my sample have a modal / breathy contrast.

(9) **Creakiness parameter**

Does the vowel system have a modal / creaky contrast?

Breathiness and creakiness may both be used contrastively in vowels, including in the same language (Silverman et al., 1995), though this can only produce a three-way contrast.

None of the languages in my sample have this contrast.

The final phonation type, voicelessness, is only ever found predictably in vowels, in certain contexts. It is not used contrastively (Gordon and Ladefoged, 2001).

4.8.5 Consonant inventory parameters

The consonant parameters have been divided into three categories, dealing with contrasts in laryngeal, place and manner features.

Laryngeal parameters

(Almost) all languages with only one type of laryngeal specification have plain voiceless stops (Maddieson, 1984). We can view this as the unmarked case of stops; in a representation using privative features, the laryngeal features are unspecified for plain voiceless stops. Represented in binary features, [-voice, -spread glottis, -constricted glottis] is the unmarked case. The first three laryngeal parameters examine deviations from this default case. The other two parameters examine voicing contrast in fricatives and nasals, since the other laryngeal contrasts are sufficiently rare to be of less importance.

I am following Honeybone (2005) in treating aspiration and voicing as two separate cases, rather than simply as two instantiations of a single underlying contrast. An alternative approach could be parameters for a single contrast and for multiple contrasts. A language with a three-way contrast and a language with a two-way contrast would then have one of two parameters in common, just as in the approach I have chosen; whereas a language without a laryngeal contrast would have zero of two parameters in common with a two-way contrasting language and a voicing language would have one of two parameters in common with an aspirating language. Whilst such a choice might align more naturally with certain applications of a distance metric, the majority of languages have at least one laryngeal contrast in stops (Henton, Ladefoged and Maddieson, 1992), so I have instead chosen parameters to more evenly partition the language space. The alternative, contrast-counting, parameters could be derived from these if required.

(1) **Stop voicing parameter**

Does the language have a contrast between voiced and voiceless stops?

This parameter only describes those languages which have a voicing contrast, rather than the aspiration contrast of Parameter 4.8.5.2.

For languages like Hindi, which has both a voicing and an aspiration contrast, or like Ostyak, with neither, this parameter is straightforward.

In languages with only a single contrast, allophones of aspirated stops may appear as plain, and allophones of plain stops as voiced. A true voicing language will have the following characteristics (Honeybone, 2005, p. 330): Are all 'voiced' stops spontaneously voiced, as opposed to only passively voiced between sonorants? Is there voicing assimilation (i.e. a voiceless stop becomes voiced in the environment of voiced stop)?

Cheke Holo has both a voicing and an aspiration contrast; there exist minimal triplets (e.g. [dao] / [tao] / [t^{h}ao]).

Matbat is transcribed with a voicing contrast, with final stops being spontaneously voiced (Remijsen, 2007). Ambel (Arnold, pc.), Dutch (Honeybone, 2005), French[16], Greek (Honeybone, 2005), Lithuanian (Steriade, 2000), Portuguese, Romanian, Spanish and Sylheti (Eden, in press) are voicing languages. Polish is phonetically a voicing language (Gussmann, 2007); there is variation in phonological behaviour between the two major dialects, with Warsaw Polish behaving as a voicing language (Cyran, 2011).

For the purposes of this parameter, Hrusso Aka does not contrast voiced and voiceless stops: there is evidence that the two-way contrast in Hrusso Aka stops is aspiration-based, but as yet none for spontaneous voicing (D'Souza, 2015). English and German do not have a voicing contrast (Honeybone, 2005), and nor does Welsh (Ball, 1984, p. 15).

(2) **Stop aspiration parameter**

Does the language have a contrast between plain and aspirated stops?

As we have seen in 4.8.5.1, languages with an aspiration contrast may have phonetic voicing. The characteristics of an aspirating language are (Honeybone, 2005, p. 329): Do the 'aspirated' stops have aspiration in any environment? Is there 'devoicing' assimilation (i.e. a voiced stop becomes voiceless in the environment of an voiceless stop)?

[16]Romance languages in general are referred to as true voiced in multiple sources, including Honeybone, 2005; Iverson and Salmons, 2008; and Cyran, 2011. French is mentioned specifically in Cyran, 2011, and Spanish in Honeybone, 2005.

Cheke Holo is an aspirating language: aspirated sonorants are pronounced with initial spread glottis, which manifests as breathy voice on a preceding vowel, or plain voicelessness utterance-initially; while aspirated stops are post-aspirated (Corretta, pc.). Hrusso Aka has a contrast between voiced and aspirated stops: voiceless plosives are aspirated before high vowels, and optionally elsewhere; high vowels are devoiced following voiceless plosives (D'Souza, 2015). English and German are aspirating languages (Honeybone, 2005), as is Welsh (Ball, 1984, p. 15).

Ambel, Dutch, French, Greek, Lithuanian, Matbat, Portuguese, and Spanish are not aspirating languages. Matbat does not have aspiration or devoicing assimilation. Polish shows voice agreement, with obstruents assimilating to the voice (or voicelessness) of the following obstruent. Given the two-way laryngeal contrast, it is assumed that there is only one active process, with devoicing 'assimilation' a process of neutralisation, "similar to word-final devoicing" (Cyran, 2011). Romanian shows final devoicing of nasals in a voiceless environment, but no such effect on obstruents (Tucker and Warner, 2010). Sylheti is unusual for an Indo-Aryan language in that it lacks an aspiration contrast (Eden, in press).

(3) **Stop glottalisation parameter**

Does the language have a contrast between plain and glottalised stops?

For this parameter, a stop is considered glottalised if the airstream mechanism is glottalic (i.e. implosives and ejectives), or if the glottis is constricted to produce creaky consonants. There are no known languages which distinguish between laryngealized pulmonic and glottalic consonants (Maddieson, 1984), so this parameter covers both interchangeably.

None of the languages in my sample have such a contrast.

(4) **Fricative voicing parameter**

Does the language have a contrast between voiceless and voiced fricatives?

The majority of languages have voiceless fricatives (Maddieson, 1984), and in general, the presence of a voiced fricative implies the presence of a voiceless counterpart. However, this does not hold for all places of articulation – e.g. bilabial fricatives are more commonly voiced; and a voiced uvular fricative may be argued to belong to the class of liquids as a rhotic, rather than that of voiced fricatives.

For this reason, this parameter deals with the contrast between voiceless and voiced fricatives at the same place of articulation, not just the presence or absence of voiced fricatives in the language's inventory.

There are very few languages with aspirated or glottalised fricatives (Maddieson, 1984), so I am not including parameters for fricatives which parallel those for stops.

This parameter depends on Parameter 4.8.5.16, the presence of fricatives in the language.

Ambel and Matbat have only voiceless fricatives.

Cheke Holo, Dutch, English, French, German, Greek, Hrusso Aka, Lithuanian, Polish, Portuguese, Romanian, Spanish, Sylheti, and Welsh have a voicing contrast.

(5) **Nasal voicing parameter**

Does the language have a contrast between voiceless and voiced nasals?

All languages with nasals have plain voiced nasals (i.e. modally voiced nasals with no secondary articulation) (Maddieson, 1984), so any language with a voiceless nasal will have this contrast.

This parameter does not cover the contrast between modal voicing and breathy or aspirated nasals, just as Parameter 4.8.5.1 does not. However, there are so few languages which contrast glottalisation or breathiness in nasals that, as with fricatives, I am not including parameters which cover those contrasts.

Welsh has a nasal voicing contrast; voiceless nasals appear in a 'nasal mutation' context, as 'reflexes of initial voiceless stops' (Hannahs, 2013).

The other languages in my sample do not; Romanian has allophonic nasal devoicing, but no contrast (Tucker and Warner, 2010).

Obstruent place parameters

The vast majority of languages have plosives at three places of articulation: labial, dental/alveolar and velar. Additional contrasting places of articulation are, in order of frequency, palatal, uvular, retroflex, labio-velar, and finally a contrast between dentals and alveolars. However,

these additional places are fairly infrequent, found in 10% of languages or fewer. The 'place-less' plosive, by contrast, divides the languages in UPSID almost equally: approximately half of languages have a glottal stop.

(6) **Glottal stop parameter**

Does the language have a glottal stop?

For this parameter to be true, the sound must be phonemic, not just be phonetically inserted into pauses; it must contrast with other stops, not just zero.

Cheka Holo has glottal stops.

The other languages in my sample do not. Various dialects of English employ glottal stops as allophones of /t/, but not the dialect on which this current analysis is based (e.g. Hughes, Trudgill and Watt, 2013).

(7) **Secondary articulation series parameter**

Does the language contain consonants which contrast solely in secondary place of articulation?

That is, does the language have a series of secondarily articulated consonants which parallels another series of consonants? E.g. Irish velarized and palatalized consonants, Russian plain and palatalized consonants. If the language contains only a single secondarily articulated obstruent, this is not considered to be a parallel series (e.g. labialized velar in Molinos Mixtec, Hunter and Pike, 1969).

Lithuanian (Kenstowicz, 1972) and Polish[17] (Gussmann, 2007) have secondary palatal series. None of the other languages in my sample have a contrasting series of secondarily articulated consonants.

Nasal place parameters

The vast majority of languages with nasals have both a bilabial nasal and a dental or alveolar nasal. Since these are so prevalent, parameters examining them would not evenly partition the

[17]Whether the Polish series is a feature of the inventory or morphophonology is a subject of some debate; I am here following Gussmann(2007, p. 99) in assuming it is 'lexical, unpredictable, underlying'; i.e. a contrast located in the obstruents.

language space. I therefore include parameters for whether a language has the next most common types: velar nasals or palatal nasals.

(8) **Velar nasal parameter**

Does the language have velar nasal phonemes?

Approximately half of languages use velar nasals (Maddieson, 1984). The majority of languages with only three nasals have a velar nasal as the third.

Cheke Holo, English, Dutch, German, Hrusso Aka, Matbat and Welsh have velar nasals. French has velar nasals only in English loan items, with a total frequency of 105 items / million.

Spanish has velar nasals after nasal place assimilation, but not in contrast to other nasals (J. W. Harris, 1984). Ambel, Greek, Lithuanian, Polish, Portuguese, Romanian, and Sylheti do not have velar nasals.

(9) **Palatal nasal parameter**

Does the language have palatal or palato-alveolar nasal phonemes?

Few, if any, languages contrast palatal with palato-alveolar nasals. Whilst they are less common than velar nasals, palatals may form the third nasal in an inventory, or, more commonly, the fourth.

Cheke Holo, French, Hrusso Aka, Lithuanian, Polish, Portuguese and Spanish have palatal nasals.

Ambel, Dutch, English, German, Greek, Matbat, Romanian, Sylheti and Welsh do not.

(10) **Word-final nasal place parameter**

Do nasal stops contrast in place word-finally?

Whilst most languages have some contrast between bilabial and dental/alveolar nasals, many lose that contrast word-finally, particularly those which do not otherwise have word-final obstruents; e.g. Japanese (Vance, 2008).

This parameter depends on Parameter 4.8.3.4, the presence of word-final consonants.

All the languages in my sample have a contrast in place between word final nasals, excepting Cheke Holo, which does not have word-final consonants at all.

Fricative place parameters

These parameters depend on Parameter 4.8.5.16, the presence of fricatives in the language.

(11) **Dental/alveolar fricative parameter**

Does the language have an interdental, denti-alveolar ('dental') or laminal alveolar fricative?

The most common place of articulation for a fricative is dental/alveolar, with the majority of languages not distinguishing between these two places. In terms of distinctive features, most languages have a [+anterior] fricative, but few distinguish between [+anterior, +distributed] and [+anterior, −distributed].

All the languages in my sample have at least one of these fricatives.

(12) **H parameter**

Does the language have /h/?

The glottal or 'placeless' fricative is the next most common fricative, with two-thirds of languages having some kind of voiceless laryngeal continuant.

Cheke Holo, Dutch, English, German, Hrusso Aka, Matbat, Romanian, Sylheti, Welsh have a glottal fricative.

Ambel, French, Greek, Lithuanian, Polish, Portuguese, Spanish do not.

(13) **Palato-alveolar fricative parameter**

Does the language have a palato-alveolar fricative?

The next most common place of articulation for a fricative is palato-alveolar.

Palatal fricatives are uncommon enough that I am not including a parameter examining them here. However, their appearance is independent of palato-alveolars, since the probability

of there being palatal fricatives is the same in languages with and without palato-alveolars (Maddieson, 1984), so they are not considered to contribute to this parameter. In terms of distinctive features, this parameter examines [+coronal] segments, not [+high] ones.

Ambel, Greek, Matbat and Spanish do not have a palato-alveolar fricative. Cheke Holo does not have palato-alveolar fricatives, provided that the sounds transcribed [tʃ] and [dʒ] are affricates. This is supported by distributional data: the sounds [ʃ] and [ʒ] only occur as components of [tʃ] and [dʒ] respectively, and the only 'three segment sequence' in the language is [tʃr]. Polish has an alveolo-palatal fricative, which Maddieson (1984) classes with palatals; I shall follow this convention here.

Dutch, English, French, German, Hrusso Aka, Lithuanian, Portuguese, Romanian, Sylheti and Welsh do have a palato-alveolar fricative.

(14) **Labial fricative parameter**

Does the language have a labial fricative?

The labio-dental fricative /f/ is the third most common fricative. Since very few languages contrast bilabial and labiodental fricatives, this parameter also includes bilabial fricatives. This scarcity may also be why there is no general consensus on which distinctive features are necessary to represent labiodentals (Odden, 2005, Hayes, 2008).

This also avoids the necessity of deciding which articulation is the underlying form in any given language. For example, Dizi has the fricative inventory [s], [z], [ʃ], [ʒ], [f], [β] and [h] (Maddieson, 1984). [f] and [β] could be considered to contrast in place, with predictable voicing, or, given the patterning of the other fricatives, to pattern in voicing with predictable place of articulation. In the latter case, much more data is required to decide which of the two places is underlying.

All the languages in my sample have a labial fricative.

(15) **Velar fricative**

Cheke Holo, Dutch, German, Greek, Hrusso Aka, Lithuanian, Polish, Spanish and Sylheti have a velar fricative. Ambel, English[18], French, Matbat, Portuguese, Romanian and Welsh do

[18] CELEX contains the single example ⟨ugh⟩, though this is para-linguistic.

not.

Manner parameters

(16) **Fricative parameter**

Does the language have fricatives?

A fricative is defined as a continuant, produced throughout with constriction leading to turbulent airflow, and acoustically, to noise. This excludes both affricates and fricative vowels.

Over 90% of languages have fricatives (Maddieson, 1984), so two languages both having fricatives is not particularly meaningful. However, this parameter is a necessary prerequisite to the larygneal and place fricative parameters. (For this reason, laryngeal continuants are included under this parameter, despite their variable classification.)

All the languages in my sample have fricatives.

(17) **Sonorant laterality parameter**

Is there a contrast between lateral and non-lateral sonorants?

That is, does the language have sonorants which have the same manner and place of articulation, and laryngeal specification, and contrast only in lateral articulation?

This contrast exists in English between /ɹ/ and /l/ and Spanish between /j/ and /ʎ/.

Dutch and German have lateral and non-lateral sonorants at different places of articulation. The French rhotic is a uvular fricative, so does not contrast with the alveolar lateral approximant for this parameter.

Ambel, Cheke Holo, Greek, Lithuanian, Matbat, Polish, Romanian, and Welsh have alveolar rhotic and lateral sonorants, but the rhotics differ in manner, being trills. Likewise, Portuguese and Sylheti have alveolar taps as counterparts to alveolar lateral approximants. Hrusso Aka has two laterals and two rhotics: /l/, /ʎ/, /ʁ/ and marginally /ɾ/ (D'Souza, 2015) but they do not contrast in manner and place simultaneously.

(18) **Contrasting lateral sonorants parameter**

Are there two or more contrasting lateral sonorants?

That is, are there sonorants with lateral articulation which contrast in place of articulation?

There are very few languages which contrast more than two lateral sonorants, so I am not distinguishing here between those languages with only one contrast and those few languages with more than one.

Ambel, Cheke Holo, Dutch, French, German, Greek, Matbat, Polish, Romanian, Sylheti and Welsh each have only a single lateral sonorant. The English alveolar lateral approximant may be syllabic or non-syllabic, but does not contrast in place of articulation.

Hrusso Aka, Portuguese, and Spanish have a contrast between alveolar and palatal lateral approximants. Lithuanian has a contrast between alveolar and palatalised alveolar lateral approximants.

(19) **Contrasting non-lateral liquids parameter**

Are there two or more contrasting non-lateral liquids?

Since over 97% of languages have two or fewer r-sounds (Maddieson, 1984), I am not including a parameter to separate out the small minority of languages which have more than two.

Lithuanian has a contrast between a palatalised and non-palatalised alveolar trill. Portuguese has a contrast between an alveolar flap and a uvular trill. Spanish has a contrast between tapped and trilled alveolars. Sylheti has a contrast between dental and retroflex flaps[19]. Welsh has a voicing contrast in its alveolar trills.

Ambel, Cheke Holo, Dutch, English, French, German, Greek, Hrusso Aka, Matbat, Polish and Romanian do not have a contrast within the category of non-lateral liquids.

(20) **Lateral obstruent parameter**

Does the language contain any lateral obstruents?

For example, fricatives, as in Welsh, or affricates, as in Navajo. Of the languages in UPSID, only 42 – 11% of languages with laterals – have lateral obstruents (Maddieson, 1984), but this parameter is included for completeness.

Of the languages in my sample, only Welsh has lateral obstruents.

[19]The retroflex flap [ɽ] is mostly allophonic with the voiced retroflex stop [ɖ] except in certain loan items, as in many Indo-Aryan languages (Masica, 1991). In neighbouring Assamese, the retroflex and dental flaps have merged into a single rhotic.

TABLE 4.11: Syllable structure parameter values

	Ambel	Cheke Holo	Dutch	English	French	German	Greek	Hrusso Aka	Lithuanian	Matbat	Polish	Portuguese	Romanian	Spanish	Sylheti	Welsh
Obligatory onset	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Double onset	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓
Complex nucleus	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗	✗	✓	✓	✓	✓	✓
Coda	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Triple onset	✗	✗	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✓
Double coda	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗	✓	✗	✓	✗	✗	✓
Superheavy rime	-	-	✓	✓	✓	✓	-	-	✓	-	-	✓	✓	✓	✓	✓
Syllabic liquid	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Syllabic nasal	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Syllabic fricative	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Syllabic stop	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Word-initial sonority violation	✗	✗	✓	✓	✓	✓	✓	✗	✓	✗	✓	✗	✓	✗	✗	✓
Initial sonority violations set	-	-	✓	✓	✓	✓	✓	-	✓	-	✗	-	✓	-	-	✓
Onset structure violations set	✗	✗	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✓	✗	✗	✓
Word-final sonority violation	✗	✗	✓	✓	✓	✓	✓	✗	✗	✗	✓	✗	✓	✗	✗	✓
Final sonority violations set	-	-	✓	✓	✗	✓	✓	-	-	-	✓	-	✓	-	-	✓
Coda structure violations set	-	✗	✓	✓	✗	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗
Onset gap 0	✗	✗	✓	✗	✓	✓	✓	✓	✓	✗	✓	✗	✓	✗	-	✗
Onset gap 1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	-	✓
Onset gap 2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	-	✓
Onset gap 3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	-	✓
Obligatory glide	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	-	✗
Obligatory liquid	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	-	✗

Table 4.12: Vowel parameter values

	Ambel	Cheke Holo	Dutch	English	French	German	Greek	Hrusso Aka	Lithuanian	Matbat	Polish	Portuguese	Romanian	Spanish	Sylheti	Welsh
Height	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ATR contrast	✗	✗	✓	✓	✓	✓	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗
Multiple ATR contrasts	-	-	✓	✓	✗	✓	-	-	✓	✗	-	✗	-	-	-	-
Back	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Front rounded	✗	✗	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Back unrounded	✗	✗	✓	✓	✗	✓	✗	✓	✓	✗	✓	✓	✓	✗	✗	✓
Nasality	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗
Breathiness	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Creakiness	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

Table 4.13: Consonant parameter values

	Ambel	Cheke Holo	Dutch	English	French	German	Greek	Hrusso Aka	Lithuanian	Matbat	Polish	Portuguese	Romanian	Spanish	Sylheti	Welsh
Stop voicing	✓	✓	✓	✗	✓	✗	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗
Stop aspiration	✗	✓	✗	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓
Stop glottalisation	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Fricative voicing	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓
Nasal voicing	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Glottal stop	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Secondary series	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗
Velar nasal	✗	✓	✓	✓	✓	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗	✓
Palatal nasal	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✓	✓	✗	✓	✗	✗
Final nasal	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Dental fricative	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Glottal fricative	✗	✓	✓	✓	✗	✓	✗	✓	✗	✓	✗	✗	✓	✗	✓	✓
Palato-alveolar fricative	✗	✗	✓	✓	✓	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓	✓
Labial fricative	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Velar fricative	✗	✓	✓	✗	✗	✓	✓	✓	✓	✗	✓	✗	✗	✓	✓	✗
Fricatives	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Sonorant laterality	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗
Contrasting laterals	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗	✓	✗	✗
Contrasting non-laterals	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	✓	✓	✓
Lateral obstruent	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓

4.9 Hamming Distance

4.9.1 Method

Given the 16 languages described above, there are 120 unique language pairs. For each pair of languages under examination, I have assigned each parameter a value of 1 (if its value differs between the languages), 0 (if it is the same), or N/A. The Hamming Distance H between the two languages is calculated using $H = \frac{d}{i+d}$, where d is the number of differently-valued parameters, and i is the number of identically-valued parameters, as explained in Subsection 4.2.2 on page 41.

These values can be found in Table 4.14, and plotted in Figure 4.1. Since Hamming Distance produces a symmetric result, the values are mirrored across the diagonal.

Another possible visualisation of the resulting similarities is Figure 4.2 on page 101. This unrooted tree was calculated using the 'Fitch' and 'DrawTree' programs of the PHYLIP package (Felsenstein, 1989).[20]

TABLE 4.14: Hamming distances

	Ambel	Cheke Holo	Dutch	English	French	German	Greek	Hrusso Aka	Lithuanian	Matbat	Polish	Portuguese	Romanian	Spanish	Sylheti	Welsh
Ambel		0.17	0.32	0.32	0.30	0.36	0.15	0.26	0.32	0.13	0.26	0.23	0.23	0.15	0.17	0.34
Cheke Holo	0.17		0.34	0.34	0.34	0.34	0.26	0.17	0.36	0.21	0.32	0.32	0.32	0.19	0.22	0.34
Dutch	0.32	0.34		0.12	0.15	0.04	0.16	0.29	0.18	0.35	0.20	0.32	0.10	0.35	0.28	0.20
English	0.32	0.34	0.12		0.23	0.08	0.24	0.29	0.25	0.35	0.28	0.32	0.14	0.35	0.35	0.12
French	0.30	0.34	0.15	0.23		0.19	0.22	0.29	0.20	0.33	0.22	0.22	0.16	0.31	0.33	0.25
German	0.36	0.34	0.04	0.08	0.19		0.20	0.25	0.22	0.39	0.24	0.36	0.14	0.39	0.33	0.16
Greek	0.15	0.26	0.16	0.24	0.22	0.20		0.29	0.22	0.29	0.16	0.31	0.14	0.23	0.24	0.28
Hrusso Aka	0.26	0.17	0.29	0.29	0.29	0.25	0.29		0.27	0.29	0.23	0.23	0.27	0.23	0.24	0.29
Lithuanian	0.32	0.36	0.18	0.25	0.20	0.22	0.22	0.27		0.41	0.18	0.18	0.16	0.20	0.26	0.26
Matbat	0.13	0.21	0.35	0.35	0.33	0.39	0.29	0.29	0.41		0.40	0.27	0.31	0.27	0.19	0.38
Polish	0.26	0.32	0.20	0.28	0.22	0.24	0.16	0.23	0.18	0.40		0.25	0.18	0.29	0.31	0.32
Portuguese	0.23	0.32	0.32	0.32	0.22	0.36	0.31	0.23	0.18	0.27	0.25		0.24	0.16	0.21	0.31
Romanian	0.23	0.32	0.10	0.14	0.16	0.14	0.14	0.27	0.16	0.31	0.18	0.24		0.29	0.21	0.14
Spanish	0.15	0.19	0.35	0.35	0.31	0.39	0.23	0.23	0.20	0.27	0.29	0.16	0.29		0.14	0.35
Sylheti	0.17	0.22	0.28	0.35	0.33	0.33	0.24	0.24	0.26	0.19	0.31	0.21	0.21	0.14		0.30
Welsh	0.34	0.34	0.20	0.12	0.25	0.16	0.28	0.29	0.26	0.38	0.32	0.31	0.14	0.35	0.30	

[20] Figure 4.2 is a representation on a two-dimensional page of a multi-dimensional web of distances, and so cannot be used to infer relationships between languages. For example, similarity is not transitive; just because two languages A and B are similar, and A is similar to a third language C, this does not necessarily mean that B and C are similar, despite the visualisation. It depends whether the parameters that A and B share are the same parameters that A and C share. For example, the Hamming Distance between Greek and Ambel is small (0.15), as is the Hamming Distance between Greek and Dutch (0.16). But the distance between Ambel and Dutch is not small (0.32).

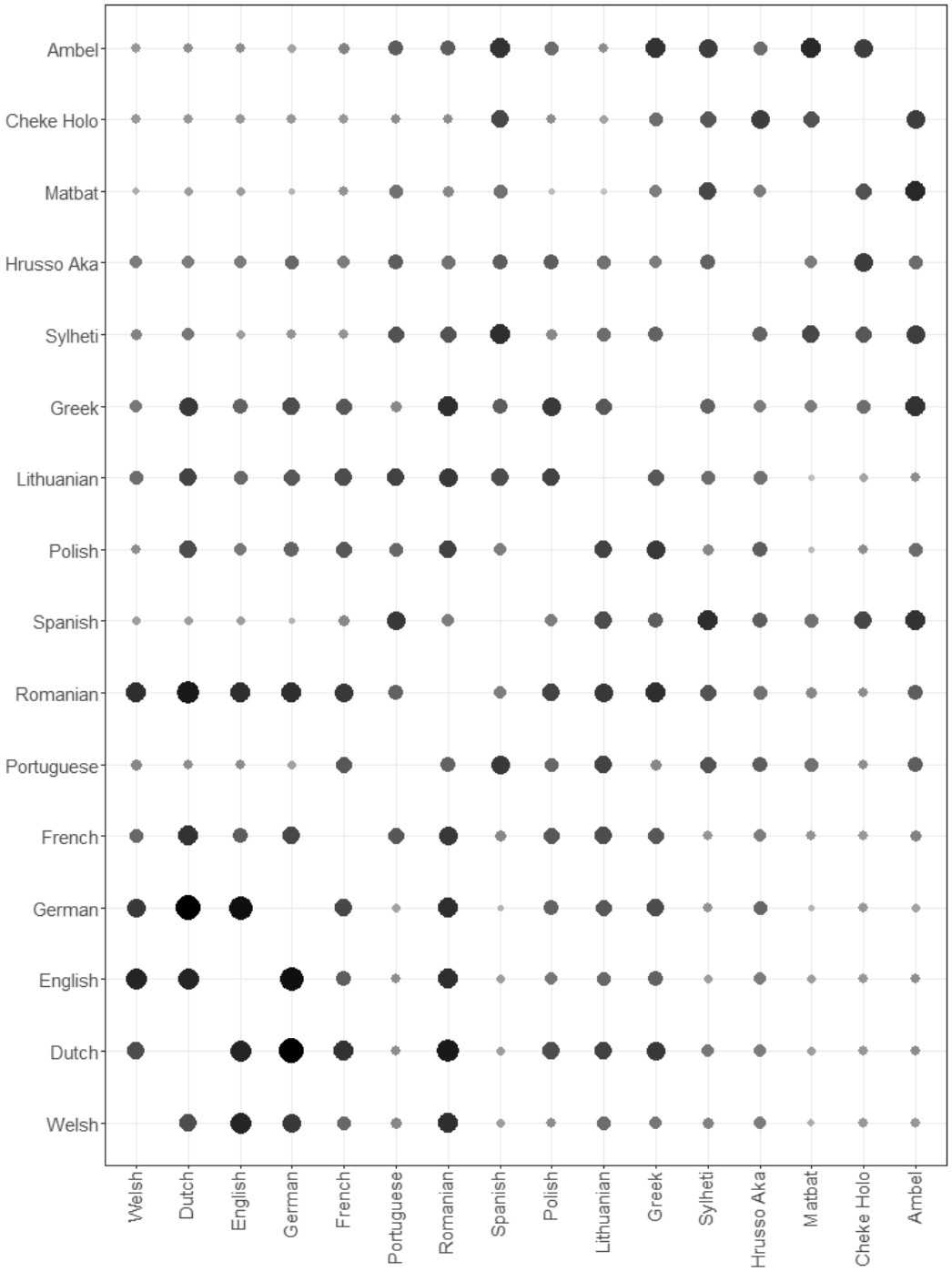


FIGURE 4.1: Heatmap of Hamming Distances; larger, blacker points are closer, smaller, greyer points are further.

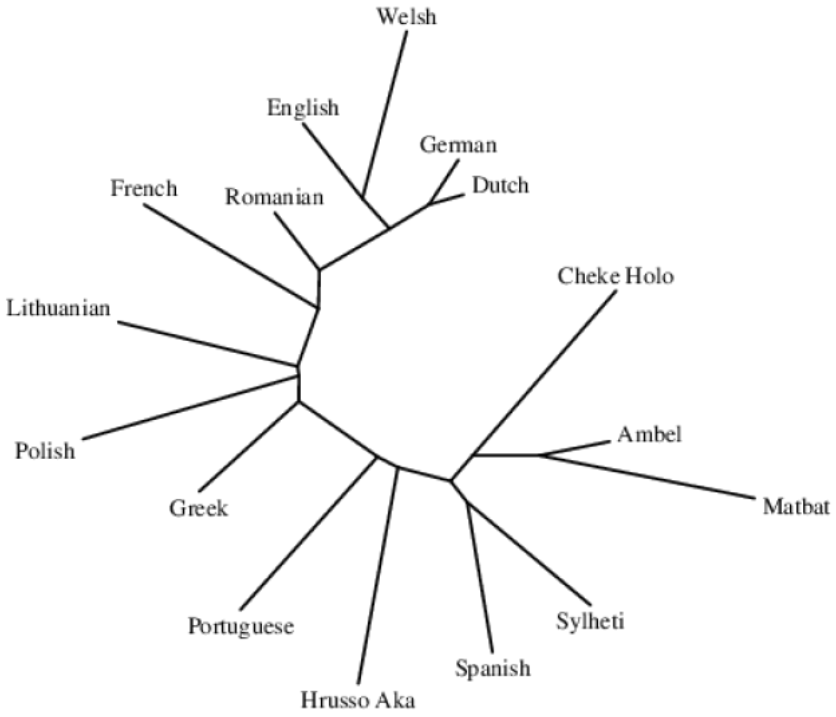


FIGURE 4.2: Visualisation of Hamming Distances (Felsenstein, 1989)

4.9.2 Significant similarity

As we saw in Subsection 4.2.2, parametric similarity between languages is only significant when these values are drawn from a sufficient total number of parameters.

Table 4.15 lists those language pairs where the probability of their high similarity arising at random is <1 in 10^{-5}, assuming both values of each binary parameter are equally likely, and all parameters are strictly independent. In that case, such similarity would imply a relationship between those language pairs. Indeed, some pairs are sisters (West Germanic, Raja Ampat, Iberian Romance); some are neighbours (English/Welsh; Dutch/French; Lithuanian/Polish).

However, not all parameter values are equally likely to occur. Deviations from the canonical CV syllable structure are not as common as the default, by definition. Ambel and Greek share a syllable shape inventory of (C)CV(C), lacking variations such as three-consonant initial sequences or two-consonant final sequences, falling initial sonority or rising final sonority sequences, or syllables which are exceptions to the standard shapes. Likewise, "considerably more languages have an inventory of five vowels than any other number" (Maddieson, 2011),

and Ambel and Greek share the 'average vowel inventory' of five monopthongs.

Cheke Holo and Hrusso Aka similarly have fewer deviations from the canonical syllable structure. With much more limited lexicons available compared to Indo-European languages, the negative evidence for sonority or structure violations is also less compelling than for Spanish or Portuguese, for example.

The remaining language pairs have statistically insignificant similarities, so nothing can be inferred about the historical relationship between them from these results.

TABLE 4.15: Language pairs with significant overlap in parameter similarity

Languages		Hamming distance	Identically valued parameters	Differently valued parameters	Total relevant parameters	
Dutch	German	0.04	50	2	52	West Germanic
English	German	0.08	48	4	52	West Germanic
Dutch	Romanian	0.10	46	5	51	Indo-European
Dutch	English	0.12	46	6	52	West Germanic
English	Welsh	0.12	45	6	51	*neighbours*
English	Romanian	0.14	44	7	51	Indo-European
German	Romanian	0.14	44	7	51	Indo-European
Romanian	Welsh	0.14	44	7	51	Indo-European
Ambel	Matbat	0.13	41	6	47	Raja Ampat
Greek	Romanian	0.14	43	7	50	Indo-European
Dutch	French	0.15	44	8	52	*neighbours*
French	Romanian	0.16	43	8	51	Romance
German	Welsh	0.16	43	8	51	Indo-European
Ambel	Greek	0.15	40	7	47	
Ambel	Spanish	0.15	40	7	47	
Dutch	Greek	0.16	42	8	50	Indo-European
Greek	Polish	0.16	42	8	50	Indo-European
Lithuanian	Romanian	0.16	42	8	50	Indo-European
Spanish	Sylheti	0.14	37	6	43	Indo-European
Portuguese	Spanish	0.16	41	8	49	Iberian Romance
Dutch	Lithuanian	0.18	42	9	51	Indo-European
Cheke Holo	Hrusso Aka	0.17	39	8	47	
Lithuanian	Portuguese	0.18	41	9	50	Indo-European
Polish	Romanian	0.18	41	9	50	Indo-European
French	German	0.19	42	10	52	*neighbours*
Ambel	Cheke Holo	0.17	38	8	46	Malayo-Polynesian
Lithuanian	Polish	0.18	40	9	49	Balto-Slavic
Dutch	Welsh	0.20	41	10	51	Indo-European
French	Lithuanian	0.20	41	10	51	Indo-European

4.9.3 Weighting

It is possible to account for asymmetries in typology by, for example, assigning a weighting proportional to the percentage of languages which share that parametric value. This system would assign a greater similarity to languages which shared a marked value than an unmarked one. This would compensate for the effect described above, making this metric more useful for probing the historical relationship between languages.

However, a metric is by definition symmetric; measuring from a phonologically 'standard' language to an unusual one should give the same distance as the reverse. It is possible to apply a weighting asymmetrically, so as to be useful in asymmetric processes such as intelligibility or acquisition (see Chapter 2). But in using a weighting for synchronic, rather than historical, research, there is the risk of begging the question: using acquisition observations to establish weightings to derive a distance metric to explain acquisition observations.

4.10 Conclusion

It is possible to measure the similarity of phonological representation systems using typological observations, formulated in either parameters or constraints.

Nidaba is a computational tool for assisting in making typological observations, and is designed to be configurable and extensible software, so as to enable users to make differing theoretical assumptions based on the same data.

I have applied a test set of 52 syllable structure and segment inventory parameters to 16 languages, and measured the resulting Hamming Distances between each language pair. The findings broadly accord with intuitive observations[21]: Dutch, English and German resemble each other more than French, Portuguese and Spanish do; and Portuguese is very similar to a Baltic language. This method is equally applicable to Tibeto-Burman and Austronesian languages as Indo-European, with no dependency on cognacy or historically significant features.

This method therefore provides a reproducible way of quantifying similarity between any pair of languages using only a lexicon of ~1000 items.

[21] Both my own impressions, and an informal survey of phonologists.

Chapter 5

Cross-Entropy

In this chapter, I move away from a static representation of phonological systems, to a similarity metric based on the cross-entropy of phonemically and featurally transcribed example texts. The advantage of such a metric is a large reduction in the amount of input data required, and in the completeness of analysis of a given language.

In Section 5.1 I discuss the basic concept of cross-entropy; in Section 5.2 the choice of notation to use in representing an extract of speech; in Section 5.3 the different approaches for calculating entropy. Section 5.4 summarises the methodology, with the prototype described in Section 5.5 and the full application in Section 5.6.

5.1 Background

5.1.1 What is entropy?

Entropy is a measure of randomness. It is used in physics to describe the disorder of a system, and in information science to describe the efficiency of information transfer.

For example, let us take a message like: 'aaaaaaaaaaaaaaaaaaaaaaaaaa'. This can either be transmitted as 26 individual characters, or as 'a, 26 times'.

The message 'abcdefghijklmnopqrstuvwxyz' cannot be compressed like that, since every character is different. However, it can be transmitted as 'the Roman alphabet'. That is, given some existing knowledge of the system - the order in which letters usually appear in the Roman alphabet - the new message is more predictable than if you had to guess the order of 26 characters at random.

The same is true for the transmission of any string of characters. The character 't' has a much higher probability of being followed by 'h' in English than by 'g', so receiving the message 'the thing' is much more likely than 'tge tging', which is more likely than 'tgb tglkh', and so on.

5.1.2 Shannon entropy

Shannon (1948) shows that the most efficient encoding is where the length of the representation in bits[1] is $-log_2 p_i$, where p_i is the probability of some unit of representation i. That is, compared to a standard word W, a word that is half as frequent as W should have a representation twice as long; a word which occurs twice as frequently as W should have a representation which is half as long.

If English were efficiently encoded, we could represent 'the' with 1 bit, ("Is this word 'the'?), 'of' with 2 bits ("Is this word 'the'? Is this word 'of'"?) and so on. By contrast, since the English alphabet requires 5 bits per letter[2], English encoded as a series of letters requires 15 bits for 'the', 10 for 'of', and so on. Therefore, written material can be compressed to require fewer bits, without loss of information. Substituting a word-frequency based representation for the written representation is just one of the possible techniques.

The maximally efficient encoding corresponds to Shannon's entropy $H(M)$, and is:

$$H(M) = \sum_i (-log_2 p_i) \cdot (p_i)$$

That is, the entropy of a message is the sum of the lengths of the efficiently encoded representations, each multiplied by the probability of its occurrence.

An entirely predictable system has an entropy of zero, since the probability of that system is 1, and $-log_2(1) = 0$; i.e. no question needs to be answered for the state of the system to be known. A system which has two equally likely states – e.g. the answer to a yes/no question – has an entropy of 1 bit (H(M) = $(-log_2(0.5) \times 0.5) \times 2$); the answer to 1 binary decision is needed to know the state of the system.

[1] A bit is a binary digit, whose two values are frequently represented as 0 or 1. It can be viewed as the answer to a yes-no question.

[2] 5 bits gives 2^5 possibilities, which can represent up to 32 characters. 2^4 can represent up to 16, which is obviously insufficient.

A system cannot have negative entropy; knowing its state cannot require fewer than 0 binary questions to be answered.

5.1.3 Cross-entropy

In order to achieve the maximally efficient encoding, we must perfectly know the probability of occurrence of every word in the message. Assuming we do not, we must use an estimated distribution Q to decide on the lengths of the encodings. Q will not be the same as the actual probability distribution P, and so the encoding it produces is less efficient.

The entropy of a system which has been encoded using the estimated distribution Q is called the cross-entropy:

$$H(P,Q) = \sum_i (-log_2 q_i) \cdot (p_i)$$

That is, the cross-entropy is the sum of the lengths of the representations (derived from the estimated probabilities), each multiplied by the true probability. This cross-entropy is minimised when $P = Q$ (i.e. the estimated distribution is the same as the true distribution).

The difference in entropy between a system encoded using Q and one using P is called the Kullback-Leibler divergence:

$$\begin{aligned} D_{KL}(P||Q) &= H(P,Q) - H(P) \\ &= \sum_i (-log_2 q_i) \cdot (p_i) - \sum_i (-log_2 p_i) \cdot (p_i) \end{aligned}$$

Since there is no theoretical difference between an accurate and an inaccurate distribution, the same technique can be applied to any two distributions P and Q, whether P is actually the true distribution, or is in reality just another estimate. This should produce a positive Kullback-Leibler divergence; if not, then the approximation Q is in fact more accurate than the 'true' distribution P.

Two languages can therefore be compared using the Kullback-Leibler divergence even if we do not know the true probability distribution of character sequences for them, provided we have a reasonable estimate for each.

For example, we can derive two encodings for English: the first based on our reasonable estimate of the probability distribution of English, and the second based instead on German. We shall label the distribution derived from English the 'true' distribution P, and the one derived from German an approximate distribution Q. The cross-entropy of these estimates will be called $H(\text{English, German})$.

If instead we wanted to apply these distributions to a German text, we would label the German-based 'true' distribution P, and the English-based approximation Q. The cross-entropy of these estimates would be called $H(\text{German, English})$.

The Kullback-Leibler divergence for each of these situations will not necessarily be the same:

$$H(\text{English, German}) - H(\text{English}) \not\equiv H(\text{German, English}) - H(\text{German})$$

Therefore, the Kullback-Leibler divergence cannot strictly be called a metric, since it is not symmetrical. However, this may be beneficial in modelling human understanding and acquisition of language, which can also be asymmetrical between language pairs. To produce a metric which is comparable to those derived in Chapter 4 and Chapter 6, the average of the two can be used.

We can then compare the cross-entropy $H(\text{English, German})$ with $H(\text{English, Spanish})$ and $H(\text{English, Dutch})$. The pair of languages with the smallest Kullback-Leibler divergence have more similar encodings, which means that knowledge of one system in that pair is likely to translate accurately into knowledge of the other. For example, if $H(\text{English, Dutch})$ had the smallest Kullback-Leibler divergence, and $H(\text{English, Spanish})$ the largest, a Dutch speaker would be more likely to correctly guess whether [#st-] occurs in English than a Spanish speaker, based solely on their own language.

The above examples kept one language (English) constant across the comparisons. However, this is not a requirement of the metric. By using the Kullback-Leibler divergence, we can

control for a system having an inherently higher or lower entropy, and compare across all language pairs, even if they do not have a language in common; e.g. comparing $H(\text{English, Dutch})$ with $H(\text{German, Spanish})$.

Whether phonemic representations vary in redundancy depending on the language will be examined below.

Finally, this language distance can be normalised to a scale between 0 and 1. When one estimate is as good as the other, the Kullback-Leibler divergence will be 0. Since entropy cannot be negative, the maximum Kullback-Leibler divergence occurs when one estimate predicts an entropy of 0, and the other estimate predicts the maximum possible entropy of that system. Maximum entropy means maximum uncertainty, i.e. every possibility is equally likely.

$$H_{max} = \sum_{i}^{N}(-log_2\frac{1}{N}) \cdot (\frac{1}{N}) = -log_2\frac{1}{N}$$

where N is the number of possible states.

We normalise the metric by dividing the Kullback-Leibler divergence by this maximum.

5.2 Representation

There are myriad options for representing languages in a suitable format for entropy estimation. For our purposes, entropy estimation requires a linear sequence of characters, known as a string. A character is any discrete representation of a concept. The most common characters are orthographic - letters, punctuation and numerals - but characters may also be concepts without a standard visual representation. Possible phonological characters include phonemes, tones, stress, distinctive features (voicing, syllabicity, nasality, etc), and combinations of distinctive features (which I shall call feature bundles).

There is broad consensus that phonological representations are discrete, so I shall not here examine the measurement of entropy in a continuous system. It is however a possibility for anyone wishing to apply the same methodology to phonetic variables, for example.

5.2.1 Orthography

Classification of written documents using entropy-based algorithms has been an active area of research for decades. But there are three obvious limitations of using orthographic texts as the input to a model of phonological similarity. Firstly, languages with different orthographies cannot be compared. Secondly, the results shed more light on historical written contact than on modern phonological similarity. Thirdly and most importantly, most orthographies are fairly inadequate representations of phonology.

5.2.2 Example phonological characters

Here are some examples of the same phrase represented using different types of characters (with IPA transcription to aid the reader (International Phonetic Association, 1999)):

(5.1) Orthographic characters

The North Wind and the Sun were disputing which was the stronger
ðə ˈnɔθ ˌwɪnd ən ðə ˈsʌn wə dɪsˈpjutɪŋ ˈwɪʧ wəz ðə ˈstɹɒŋgə

(5.2) Phrasal stress characters

0 2 1 0 0 2 0 020 2 0 0 2
ðə ˈnɔθ ˌwɪnd ən ðə ˈsʌn wə dɪsˈpjutɪŋ ˈwɪʧ wəz ðə ˈstɹɒŋgə

(5.3) Voicing features

++ ++− ++++ ++ ++ −++ ++ ++−+−++−++ ++− +++ ++ −−+++++
ðə ˈnɔθ ˌwɪnd ən ðə ˈsʌn wə dɪsˈpjutɪŋ ˈwɪʧ wəz ðə ˈstɹɒŋgə

(5.4) Nasality features

−− +−− −−+− −+ −− −−+ −− −−−−−−+ −−− −−− −− −−−−+−−
ðə ˈnɔθ ˌwɪnd ən ðə ˈsʌn wə dɪsˈpjutɪŋ ˈwɪʧ wəz ðə ˈstɹɒŋgə

(5.5) Feature bundles (voicing and nasality combined)

$\begin{matrix}+\\-\end{matrix}\begin{matrix}+\\-\end{matrix}$	$\begin{matrix}+\\+\\-\end{matrix}\begin{matrix}+\\-\\-\end{matrix}$	$\begin{matrix}+\\-\\-\\+\end{matrix}\begin{matrix}+\\-\\+\\-\end{matrix}$
BB	ABC	BBAB
ðə	ˈnɔθ	ˌwɪnd

Each feature bundle forms a single character, so $\begin{matrix}+\\+\end{matrix}$, $\begin{matrix}+\\-\end{matrix}$ and $\begin{matrix}-\\-\end{matrix}$ could alternatively be represented as 'A', 'B' and 'C'. Using this representational choice, it does not matter how many feature values the bundles have in common, only whether each bundle is identical or not. A phonemic representation is a particular kind of feature bundle: if feature bundles comprise all the relevant features, they are abstract phonemes.

This choice of character type for the algorithm can therefore be used to compare feature theories; by choosing different representations (e.g. SPE features, Elements) we can compare which theory of representation gives a more insightful result.

For my prototype in Section 5.5, I use orthographic and phonemic characters. In Section 5.6, I then move on to using various subphonemic features, described below.

5.2.3 Static IPA-feature mapping

The first phonological representation I examine is that of Hayes, 2008. This is a set of binary features which map statically to the IPA. All segments are fully specified for all relevant features. Whilst this has obvious problems in accounting for natural classes cross-linguistically, it is straightforward to apply to IPA-transcribed texts from multiple languages, and is therefore a useful starting place.

5.2.4 Language-specific binary features

The second representation is a set of binary features formed from the consensus of Gussenhoven and Jacobs (2013), Hayes (2008), and Odden (2005). There is variation both in the inclusion or exclusion of features in a given feature system, and in the criteria used to decide on their values. Where possible, I have relied on the criteria found in the three textbooks, for consistency.

Feature set

Table 5.1 lists all the features found in Gussenhoven and Jacobs (2013), Hayes (2008), and Odden (2005); and the set of features I have chosen to include. The criteria for deciding on the values of these features are in Section A.1 on page 195, and the values for each languages listed in Section A.3 on page 201.

I am not including [syllabic], assuming that structural information is represented separately from melodic information (Goldsmith, 1976). This means that in the representations below, glides are indistinguishable from high vowels, since structural information is not included. Likewise, I am not including [long] or [delayed release]; these are better represented by one-to-many / many-to-one relationships between the melodic and segmental tiers.

Backness and rounding give a four-way contrast; including [front] to generate a six-way contrast is unnecessary, at least for the languages sampled.

Implosives can be specified with a combination of constricted glottis and voicing.

Whilst not strictly necessary for distinguishing segments, [labial] rationalises observable patterns. For this set of languages, including it makes [round] redundant.

Labiodentals can be specified with [distributed] and [strident]. Furthermore, per Odden and other authors, [strident] is redundant for all the languages under examination.

[RADICAL] contrasts pharyngeal with other places of articulation, and is likewise redundant for the languages in my sample.

[tap] and [trill] are specified with [distributed] and [continuant].

I used a Python program to analyse a feature specification for a given language, and indicate where there are redundancies, or where two segments have the same specification.[3] I found that [distributed] and [constricted glottis] are redundant features with this choice of languages, and they are therefore not included in the entropy calculations. More details are available in Section A.2 on page 198.

5.2.5 Element Theory representation

SPE-style binary features are not the only system of phonological representation currently in use. One alternative to using articulatory features is Element Theory. Elements correspond to

[3]The source code is available at https://github.com/ElizabethSEden/NaturalClasses.

Odden	**Gussenhoven & Jacobs**	**Hayes**	Consensus
anterior	anterior	anterior	✓
	approximant	approximant	
back	back	back	✓
consonantal	consonantal	consonantal	✓
constricted glottis	constricted glottis	constr glottis	(✓)
continuant	continuant	continuant	✓
coronal	CORONAL	coronal	✓
delayed release		delayed release	
distributed	distributed	distributed	(✓)
	DORSAL	dorsal	
		front	
high	high	high	✓
		implosive	
labial	LABIAL	labial	✓
		labiodental	
lateral	lateral	lateral	✓
low	low	low	✓
nasal	nasal	nasal	✓
	RADICAL		
round	round	round	(✓)
sonorant	sonorant	sonorous	✓
spread glottis	spread glottis	spread glottis	✓
strident	strident	strident	
		tap	
ATR	tense	tense	✓
		trill	
voice	voice	voice	✓
syllabic		syllable	
long		long	

TABLE 5.1: Features consensus; highlighted features are included; constricted glottis, distributed and round are excluded as redundant.

acoustic signatures, though there is no one-to-one mapping to the phonetic signal; the elements of a language are discovered through its phonological behaviour.

I will derive the elements for the seven languages in question based on the principles in Backley (2011). There are six elements, each of which can be a head or a dependent in Backley's approach. A headed element plays a greater role in determining the overall acoustic shape. Headedness is represented by underlining.

The element assignments that I have chosen are in Section A.4. Element values for English are adapted from the values for Received Pronunciation English in Backley (2011), as are element values for the other languages which Backley discusses explicitly[4]. The remaining element values are only a first approximation, and open to amendment. However, they are sufficient to test a proof of concept - namely that such representations will give rise to language-dependent cross-entropy differences.

The six elements and their characteristics are:

(1) **|A|**

|A| is characterised by a lower-central energy peak, around 1kHz. |A| as a single element in an expression will be a sound like [a].

|A| contributes to place in coronals, labiodentals and gutterals. Simplex |A| is used in retroflexes or pharyngeals.

(2) **|I|**

|I| is characterised by energy peaks around 500Hz and 2.5kHz, with a dip between them. |I| as a single element in an expression will be a sound like [i].

As a consonantal place element, |I| is used in coronals and |I| in palatals. |I A| is used in alveolo-palatals. Non-high front vowels are |I| with |A|.

(3) **|U|**

[4] See in particular p.52 for English vowels, p.109 for place in obstruents, p.161 for manner in consonants, and p.184 for glides.

|U| is characterised by low frequency energy, under 1kH. |U̲| as a single element in an expression will be a sound like [u].

Non-high back vowels are |U| with |A|. Front rounded vowels are |I| with |U|. |U| plays a similar rounding role in consonantal place: |U̲| for labials, |U̲ A| for labiodentals, |U| for velars, |I̲ U| for palato-velars and |U A̲| for uvulars.

A central vowel such as schwa may be empty, containing none of the three vowel elements.

(4) **|H|**

|H| is characterised by high-frequency aperiodic noise, such as frication and release bursts. |H| in isolation is placeless frication noise, i.e. [h]. Dependent |H| indicates a fricative, with |H̲| a fortis or aspirated fricative. In a nasal, |H| indicates breathiness or voicelessness. In a stop, |H̲| indicates aspiration and |H| breathiness or an ejective.

I have followed Backley's simplifying assumption that a language with a two-way laryngeal contrast is either aspirating (an H language) or voicing (an L language), with no variation between stops and fricatives.

(5) **|L|**

|L| is characterised by murmur, a band of low frequency energy found most prevalently in nasals. |L| in isolation is a placeless nasal, such as the moraic nasal of Japanese. |L| is found in nasal consonants and vowels; |L̲| is found in voiced obstruents.

(6) **|ʔ|**

|ʔ| is characterised by "a sudden and sustained drop in acoustic energy". |ʔ| is isolation is a placeless (i.e. glottal) stop, [ʔ]. |ʔ| is found in stops, and some nasals and laterals. |ʔ̲| is found in ejectives.

(7) **Syllabicity**

Unlike with SPE-style features, the same elements are used to represent all vowels and consonants. Glides and liquids, lacking frication or closure, are separated from vowels only by syllable

structure.

However, for the purposes of a linear representation, I have included an additional bit of information for each segment: its syllabicity. An alternative to this approach would be to include empty nuclei where necessary to give rise to an entirely predictable onset-nucleus structure. From a conservation of information perspective, the outcome is comparable, if not identical.

(8) **Length**

Length is expressed structurally, with elements associating to multiple timing slots. For the purposes of entropy calculation, I have expressed this as duplicate element bundles. Backley encodes the Germanic tense-lax distinction solely through length, with long-short pairs having the same element structure, despite the quality difference. I have kept to this principle, since this results in no loss of contrast.

5.3 Algorithms

We have seen that language distance can be measured using entropy estimation, and reviewed some potential phonological representations of language. In this section, I will give an overview of the algorithms that I am using to estimate values for entropy.

Since entropy is a measure of predictability, it can be estimated using the results of compression algorithms. The aim of a compression algorithm is to remove any redundant information from a message, whether that be an audio recording, a text file, or something else.

5.3.1 Unigram model

The most basic algorithm for estimating P calculates entropy directly from the probability distribution P of characters in a text, using Shannon's formula:

$$H(P) = \sum_i (-log_2 p_i) \cdot (p_i)$$

where p_i is the probability of a given character.

It uses a basic unigram model of probability, based simply on the frequency of each character observed in a sample of text. In its simplest form, this model is:

$$p_i = \frac{n_i}{N}$$

where n_i the number of times it is observed in a sample of N characters.

This model gives a probability of 0 for characters which are not found in the text sample, so to account for inevitable low frequency items, a smoothing function is applied:

$$p_i = \frac{n_i + \lambda}{N + A \times \lambda}$$

where A is the number of different potential characters ('alphabet size'). λ is the smoothing parameter. A greater value of λ means that a greater number of previously unseen items are expected. I have set λ to 0.5, a commonly used value in Natural Language Processing (Manning and Schütze, 1999).

I implemented this algorithm to prototype the cross-entropy approach in Section 5.5.

5.3.2 Prediction by partial matching

A more complex model estimates p_i using the surrounding context. Instead of the probability of a character or a word being fixed, it is dependent on the preceding n characters or words (n-gram models) or words and their parts of speech (n-pos models).

A Markov model lists the possible states (e.g. 't', 'h', 'g'), and the probability of transitioning from one to another (e.g. 't' $\rightarrow$ 'h' = 0.5; 't' $\rightarrow$ 'g' = 0.01; 't' $\rightarrow$ 't' again = 0.1.). The model is memoryless - only the current state matters, and the probabilities do not depend on previous states. To take larger contexts into account, each longer string must be treated as an independent state (e.g. 'th', 'he', 'gh').

In prediction by partial matching (PPM), several of these fixed-order context models are combined, with the process starting with the longest matching model, and falling back to shorter contexts if no match can be found.

Teahan (2000) finds that PPM can be used to successfully identify the dialect of orthographic text as British or American English. Teahan's (1999) Text Mining Toolkit which implements this scheme is therefore a reasonable starting point for examining phonological language identification, and the source of entropy calculations in Section 5.6.

The longest useful context with orthographic characters has been found to be 5 characters (Cleary and Teahan, 1997). Beyond this length, predictions that do exist are more specific, but many contexts do not give rise do any predictions at all; this uncertainty increases entropy. I have therefore used the Text Mining Toolkit's default maximum context of 5 characters in the investigation below. Further research is needed to determine if other representations have the same optimal context length as orthographic characters.

5.3.3 Alternative algorithms

There are several text compression schemes besides PPM, including the match-length approach used by Juola (see Subsection 3.3.2 on page 37). However, comparing their performance is beyond the scope of this investigation.

5.4 Methodology

To recap, entropy is a measure of predictability of a sequence of characters. The cross-entropy of two sequences is how good a measure the entropy as calculated from one sequence is at predicting the other sequence. The maximum possible entropy of a sequence is constant for a single set of characters; to compare between entropies derived from different sets of characters, we can divide by this maximum value. The true entropy of each sequence, however, is not constant. It must be subtracted from the calculated cross-entropy, so that the final value is directly comparable across different pairs of sequences. This final value is called the Kullback-Leibler divergence.

The Kullback-Leibler divergence tells us how predictable a sequence A is, given a sequence B. By definition, the more similar the two sequences are, the smaller the Kullback-Leibler divergence will be. This method can be applied to any pairs of sequence of characters.

5.4.1 Hypotheses

For each system of representation, I test the following hypotheses:

1. The language of a test string can be reliably identified.

2. The minimum required test string length for reliable language identification is consistent across multiple samples of text.

3. If the language of the test string can be reliably identified, the Kullback-Leibler divergences between each pair of languages will be consistently ranked across multiple samples of text.

4. The Kullback-Leibler divergence is symmetrical for all language pairs.

5. Languages do not differ in their segmental predictability

6. Every feature encodes the same amount of information

Hypotheses 1 - 3 are requirements for Kullback-Leibler divergence to be a viable method of measuring language distance, with Hypothesis 1 a prerequisite for Hypothesis 2, and 2 for 3.

Hypotheses 4 - 6 examine the relative information content of components of a text in a given representational system. The default position is that each component is homogenous with regards to predictability. Language acquisition and intelligibility can be asymmetric between language pairs, suggesting Hypothesis 4 may be false. Languages vary in their use of suprasegmental information, implying Hypothesis 5 to be false. Representational theories differ in which features they privilege, so variation in this area provides a means of comparing theories based on observable information content.

5.4.2 Language distance

For each system of representation, the algorithm to generate a language distance metric is as follows:

1. For each language, obtain multiple samples of text transcribed in the desired character set.

2. Calculate the cross-entropy $H(P, Q)$ of each pair of samples.

   (a) Calculate the probability distribution P for each sample of text.

   (b) Calculate the cross-entropy $H(P, Q) = \sum_i (-log_2 q_i) \cdot (p_i)$ for each pair of samples. Where a sample is paired with itself, this gives the entropy of that sample: $H(P, Q) = H(P, P) = H(P)$.

(c) Verify that this gives consistent results

3. Calculate the cross-entropy for each pair of languages by grouping sample pairs by their languages, and taking the average of the group.

4. Normalise the cross-entropies by dividing by the maximum possible entropy.

5. Calculate the Kullback-Leibler divergence $KL(P, Q)$ of each language pair.

6. Take the average of $KL(P, Q)$ and $KL(Q, P)$ to get a symmetrical language distance.

5.5 Prototype

5.5.1 Input data

For the prototype, I have used orthographically transcribed data from the Europarl corpus (Koehn, 2005), as a starting point. This corpus contains text from 20 European languages, taken from the proceedings of the European Parliament, of which I am examining six: three Germanic (Dutch, English, German) and three Romance (French, Portuguese, Spanish). I have used eight samples of 1000 lines per language, sampled at random from the proceedings in that language, putting aside the question of minimum required sample size during this experiment.

5.5.2 Replication of orthographic work

Firstly, I present a partial reproduction of Juola's orthographic results, showing that the unigram-based algorithm works as intended.

Table 5.2 shows the cross-entropy $H(P, Q)$ of each language pair. This has been averaged across all samples of each language, and normalised. Since I used a unigram probability distribution of 26 segments, the maximum entropy was 4.7 ($= -log_2 \frac{1}{26}$),

The row name refers to the source language of 'true' probability (P), i.e. a probability distribution generated from the text itself. The smallest value in a row is in the column of the model which best predicts it. Column names refer to the source language of 'estimated' probability (Q) i.e. probability distributions generated from other texts. The smallest value in a column is the sample of text which the model best predicts.

P \ Q	Portuguese	French	Spanish	German	English	Dutch
Portuguese	**0.71**	0.73	0.74	0.81	0.83	0.85
French	0.72	**0.71**	0.74	0.79	0.81	0.84
Spanish	0.71	0.72	**0.72**	0.79	0.8	0.83
German	0.77	0.76	0.78	**0.73**	0.79	0.78
English	0.73	0.73	0.75	0.75	**0.75**	0.8
Dutch	0.75	0.75	0.77	0.74	0.78	**0.75**
Average $H(P,Q)$	0.73	0.73	0.75	0.77	0.79	0.81

TABLE 5.2: Cross-entropy $H(P,Q)$ of orthographic texts

We can see that for every column, the smallest value is that where the source and model language are the same, where $H(P,Q) = H(P)$. This value has been highlighted for each model language. It is not a constant across languages; languages vary in their predictability, which is why the Kullback-Leibler divergence is required.

For example, the Portuguese and French models have inherently lower entropy than the English and Dutch. The cross-entropy of those models with all source languages therefore tends to be lower, and in some cases even lower than when the source and model languages match.

Looking at the Kullback-Leibler divergence in Table 5.3, we can see that it is minimised when source and model language match, whether in comparison to alternative source languages (rows) or model languages (columns).

The resulting 'distances' are visualised using Phylip (Felsenstein, 1989) in Figure 5.1.

P \ Q	Portuguese	Spanish	French	English	Dutch	German
Portuguese	**0.00**	0.02	0.02	0.08	0.08	0.10
French	0.01	**0.00**	0.02	0.06	0.06	0.09
Spanish	0.00	0.01	**0.00**	0.06	0.05	0.08
German	0.06	0.05	0.06	**0.00**	0.04	0.03
English	0.02	0.02	0.03	0.02	**0.00**	0.05
Dutch	0.04	0.04	0.05	0.01	0.03	**0.00**

TABLE 5.3: Kullback-Leibler divergence of orthographic texts

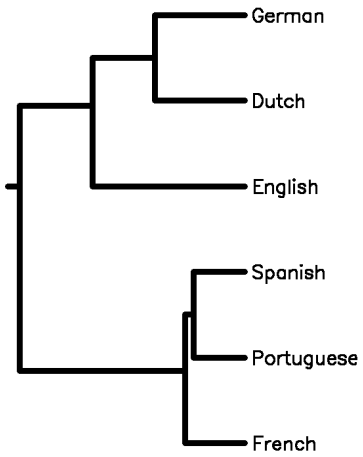


FIGURE 5.1: Language distance based on average Kullback-Leibler divergence of orthographic texts

These results replicate Juola's findings, that the orthographic distances as measured by cross-entropy are a good proxy for historical relatedness of Indo-European languages: the Iberian languages are grouped together, then Romance; and Germanic separately.

5.5.3 Transcribed results

Next, I present the results of the unigram method as applied to IPA-transcribed text from Dutch, English and French. I include the intermediate steps of the probability values and cross-entropy, and the resulting language distance.

To get samples approximating phonemically transcribed data, I automatically replaced the orthographic Europarl text with IPA transcriptions of each individual word, drawn from the lexicons used in Nidaba (CELEX for English and Dutch, and Lexique3 for French). For these languages, more than 85% of instances of orthographic words could be replaced with IPA transcriptions. The remaining words - mostly proper nouns - were not included. This is obviously a very crude technique, but it gives some indication of the feasibility of using IPA-based texts.

Probability distributions

Table 5.4, Table 5.5 and Table 5.6 show samples of probability distributions for each IPA segment. The different text samples for English produce slightly different probability distributions, but they are much more similar to each other than to the French distribution. Segments which have a probability of less than 0.001 have not been shown; the differing inventories obviously produce the largest disparity. Using distinctive features rather than phonemes will eliminate this effect (see Section 5.6).

5.5.4 Average cross-entropy per language pair

Table 5.7 shows the average cross-entropy of each language pair. The normalising constant was 8.54.

The minimum cross-entropy for each language occurred when the true language was used to generate the estimate, as expected. The results are approximately symmetrical for each pair, i.e. there is little or no difference between Dutch being the source of the 'true' model, and English being the source of the estimate, and vice versa. The cross-entropy between Dutch and English is much smaller than the cross-entropy of either with French.

5.5.5 Kullback-Leibler divergence

Table 5.8 shows the Kullback-Leibler divergence for Dutch, French and English texts. It is visualised using Phylip (Felsenstein, 1989) in Figure 5.2. I find that English and Dutch are most similar, as expected, followed by French and English, then French and Dutch.

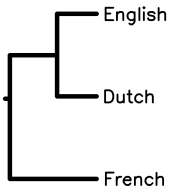


FIGURE 5.2: Language distance based on average Kullback-Leibler divergence of IPA transcribed texts

Segment	Probability
ɪ*	0.093
t	0.054
n	0.053
ə	0.048
s	0.034
l	0.027
iː	0.026
ð	0.026
d	0.026
k	0.024
æ	0.023
ɹ	0.022
z	0.021
m	0.020
ɒ	0.018
p	0.018
ɛ	0.017
v	0.016
e	0.015
ʊ	0.014
uː	0.014
w	0.014
a	0.013
b	0.012
f	0.011
ɔː	0.011
ʌ	0.009
ʃ	0.008
ŋ	0.006
h	0.006
j	0.006
g	0.005
ɑː	0.005
t͡ʃ	0.004
d͡ʒ	0.004
ɜː	0.004
θ	0.002
ɔ	0.001
ʒ	0.001

Table 5.4: Unigram probabilities for English sample 1

Segment	Probability
ɪ*	0.096
n	0.055
t	0.052
ə	0.048
s	0.035
l	0.028
d	0.027
iː	0.026
ð	0.025
æ	0.024
k	0.023
ɹ	0.022
z	0.021
p	0.021
m	0.020
ɛ	0.018
ɒ	0.018
v	0.016
ʊ	0.015
e	0.014
uː	0.014
w	0.014
a	0.013
ɔː	0.012
b	0.011
f	0.010
ʃ	0.009
ʌ	0.008
ŋ	0.006
h	0.006
j	0.005
g	0.005
ɑː	0.005
t͡ʃ	0.004
ɜː	0.003
d͡ʒ	0.003
θ	0.003
ɔ	0.001
ʒ	0.001

Table 5.5: Unigram probabilities for English sample 2

Segment	Probability
ʁ	0.064
e	0.051
a	0.051
s	0.045
l	0.044
ø	0.042
d	0.040
i	0.038
t	0.038
p	0.032
k	0.032
ɛ	0.031
o	0.025
ɑ̃	0.024
m	0.021
õ	0.019
n	0.019
y	0.019
j	0.017
v	0.013
u	0.011
f	0.009
z	0.008
ɔ	0.008
ɛ̃	0.007
b	0.006
ʒ	0.006
g	0.005
w	0.005
ʃ	0.004
œ	0.004
ɥ	0.003
œ̃	0.002
ɲ	0.001

Table 5.6: Unigram probabilities for French sample 1

*Note that the happY vowel is transcribed as [ɪ] in CELEX

P \ Q	Dutch	English	French
Dutch	**0.31**	0.46	0.64
English	0.46	**0.31**	0.60
French	0.66	0.66	**0.31**
Average	0.48	0.52	0.48

TABLE 5.7: Cross-entropy $H(P, Q)$ of IPA transcribed texts

Dutch	0		
French	0.34	0	
English	0.15	0.32	0
	Dutch	French	English

TABLE 5.8: Average Kullback-Leibler divergence of IPA transcribed texts

5.5.6 Conclusion

Applying the basic unigram calculation of entropy to orthographic data reproduces Juola's results, so I am confident that this algorithm functions as intended.

Applying it to even simplistically auto-transcribed samples gives internally consistent results, which accord with both historical and intuitive measures of distance. In Section 5.6, I therefore use Teahan's Text Mining Toolkit to gain a more sophisticated and accurate measure of the cross-entropy of a variety of phonological representations.

5.6 Text Mining Toolkit

5.6.1 Input data

The entropy of orthographic texts is affected by factors including dialect, genre, author and topic (Teahan, 2000) – in short, everything that alters the content of a text. I therefore used translations of a single text, to minimise the impact of these factors. An investigation using non-translated (and hence possibly more representative) texts will require many more texts from across a wide range of genres and authors.

For training texts, I used the first chapter of the gospel of Mark. This is a text which is widely translated. In many languages, it may be the only published material, or the only text available

in both English and a minority language. It is also a text which tends to be available as an audio recording as well as – or even in preference to – an orthographic text.

Phonemic transcriptions were created by performing substitutions on an orthographic text, using data from the lexicons in Nidaba. The resulting transcription was then verified against an audio recording where possible.

For test data, I initially used The North Wind and the Sun, a widely translated story used for example transcriptions by the International Phonetic Association. However, to examine the effects of varying the length of the test string, I instead used a longer text - the second chapter of the letter to the Phillipians - from the same Bible translation for each language. Some of the cross-entropy effects may therefore reflect the fact that training and test texts for a given language share translators.

These texts can be found in Section A.5.

5.6.2 Results: IPA Representation

The first representation I examine is IPA transcription. Each character is a single phoneme.

Language identification

Using samples of text from Phillipians 2, the correct language for each test string was identified reliably (i.e. in 100% of cases) for test strings of length 26 characters or longer. There is an exponential increase in mis-identification as test strings become shorter than this threshold, with Spanish being identified as Greek, then also German as Dutch, then a broader scattering of errors (see Figure 5.3).

The best fit curve has the equation: Percentage correct $\approx 100(1 - 1.3e^{-0.41L})$, where L is the length of the test string. Therefore, to achieve 100% accuracy using 100 test strings in 99% of experiments, a test string of length $L \geq 34$ is required. A test string of length 500 (used hereafter) has an identification error rate of <1 in 10^{-87}. I therefore confirm that the language of the test string can be reliably identified, and the length threshold for doing so is consistent, as per Hypotheses 1 and 2 in Subsection 5.4.1.

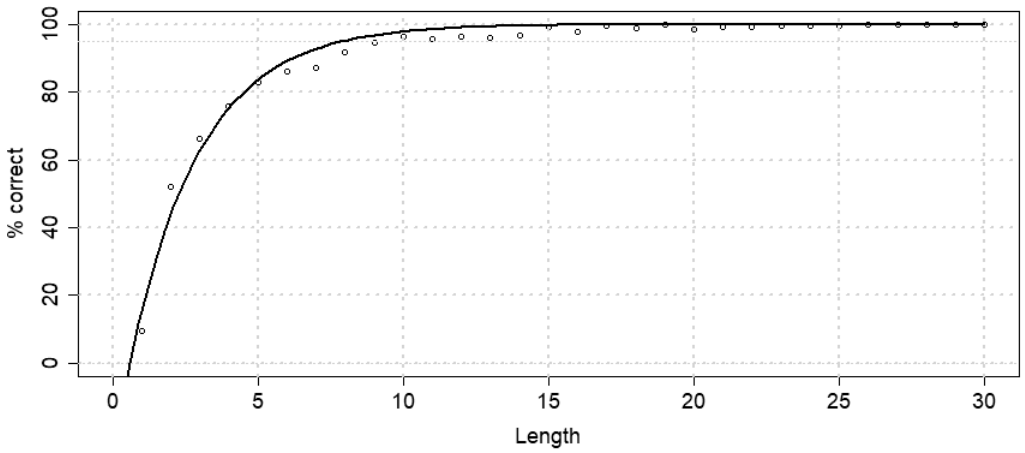


FIGURE 5.3: Percentage of test strings correctly identified by length

Language interaction as a predictor of cross-entropy

We can reliably identify the language of a test string of a given length transcribed phonemically in the IPA. The cross-entropy of a test string in a given language with a model based on that same language is therefore consistently ranked lower than the cross-entropy of different language models. But are the mean cross-entropies of non-identical language pairs distinguishable from one another?

Applying a one-way ANOVA to the cross-entropy of an ordered pairing[5] of languages for test strings of length 500±5 characters, I find that there is an effect size of $\eta^2 = 0.87$ (see Table 5.9). That is, the proportion of the variance in cross-entropy that can be explained by the combination of the test language and the model language is 87%. The proportion of the variance which is residual, not explained by this, nor by the test language or model language independently, is <0.5%. Ordered pairings of languages are a reliable predictor of cross-entropy, and so further investigation of Kullback-Leibler divergence is worth pursuing.

Figure 5.4 shows the distributions of cross-entropy for test strings of length 500±5 characters and 150±5 characters. There were approximately six test strings and 21 test strings per language, respectively.

[5]e.g. 'Dutch Spanish' refers to a Dutch test string modelled using Spanish, which as discussed is not necessarily the same as 'Spanish Dutch', a Spanish test string modelled using Dutch.

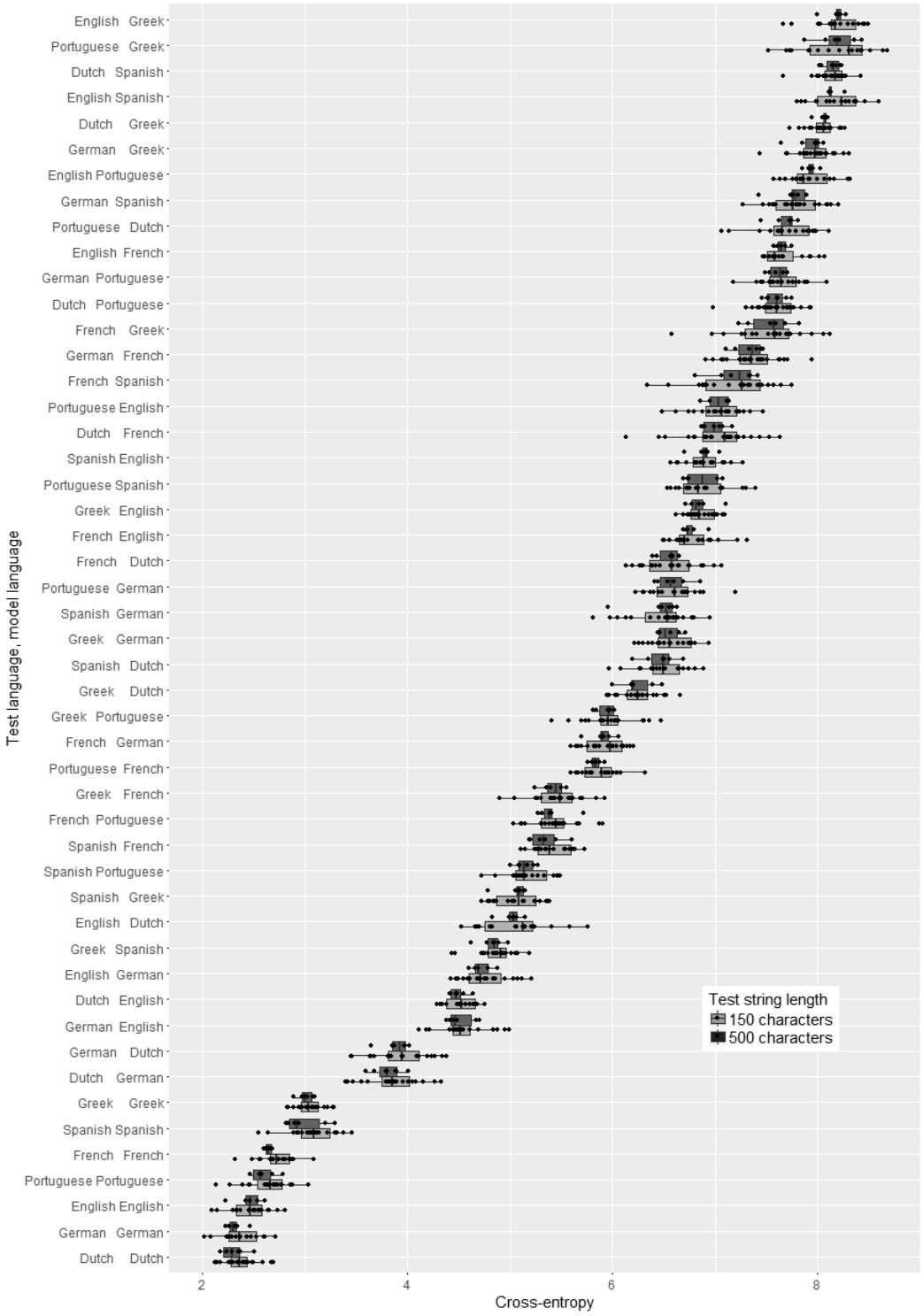


FIGURE 5.4: Cross-entropy ranking of IPA transcriptions, for test strings of length 150 and 500 characters.
Each point corresponds to a single test string. Also shown are the mean, hinges at first and third quartiles, and whiskers extending to the minimum/maximum values that are no further than 1.5 times the inter-quartile range from the hinges.

	Deg. of freedom	Sum of Squares	Mean Square	F-ratio	Pr(>F)	η^2
Language of test string	6	0.398	0.06639	267.6	$< 2 \times 10^{-16}$	0.031
Language of model	6	1.283	0.21384	862	$< 2 \times 10^{-16}$	0.099
Language of test string × Language of model	36	11.278	0.31328	1262.9	$< 2 \times 10^{-16}$	0.866
Residuals	245	0.061	0.00025			

TABLE 5.9: Factors contributing to variance in cross-entropy of IPA transcriptions, for test strings of length 500 characters.

Consistency of Kullback-Leibler divergence

Having established that cross-entropy is significantly predicated on the combination of two languages, we turn to the Kullback-Leibler divergence.

Figure 5.5 shows symmetric Kullback-Leibler divergences for all language pairs. These are calculated by pairwise means of the Kullback-Leibler divergences for a test string of language A modelled with B, and for B modelled with A, and normalised using the same constant as in the prototype (i.e. 8.54) to give values between 0 and 1.

The robustness of this ranking was tested using 10-fold cross-validation. The data were randomly divided into 10 sets. Each set in turn was treated as a test set, with the remaining 90% of data points forming a training set. The training sets were modelled using a random decision forest, and the resulting predictions compared to the relevant test set. The mean error was 0.016, the 99th percentile was 0.053, and the maximum was 0.085. For comparison, the values obtained for these languages have ranges between 0.16 and 0.63, so 99th percentile Kullback-Leibler divergences obtained from an IPA representation are accurate to ±11% of the range. For the purposes of categorical comparison, these language pairs could therefore be divided into five non-overlapping categories (see Table 5.10).

Considering Hypothesis 3 (Subsection 5.4.1), that the Kullback-Leibler divergences are consistently ranked, we see that this is false when considering the ordering of 42 language pairings as distinct items. However, we can reject the null hypothesis that there is no effect on rankings

from language pairings, since five distinct categories can be observed.

Similar	Dutch & German
	Greek & Spanish
	English & German
Somewhat similar	Dutch & English
	French & Portuguese
Middling	Portuguese & Spanish
	French & Spanish
	French & Greek
Somewhat dissimilar	French & German
	Greek & Portuguese
	Dutch & French
	German & Spanish
	Dutch & Greek
	German & Greek
	German & Portuguese
	Dutch & Spanish
	English & French
	English & Greek
	English & Spanish
Dissimilar	English & Portuguese
	Dutch & Portuguese

TABLE 5.10: Language pairs categorised by symmetric Kullback-Leibler divergence

Asymmetry of Kullback-Leibler divergence

The Kullback-Leibler divergence of IPA representations is not symmetrical (see Figure 5.7). The cross-entropy of test language A modelled by language B is significantly different from B modelled by A in all cases. However, this asymmetry varies in magnitude (Table 5.11), depending on the language of the test string and of the model.

We can therefore reject Hypothesis 4 (Subsection 5.4.1), that the Kullback-Leibler divergence is symmetrical for all language pairs.

Predictability per language

Returning to the ANOVA of cross-entropy (Table 5.9), we see that the language of the test string, the language of the model and their combination are all significant factors ($p < 10^{-16}$). I therefore reject the null hypothesis that all languages are equally segmentally predictable when

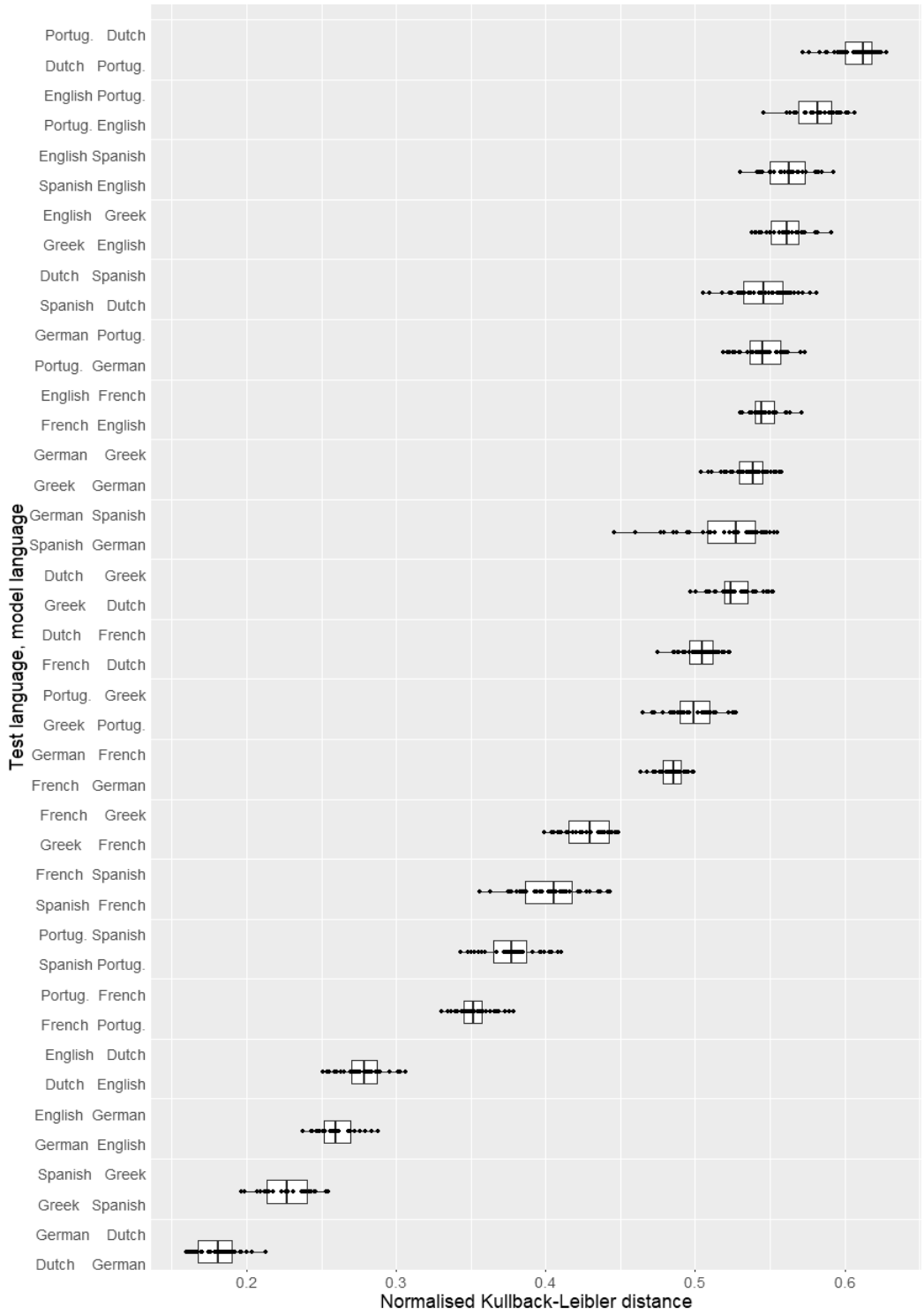


FIGURE 5.5: Symmetric Kullback-Leibler divergence of IPA representation

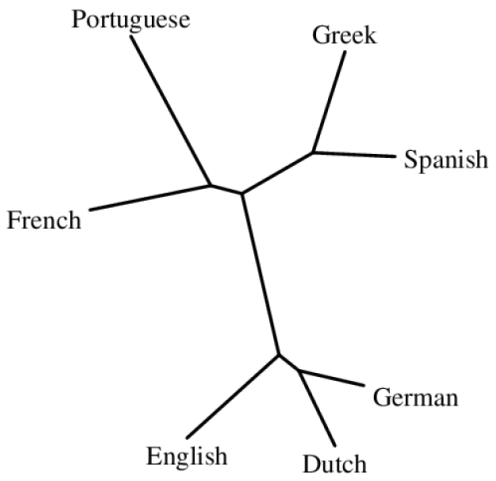


FIGURE 5.6: Visualisation of mean symmetric Kullback-Leibler divergence, IPA transcription. (Dereeper et al., 2008, Felsenstein, 1989)

represented with IPA characters. Of the three factors, the language of the test string has the smallest impact ($\eta^2 = 0.03$), the language of the model has a larger impact ($\eta^2 = 0.10$), and the combination of the two has by far the largest effect size ($\eta^2 = 0.87$).

Test strings in Spanish have the lowest entropy (see Table 5.12). For example, the average Portuguese test string of a given length requires 17% more bits than the average Spanish test string of the same length. This implies that there is more segmental information in a Portuguese phrase than in a Spanish phrase with the same number of segments, and so on for other pairs.

The models for German and Dutch result in better compression, on average, than the models for Spanish and Greek (see Table 5.13). Test strings encoded with a Greek model require $\frac{1}{3}$ more bits, averaged across all test languages, than the same test strings encoded with a German model.

The predictability per language across all four representations under examination is compared in Subsection 5.6.6 on page 153.

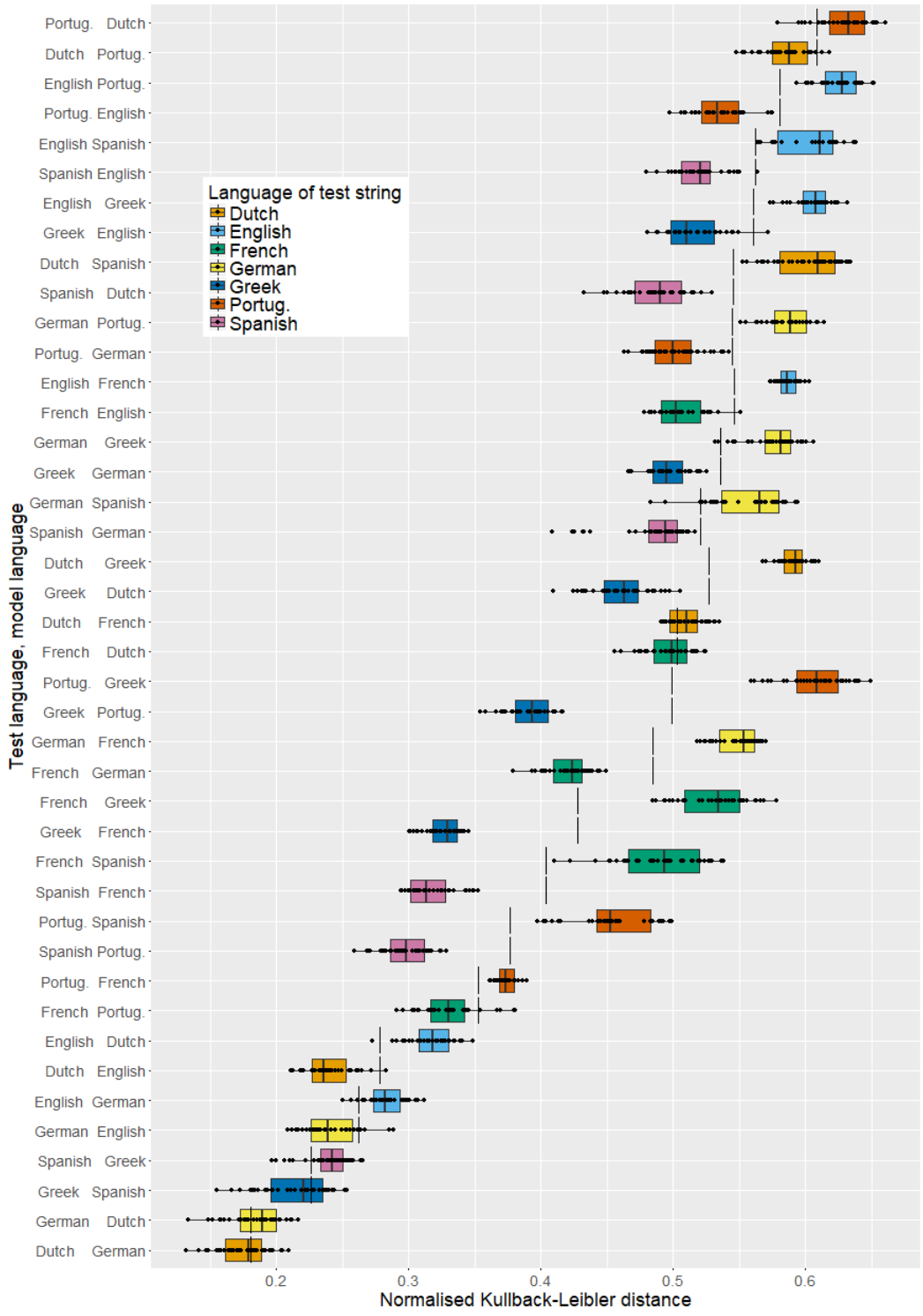


FIGURE 5.7: Kullback-Leibler divergence of language pairs and their inverse, ordered by the mean of the two, which is marked with a vertical line; IPA representation

Language pair	Inverse	Probability (order of magnitude)
Greek Portug.	Portug. Greek	-50
French German	German French	-45
Greek French	French Greek	-42
Greek Dutch	Dutch Greek	-39
Spanish Dutch	Dutch Spanish	-37
Spanish Portug.	Portug. Spanish	-35
Spanish French	French Spanish	-33
Portug. German	German Portug.	-30
Greek German	German Greek	-31
Dutch English	English Dutch	-28
Portug. English	English Portug.	-26
Greek English	English Greek	-24
French English	English French	-24
Spanish English	English Spanish	-21
Dutch Portug.	Portug. Dutch	-18
Spanish German	German Spanish	-16
French Portug.	Portug. French	-14
German English	English German	-11
Greek Spanish	Spanish Greek	-6
French Dutch	Dutch French	-4
Dutch German	German Dutch	-3

TABLE 5.11: Probability that KL distances of language pairs and of their inverses were drawn from the same distribution

	Entropy	% increase over Spanish
Spanish	0.64	0
Greek	0.65	1
German	0.69	8
Dutch	0.69	8
French	0.70	10
English	0.74	15
Portug.	0.75	17

TABLE 5.12: GLM: Contribution to cross-entropy by language of test string; IPA representation

	Entropy	% increase over German
German	0.60	-
Dutch	0.63	4
English	0.66	9
French	0.69	14
Portug.	0.71	17
Spanish	0.77	27
Greek	0.80	32

TABLE 5.13: GLM: Contribution to cross-entropy by language of model; IPA representation

5.6.3 Results: Static SPE-style features

In this next section, I repeat the procedure above using binary features from Hayes, 2008.

Language identification

All test strings of length 25 characters and above are identified as the correct language out of the seven. In the following tests, I use a test string of length 500 characters.

Consistency of Kullback-Leibler divergence

With 28 features, each of which can have one of three values, +, -, or undefined, the theoretical maximum entropy per character is 44.38 bits. Certain feature combinations are illegal, but this is not inherent in the representation. This representation is therefore given a high normalisation factor, and has a much smaller variation in language distance than the IPA representation. (See Figure 5.8 and Figure 5.9).

Applying 10-fold cross-validation, the mean error is 0.0033, the 99th percentile is 0.010, and the maximum error is 0.017. Normalised Kullback-Leibler divergences range from 0.036 to 0.12, so the 99th percentile error is ±12% of the range of values observed, much like for the IPA representation.

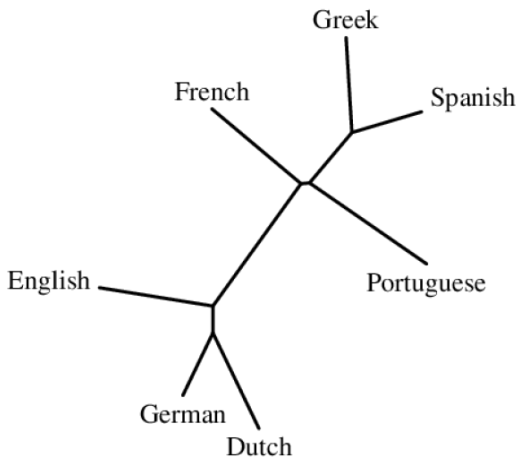


FIGURE 5.8: Visualisation of mean symmetric Kullback-Leibler divergence, Hayes' static featural representation (Dereeper et al., 2008, Felsenstein, 1989)

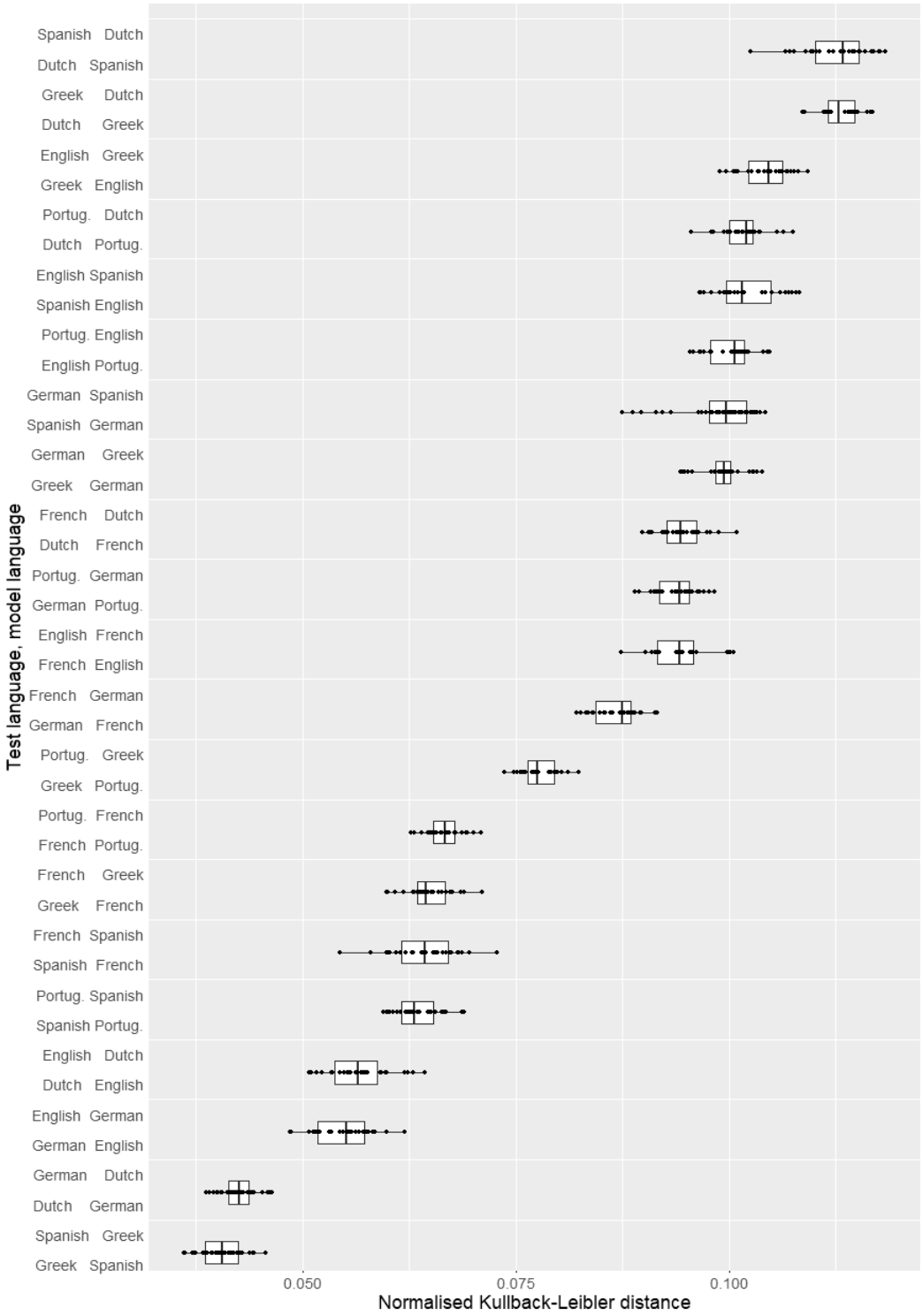


FIGURE 5.9: Symmetric Kullback-Leibler divergence of Hayes' static featural representation

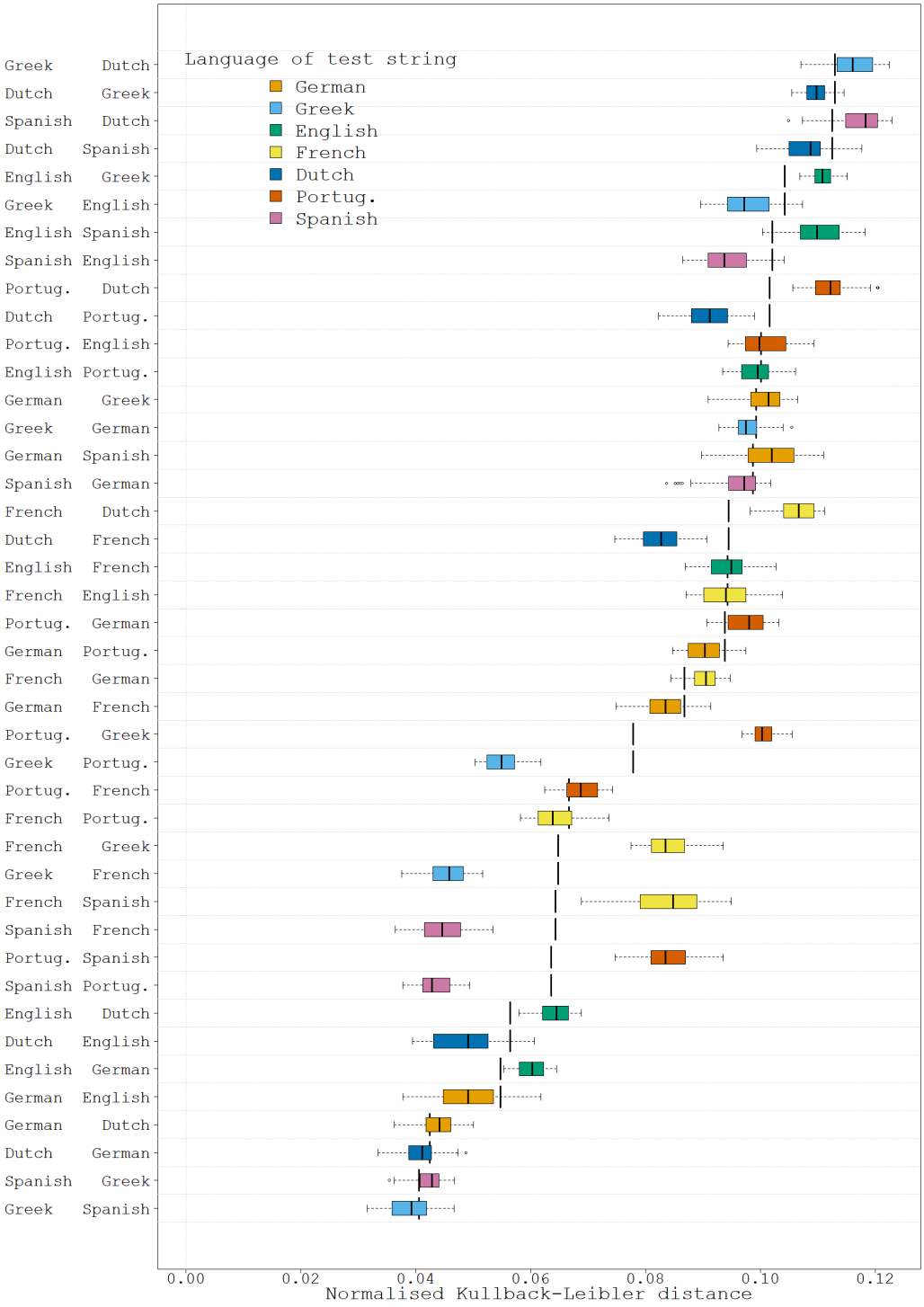


FIGURE 5.10: Kullback-Leibler divergence of language pairs and their inverse, ordered by the mean of the two;
Hayes' static featural representation

Asymmetry of Kullback-Leibler divergence

The Kullback-Leibler divergence of Hayes' static featural representations is not symmetrical (see Figure 5.10). The cross-entropy of test language A modelled by language B is significantly different from B modelled by A, except for the language pairs English and Portuguese, and English and French. As with the IPA, this asymmetry varies in magnitude (Table 5.14), depending on the language of the test string and of the model.

Language pair	Inverse	Probability (order of magnitude)
Greek Portug.	Portug. Greek	-51
Greek French	French Greek	-42
Spanish Portug.	Portug. Spanish	-39
Dutch French	French Dutch	-31
Spanish French	French Spanish	-31
Dutch Portug.	Portug. Dutch	-26
Spanish English	English Spanish	-19
Greek English	English Greek	-16
Dutch English	English Dutch	-16
Dutch Spanish	Spanish Dutch	-12
German English	English German	-11
German Portug.	Portug. German	-11
Dutch Greek	Greek Dutch	-11
German French	French German	-10
Spanish German	German Spanish	-6
French Portug.	Portug. French	-5
Greek Spanish	Spanish Greek	-5
Dutch German	German Dutch	-4
Greek German	German Greek	-3
English Portug.	Portug. English	-1
French English	English French	-1

TABLE 5.14: Probability that KL distances of language pairs and of their inverses were drawn from the same distribution; Hayes' static featural representation

Predictability per language

Applying a one-way ANOVA to the cross-entropy, the language of the test string has a small but statistically significant impact ($\eta^2 = 0.029$), as does the language of the model ($\eta^2 = 0.014$), although the combination of the two has by far the largest effect size ($\eta^2 = 0.95$).

The variation is smaller than using the IPA representation, and not identically patterned (see Subsection 5.6.6). However, there are similarities: the German model results in the best

compression; Spanish test strings require the fewest bits to encode, and Portuguese test strings require the most. (See Table 5.20 and Table 5.21.)

	Entropy	% increase over Spanish
Spanish	0.13	-
Greek	0.14	2
German	0.14	3
Dutch	0.14	5
French	0.15	9
English	0.15	11
Portuguese	0.15	13

TABLE 5.15: GLM: Contribution to cross-entropy by language of test string; Hayes' static featural representation

	Entropy	% increase over German
German	0.14	-
French	0.14	1
Spanish	0.14	6
English	0.14	6
Portuguese	0.15	7
Dutch	0.15	8
Greek	0.15	8

TABLE 5.16: GLM: Contribution to cross-entropy by language of model; Hayes' static featural representation

5.6.4 Results: Language specific SPE-style binary features

In this section, I repeat the procedure above using the consensus of binary features detailed in Section 5.2. The values for each phoneme of each language are listed in Section A.3 on page 201, as determined by the criteria listed in Section A.1 on page 195.

Language identification

The language of all test strings of length 19 characters and above are identified correctly. I use a test string of length 500 characters.

Consistency of Kullback-Leibler divergence

With 16 features, each of which can have one of three values, +, -, or undefined, the theoretical maximum entropy per character is 25 bits. Again, I have not removed illogical combinations of features when calculating this value.

Applying 10-fold cross-validation, the mean error is 0.0052, the 99th percentile is 0.016, and the maximum error is 0.023. Normalised Kullback-Leibler divergences range from 0.13 to 0.29, so the error is ±9% of the range of values observed. This representation has a similar consistency to the previous representations, with a slight improvement compared to the static binary feature representation. Symmetrical Kullback-Leibler divergence is visualised in Figure 5.11 and Figure 5.12. The resulting six categories of language pairs are listed in Table 5.17.

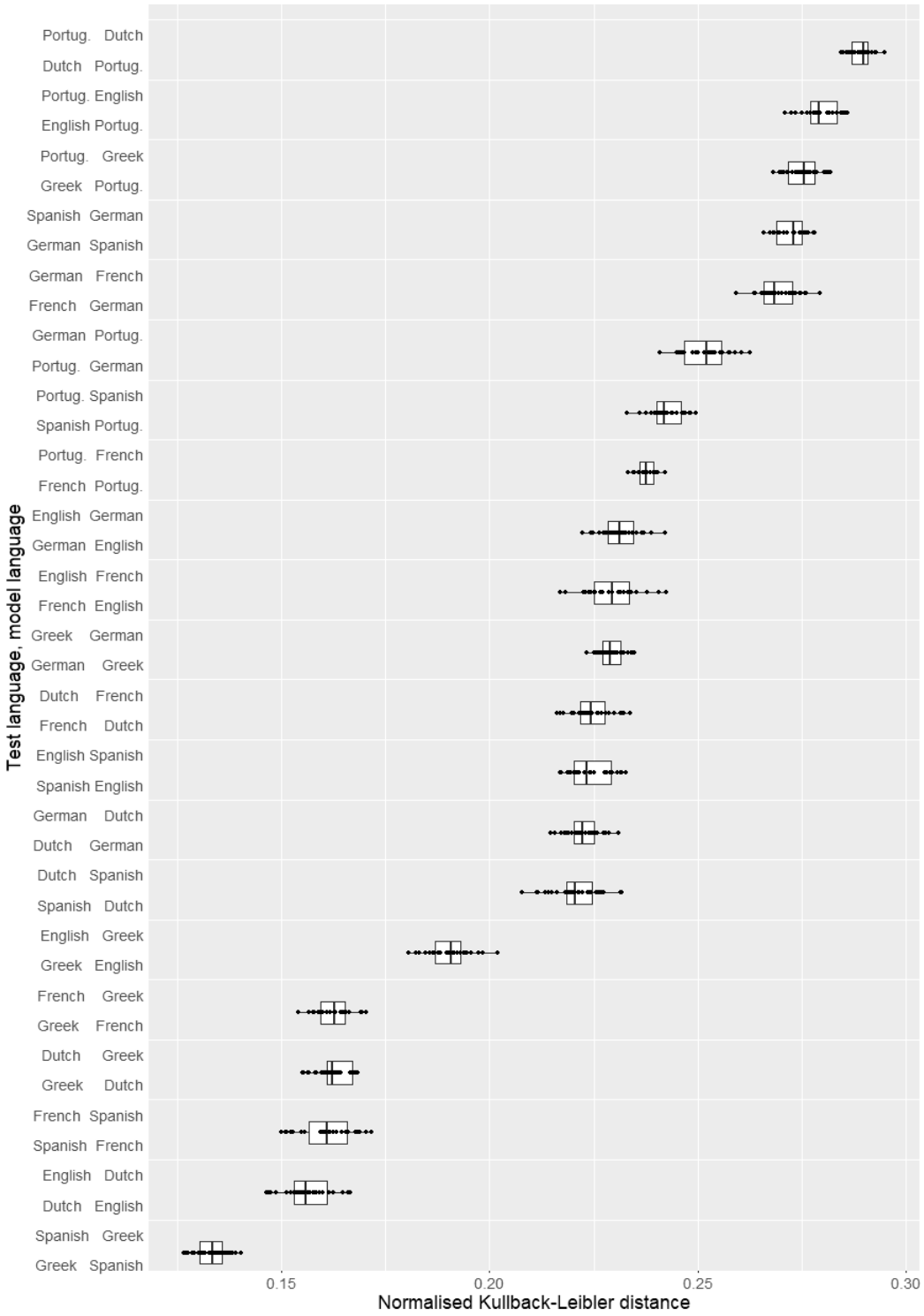


FIGURE 5.11: Symmetric Kullback-Leibler divergence; SPE-style representation

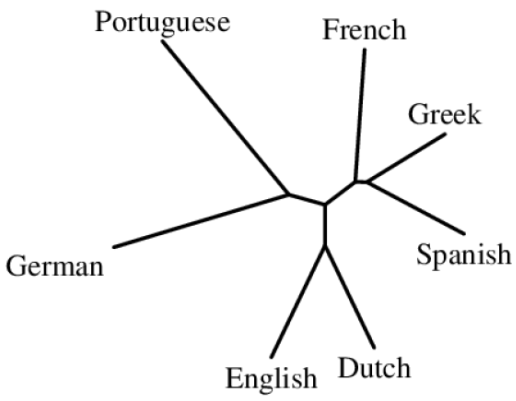


FIGURE 5.12: Visualisation of mean symmetric Kullback-Leibler divergence, SPE-style representation (Dereeper et al., 2008, Felsenstein, 1989)

Most similar	Greek & Spanish
Somewhat similar	Dutch & English
	French & Spanish
	Dutch & Greek
	French & Greek
Middling	English & Greek
Somewhat dissimilar	Dutch & Spanish
	Dutch & German
	English & Spanish
	Dutch & French
	German & Greek
	English & French
	English & German
	French & Portuguese
	Portuguese & Spanish
	German & Portuguese
Dissimilar	French & German
	German & Spanish
	Greek & Portuguese
	French & Portuguese
	English & Portuguese
Most dissimilar	Dutch & Portuguese

TABLE 5.17: Language pairs categorised by symmetric Kullback-Leibler divergence; SPE-style representation

Asymmetry of Kullback-Leibler divergence

The Kullback-Leibler divergence of language-specific featural representations is not symmetrical (see Figure 5.13). The cross-entropy of test language A modelled by language B is significantly different from B modelled by A, except for the language pairs Greek and Portuguese, and Greek and Spanish. As with the previous two representations, this asymmetry varies in magnitude (Table 5.18 and Table 5.19).

Language pair	Inverse	Probability (order of magnitude)
Greek Dutch	Dutch Greek	-40
Spanish English	English Spanish	-39
Spanish Dutch	Dutch Spanish	-34
Greek French	French Greek	-32
Spanish French	French Spanish	-30
Greek English	English Greek	-28
Dutch Portug.	Portug. Dutch	-21
French Dutch	Dutch French	-16
French English	English French	-13
German English	English German	-13
German Greek	Greek German	-10
French Portug.	Portug. French	-9
Dutch German	German Dutch	-8
Spanish Portug.	Portug. Spanish	-8
German Spanish	Spanish German	-5
English Portug.	Portug. English	-5
French German	German French	-4
Dutch English	English Dutch	-3
Portug. German	German Portug.	-3
Greek Portug.	Portug. Greek	-1

TABLE 5.18: Probability that KL distances of language pairs and of their inverses were drawn from the same distribution; language-specific SPE features

Model \ Test	Dutch	English	French	German	Greek	Portug.	Spanish
Dutch		4%	-8%	3%	-29%	7%	-20%
English	-4%		-9%	-8%	-23%	3%	-25%
French	8%	9%		2%	-25%	5%	-31%
German	-3%	8%	-2%		4%	-2%	2%
Greek	29%	23%	25%	-4%		0%	1%
Portuguese	-7%	-3%	-5%	2%	-0%		-3%
Spanish	20%	25%	31%	-2%%	-1%	3%	

TABLE 5.19: Proportional asymmetry in mean Kullback-Leibler divergences; language-specific SPE features

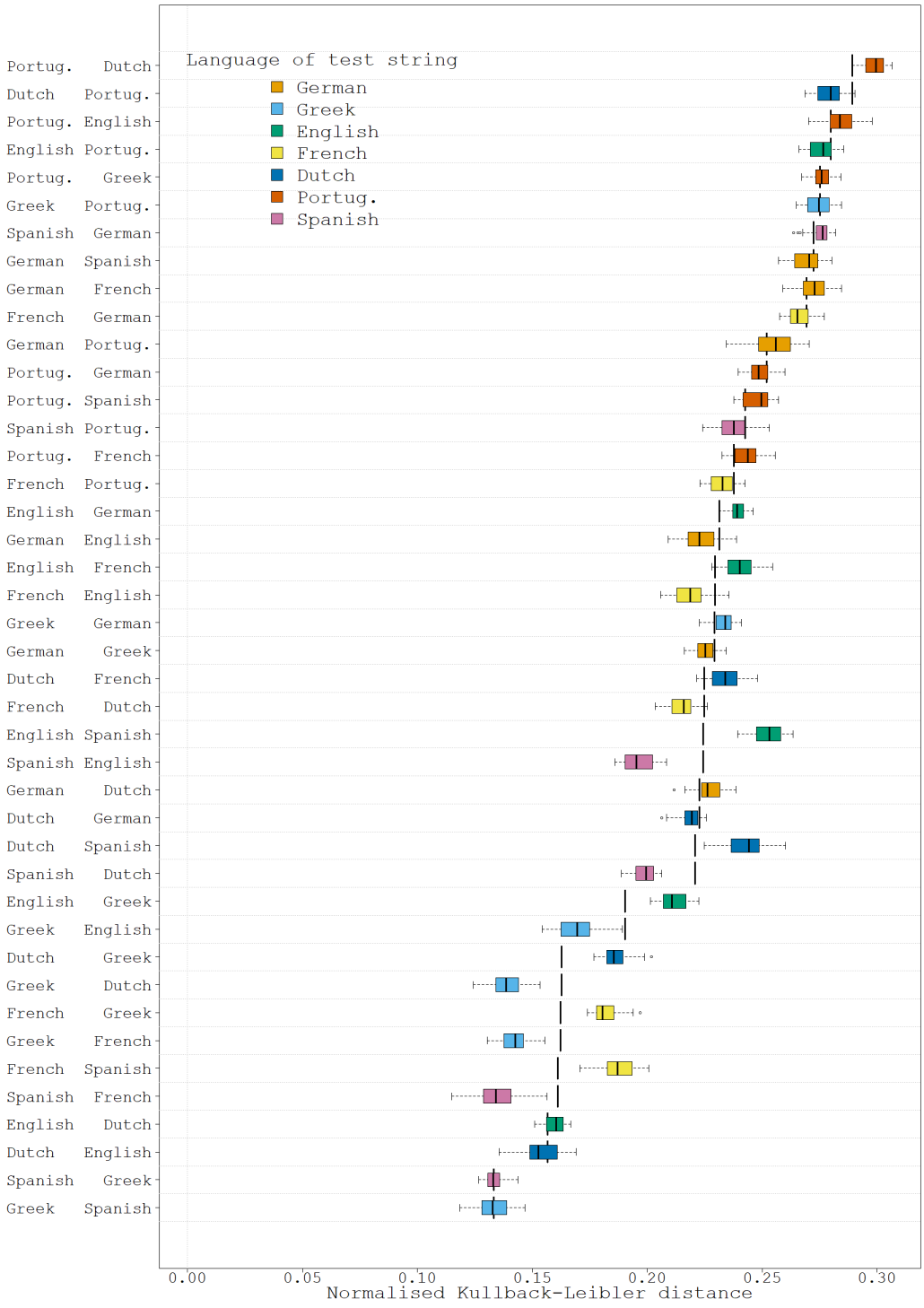


FIGURE 5.13: Kullback-Leibler divergence of language pairs and their inverse, ordered by the mean of the two; SPE-style representation

Predictability per language

Consistent with IPA and static featural representations, Spanish and Greek test strings are the most predictable, and Portuguese is the least, though the ranking is not identical (see Subsection 5.6.6 for a full comparison.)

However, the model languages contribute differently to the GLM for this representation than for the others. For language-specific binary features, text is much more efficiently encoded using a model based on Greek or Spanish than a model based on German.

	Entropy	% increase over Greek
Greek	0.28	-
Spanish	0.29	4
French	0.31	11
Dutch	0.31	11
English	0.32	15
German	0.34	19
Portug.	0.35	26

TABLE 5.20: GLM: Contribution to cross-entropy by language of test string; language-specific binary features

	Entropy	% increase over Greek
Greek	0.29	-
Dutch	0.30	1
Spanish	0.30	3
English	0.31	6
French	0.32	8
German	0.33	13
Portug.	0.35	21

TABLE 5.21: GLM: Contribution to cross-entropy by language of model; language-specific SPE features

5.6.5 Results: Elements

In this section, I examine the combination of all six elements, plus syllabicity. The values for each languages are given in Section A.4 on page 208.

Language identification

The correct language for each test string was identified reliably (i.e. in 100% of cases) for test strings of length 25 characters or longer.

The best fit curve is $y = 100(1 - 1.217e^{-0.415x})$, where y is the percentage of test strings whose language is correctly identified, and x is the length of the test string. Therefore, to achieve 100% accuracy using 100 test strings in 99% of experiments, a test string of length 34+ is required, as for the IPA. A test string of length 250 has an error rate of 1 in 10^{43}.

Consistency of Kullback-Leibler divergence

Each of the six elements can be either headed, unheaded or absent. This gives $3^6 \times 2$ possible combinations, including syllabic and non-syllabic segments, resulting in a normalisation constant of 10.5.

Applying 10-fold cross-validation, the mean error is 0.017, the 99th percentile is 0.052, and the maximum error is 0.076. Normalised Kullback-Leibler divergences range from 0.15 to 0.54, so the error is ±13% of the range of values observed. This is the least stable of the four representations tested, though not significantly different.

Symmetric Kullback-Leibler divergence is visualised in Figure 5.14 and Figure 5.15.

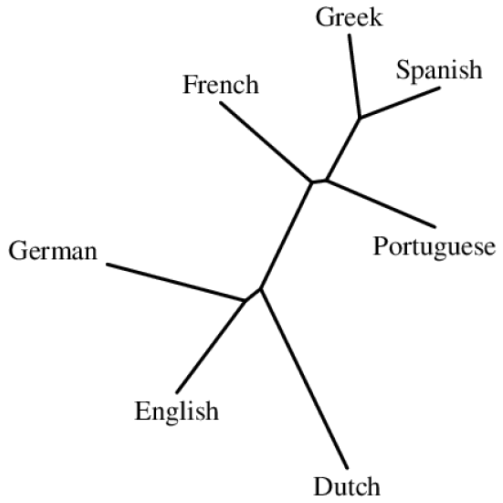


FIGURE 5.14: Visualisation of mean symmetric Kullback-Leibler divergence, Element representation (Dereeper et al., 2008, Felsenstein, 1989)

Asymmetry of Kullback-Leibler divergence

Figure 5.16 shows the asymmetry in Kullback-Leibler divergences calculated from Element representations. The only language pair which is not significantly asymmetric is English and Dutch.

As with the previous two representations, this asymmetry varies in magnitude (Table 5.22 and Table 5.23).

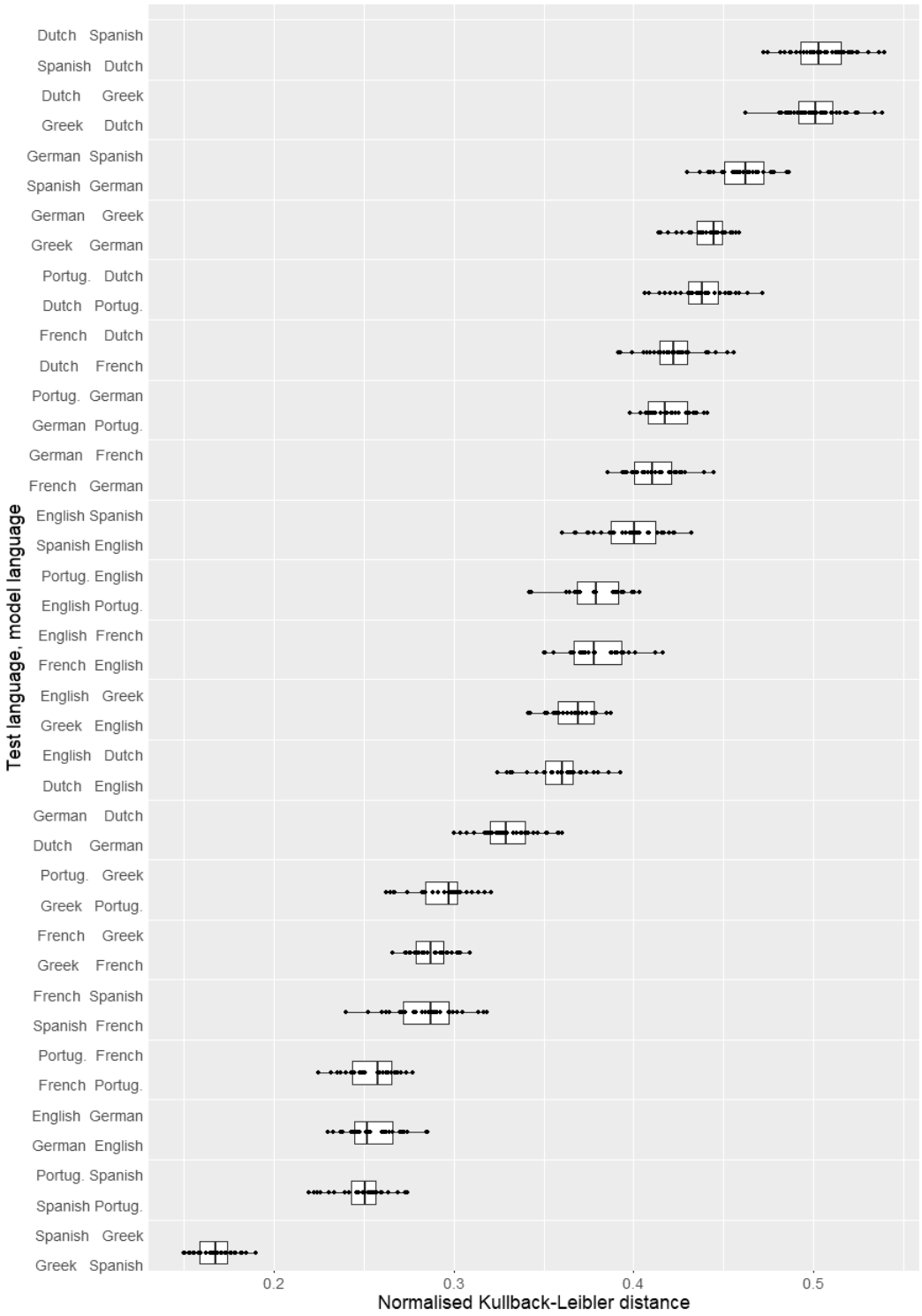


FIGURE 5.15: Symmetric Kullback-Leibler divergence; Element representation

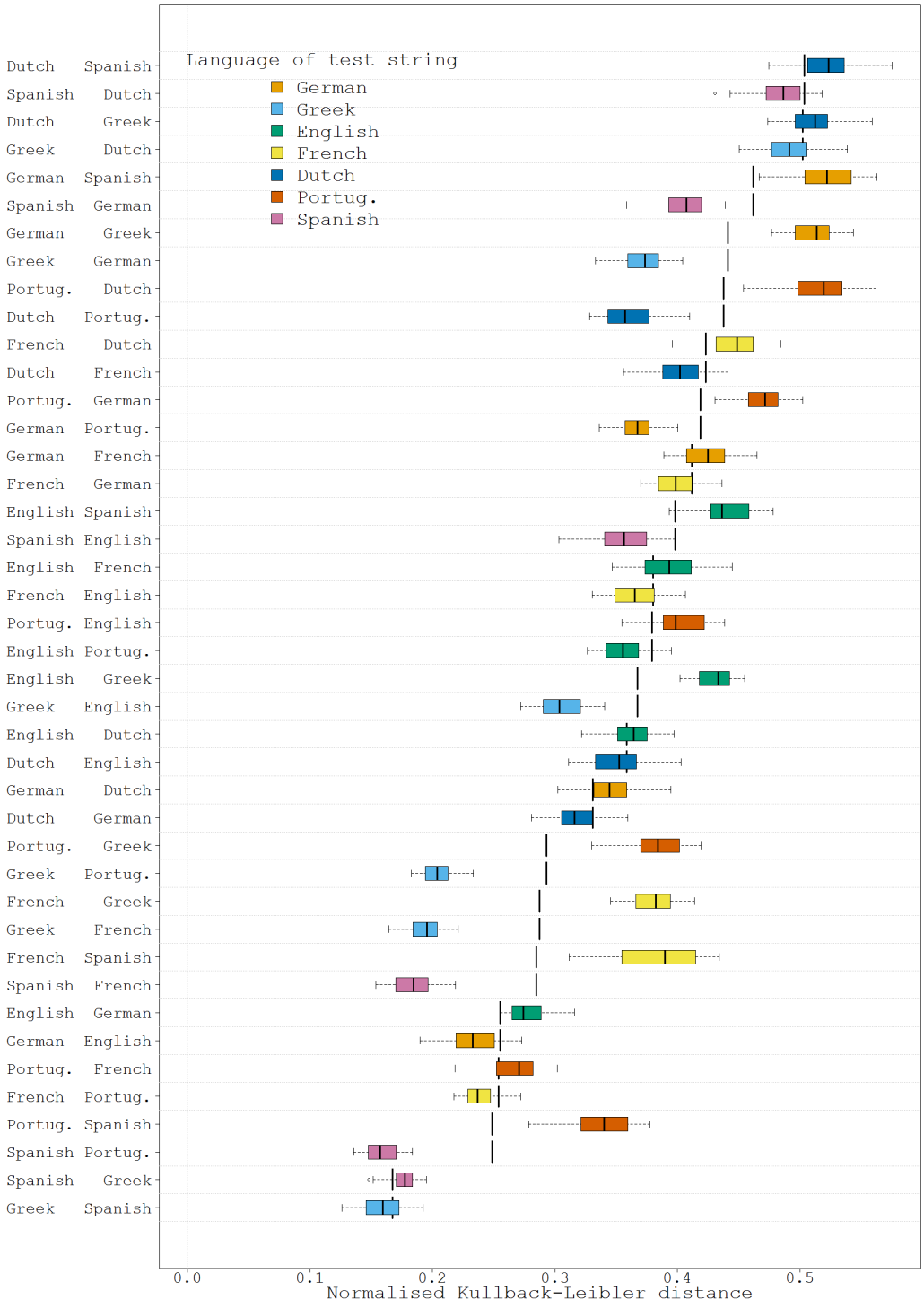


FIGURE 5.16: Kullback-Leibler divergence of language pairs and their inverse, ordered by the mean of the two; Element representation

Language pair	Inverse	Probability (order of magnitude)
Greek German	German Greek	-45
Greek French	French Greek	-44
Dutch Portug.	Portug. Dutch	-38
Greek English	English Greek	-34
Greek Portug.	Portug. Greek	-34
Spanish Portug.	Portug. Spanish	-33
Spanish German	German Spanish	-32
German Portug.	Portug. German	-29
Spanish French	French Spanish	-29
Spanish English	English Spanish	-20
Dutch French	French Dutch	-13
Spanish Dutch	Dutch Spanish	-12
German English	English German	-12
English Portug.	Portug. English	-10
Dutch German	German Dutch	-9
French German	German French	-6
Greek Spanish	Spanish Greek	-6
French Portug.	Portug. French	-6
Greek Dutch	Dutch Greek	-6
French English	English French	-4
Dutch English	English Dutch	-1

TABLE 5.22: Probability that KL distances of language pairs and of their inverses were drawn from the same distribution; Element representation

Predictability per language

The language of the model and the language of the test string are both significant factors in cross-entropy. Greek and Spanish test strings require the fewest bits, and German and Dutch the most (Table 5.24). The model for Portuguese results in the best compression, on average, and Dutch the worst (Table 5.25).

Model \ Test	Dutch	English	French	German	Greek	Portug.	Spanish
Dutch		3	11	9	-3	36	-6
English	-3		-6	-17	-34	12	-19
French	-11	6		5	-64	11	-70
German	-9	17	-5		-30	24	-24
Greek	3	34	64	30		60	12
Portuguese	-36	-12	-11	-24	-60		-70
Spanish	6	19	70	24	-12	70	

TABLE 5.23: Proportional asymmetry in mean Kullback-Leibler divergences (%); Element representation

Language	Bits required	% Increase over Greek
Greek	0.55	-
Spanish	0.55	1
French	0.62	13
English	0.63	14
Portug.	0.64	17
German	0.65	18
Dutch	0.66	19

TABLE 5.24: GLM: Contribution to cross-entropy by language of test; Entropy representation

Language	Bits required	% Increase over Portuguese
Portuguese.	0.59	-
German	0.59	0
French	0.60	1
English	0.60	1
Spanish	0.63	6
Greek	0.64	8
Dutch	0.64	8

TABLE 5.25: GLM: Contribution to cross-entropy by language of model; Entropy representation

5.6.6 Comparison between representations

In this section, I compare the results of each of the four representations for language distance, asymmetry and language predictability.

Language distances

Figure 5.17 and Figure 5.18 show the language distances for all four representations.

All four representations show strong similarities between Greek and Spanish. The Germanic languages all have small Kullback-Leibler distances, with some specific variation: the aspirating languages English and German are closer, relative to Dutch, in the element representation; English-Dutch are closer than the other pairs in the language-specific binary features representation.

The IPA, static binary features and element representations all have small Kullback-Leibler divergences where French, Spanish or Greek are the test strings, and French or Portuguese are the model strings, but not vice versa. The language-specific binary features representation has symmetrically close relationships for French, Spanish and Greek - but larger distances for Portuguese regardless of the test language.

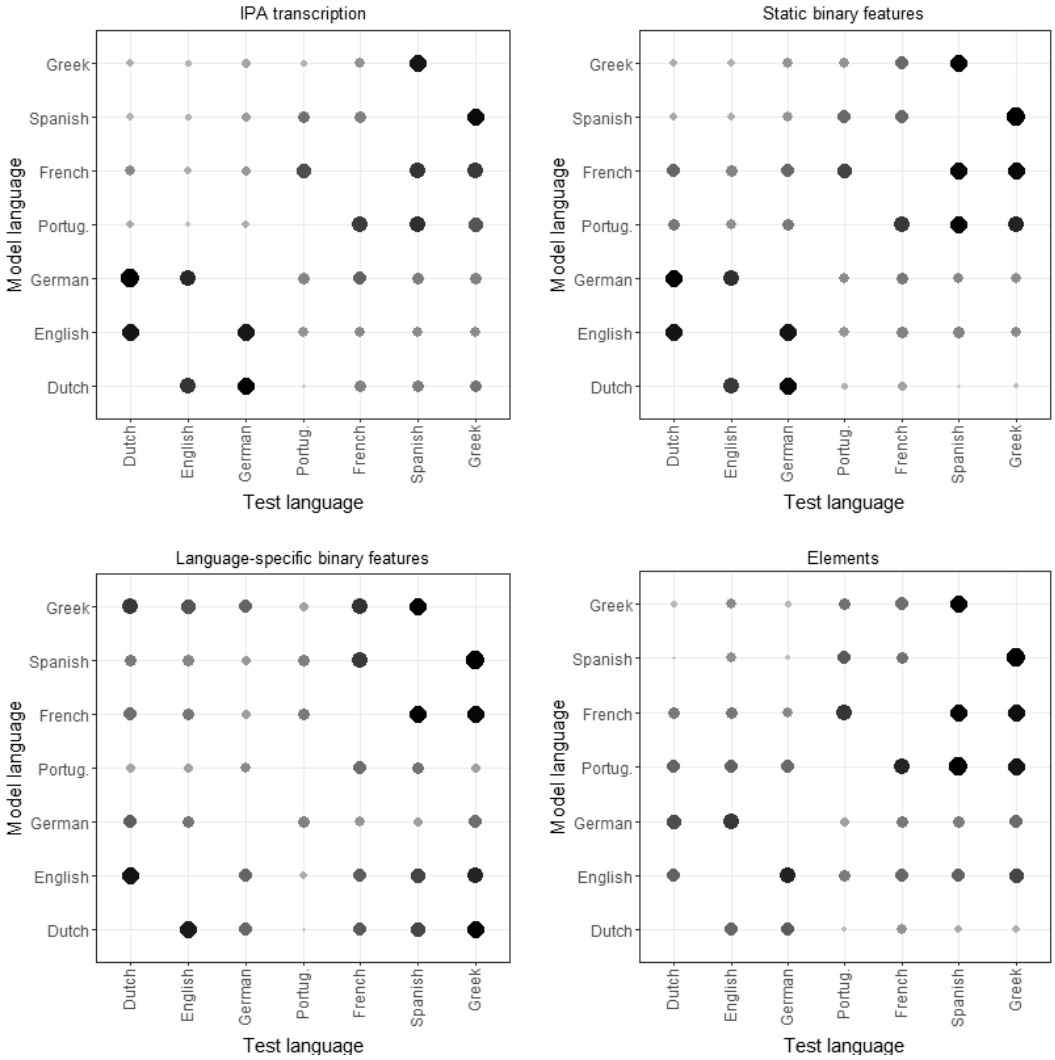
Finally, the language-specific binary features representation differs from the others in that it has relatively small Kullback-Leibler divergences between Greek and Dutch - in both directions - and between the test languages of French, Spanish and Greek and the model languages of English and Dutch. However, since this representation has larger Kullback-Leibler divergences than the other representations - the predictive power of its models is worse, on average - the absolute figures are similar for these language pairs in the other representations.

Table 5.26 shows that the IPA and static SPE representations give very strongly correlated rankings, as expected given the relationship between them. These rankings are in turn strongly correlated with the rankings from Element representation, but only moderately correlated with the rankings from language-specific SPE representation.

TABLE 5.26: Pearson's correlation co-efficient of representations, using mean Kullback-Leibler divergence for each language pair

	IPA	Static SPE	Language-specific SPE	Elements
IPA	-	0.90	0.48	0.71
Static SPE	0.90	-	0.38	0.85
Language-specific SPE	0.48	0.38	-	0.34
Elements	0.71	0.85	0.34	-

FIGURE 5.17: Kullback-Leibler divergences between language pairs for each representation, scaled for optimal visualisation. Larger, blacker points have smaller Kullback-Leibler divergences; smaller, greyer points have greater.



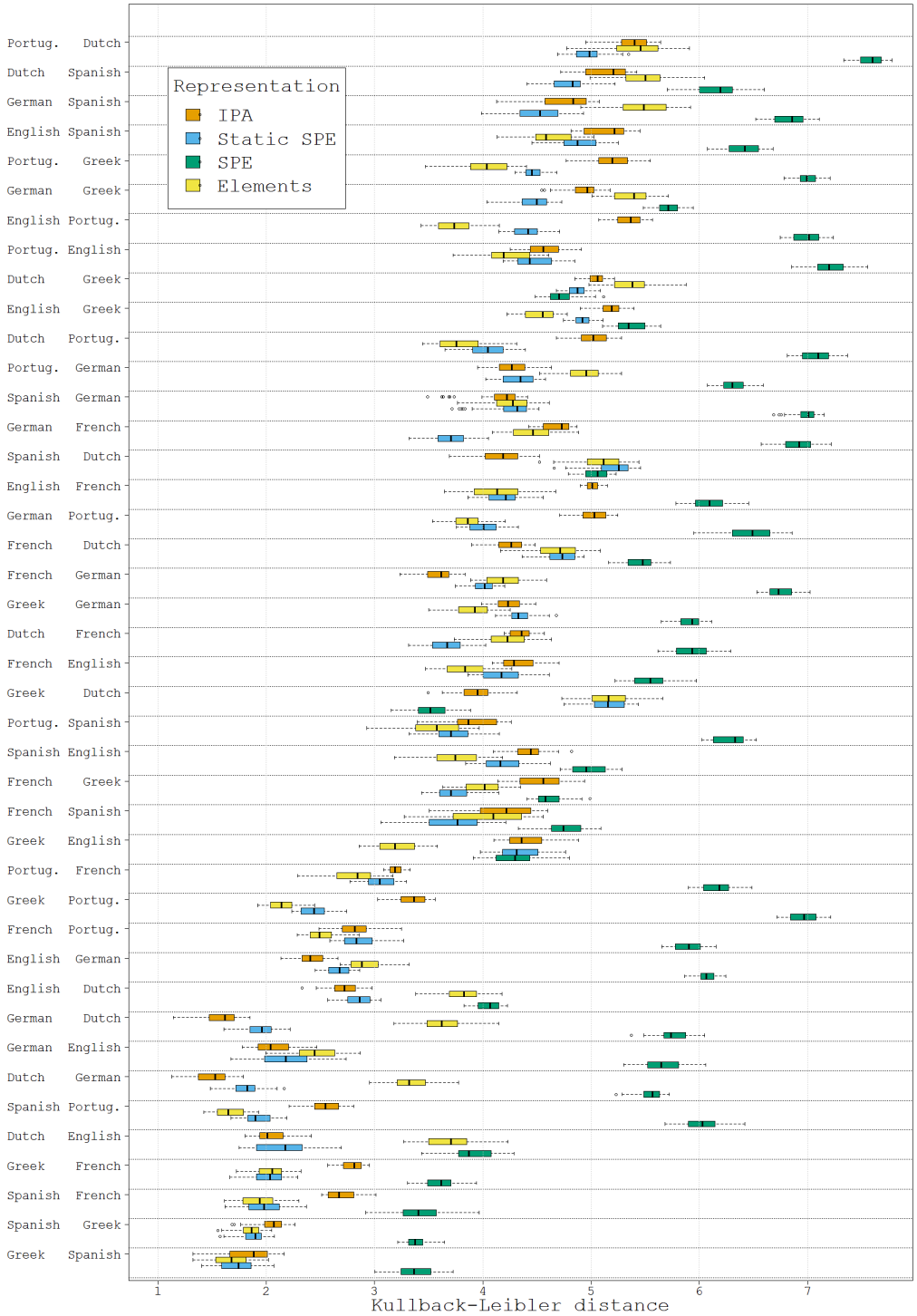


FIGURE 5.18: Kullback-Leibler divergences for each representation

Asymmetry

The observed asymmetry between language pairs is not randomly distributed: the relative asymmetry between language pairs is moderately correlated across most representations ($0.43 \leq r \leq 0.48, p < 0.025$), with strong correlation between the Element and static SPE representations ($r = 0.77, p < 0.005$). However, there was no significant correlation for language-specific SPE with static SPE or Element representations (see Table 5.27)

Given this variability, I do not have strong evidence that these results reflect underlying asymmetries in segmental predictability. However, individual language pairs may have consistent asymmetries across all representations.

The greatest asymmetries are found with Spanish or Greek as test languages and French, Portuguese or English as model languages; these have much lower Kullback-Leibler divergences than the inverse pairs. This is as expected, given the lower entropy of Spanish and Greek test strings by comparison to the other languages. By contrast, the pair Spanish and Greek and the pair German and Dutch show similar Kullback-Leibler divergences regardless of which language is the test and which is the model.

	IPA	Static binary features	Language-specific binary features	Elements
IPA		0.44	0.48	0.43
Static SPE	0.44		0.10	0.77
Language-specific SPE	0.48	0.10		0.22
Elements	0.43	0.77	0.22	

TABLE 5.27: Correlation different representations of between magnitude of asymmetry of language pairs

Language predictability

In this section, I set aside the relative similarity between combinations of languages, and examine the effect on cross-entropy of individual languages themselves. Do languages differ in their segmental predictability?

This section summarises the data previously presented in Table 5.12,Table 5.15, Table 5.20 and Table 5.24 (for test languages) and Table 5.13,Table 5.16, Table 5.21 and Table 5.25 (for model languages).

First, looking at the contribution of the test language to the generalised linear model (GLM) for cross-entropy, I find that there is a similar effect in all four representations. Greek and Spanish have the most predictable test strings, and English, German and Portuguese the least. This is related to segmental inventory size, with Greek and Spanish having 23 and 26 IPA characters respectively, whereas the other language had over 30: Dutch had 34, French 36, English 37, and German and Portuguese had 39. The effect was correlated between all four representations (see Figure 5.19), but with only six languages, most correlations between pairs of representations were not significant ($p >= 0.05$). The exception is the correlation between the IPA and static binary features representations, at $r_s = 0.86$.

The contribution of the language of the model was not correlated between different representations (Figure 5.20).

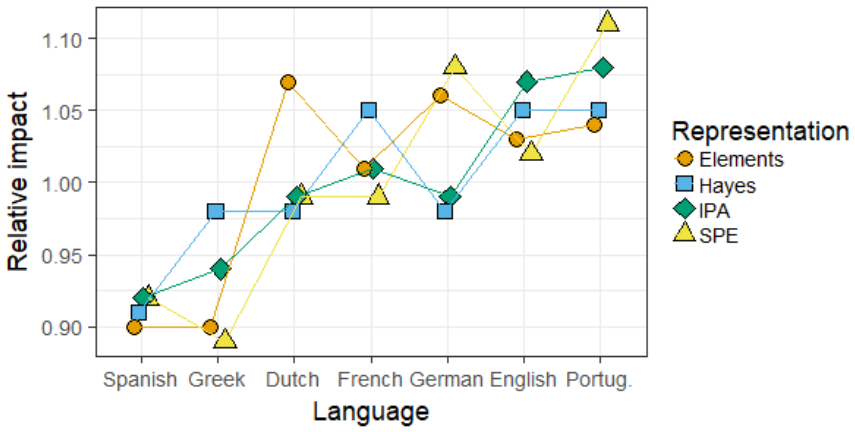


FIGURE 5.19: Relative impact on GLM of test language for each representation

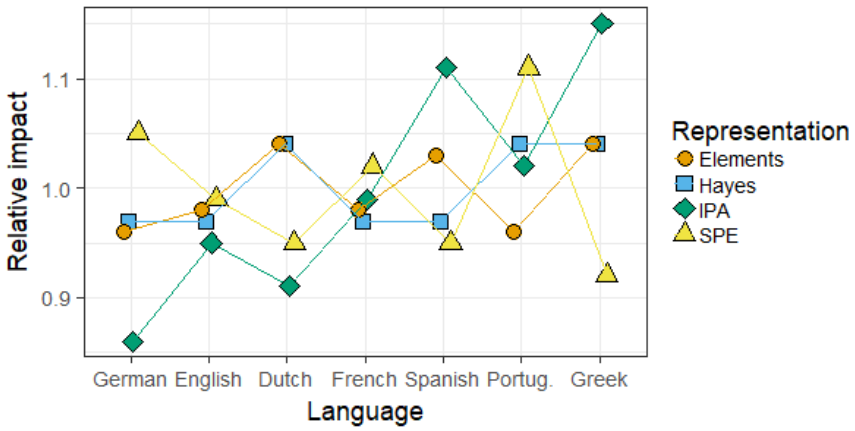


FIGURE 5.20: Relative impact on GLM of model language for each representation

5.6.7 Segments to features

So far, I have examined all features in combination for each representation, effectively using segmental representations with differing levels of phonetic detail and inventory overlap between languages. The end result of this is that all four representations give rise to similar language distances, because there is very little variation in the resulting segment inventories. It is therefore not possible to use these language distance hypotheses as predictions of the representational theories to be tested and compared. This section therefore discusses the predictions made using individual features / elements from the three non-segmental representations.

For each feature/element of each representation, I calculated cross-entropy and Kullback-Leibler divergence as I did for feature bundles in Subsection 5.6.2 - Subsection 5.6.5.

Static SPE-style features

As discussed in Subsection 5.2.2, a string representing a single binary feature would look like so:

++ ++− ++++ ++ ++ −++ ++ ++−+−++−++ ++− +++ ++ −−+++++
ðə ˈnɔθ ˌwɪnd ən ðə ˈsʌn wə dɪsˈpjutɪŋ ˈwɪʃ wəz ðə ˈstɹɒŋgə

For individual static features, the maximum entropy is 1.6, assuming three potential states per character.

Individual features require more input data than combinations; with test strings of length 500 characters, not all test strings returned the language of the test strings as the language of the model having the lowest entropy. I therefore used longer test strings which did reliably return the correct language, of length 800 characters.

Running 10-fold cross validation, the stability of Kullback-Leibler divergence calculated using individual features ranges from slightly more stable than combinations (for [round] and [tap]), to much less stable - see Table 5.28.

Language specific SPE-style features

Turning to the language-specific binary features, even with the longest available test string of 900 characters, the language of the test string cannot be reliably identified with single features.

Combining the features into laryngeal, place, and manner bundles, test strings of over 400 characters could be reliably identified as to language. The manner bundle provided the most

TABLE 5.28: Kullback-Leibler values and error for individual static features

	Mean value	Range of values	Mean error	99^{th} percentile error	Max error	Error as % of range
round	0.21	1.79	0.046	0.16	0.18	9%
tap	0.25	1.47	0.036	0.13	0.14	9%
anterior	0.33	1.16	0.042	0.13	0.17	11%
consonantal	0.32	1.12	0.044	0.13	0.19	12%
labiodental	0.19	1.13	0.041	0.14	0.18	12%
voice	0.25	0.90	0.037	0.11	0.13	12%
spread glottis	0.20	1.20	0.041	0.15	0.18	13%
constricted glottis	0.17	0.97	0.030	0.13	0.15	13%
distributed	0.32	0.89	0.039	0.12	0.17	13%
implosive	0.24	1.02	0.039	0.13	0.15	13%
lateral	0.19	1.15	0.050	0.15	0.19	13%
syllable	0.33	1.06	0.048	0.15	0.16	14%
sonorous	0.26	0.91	0.042	0.13	0.16	14%
delayed release	0.29	0.84	0.038	0.12	0.16	14%
strident	0.29	0.85	0.041	0.12	0.16	14%
continuant	0.27	0.93	0.043	0.13	0.17	14%
dorsal	0.30	0.91	0.043	0.13	0.15	14%
coronal	0.26	0.89	0.042	0.13	0.15	14%
tense	0.52	1.16	0.049	0.17	0.20	15%
trill	0.22	0.87	0.036	0.13	0.15	15%
labial	0.20	0.76	0.038	0.12	0.16	16%
approximant	0.30	0.90	0.046	0.15	0.17	17%
long	0.18	1.09	0.039	0.19	0.23	17%
back	0.36	0.85	0.041	0.15	0.17	18%
front	0.39	0.75	0.042	0.13	0.18	18%
nasal	0.21	0.89	0.053	0.16	0.20	18%
high	0.36	0.73	0.044	0.13	0.18	18%
low	0.33	0.68	0.046	0.13	0.16	19%

stable ranking, and the laryngeal bundle the least, reflecting the number of individual features which combine to produce them. (See Table 5.29 for results from 800 character test strings.)

Whilst each individual bundle gives a stable ranking, these rankings are not correlated between bundles (Figure 5.21). This means that it is possible to contrast the differing effects of different features on language distance. However, combining the bundles into average results does not result in reliable language distance calculations. This could potentially be mitigated by using longer test strings - the segmental tests run previously used strings with lengths an order of magnitude greater than the threshold for language identification.

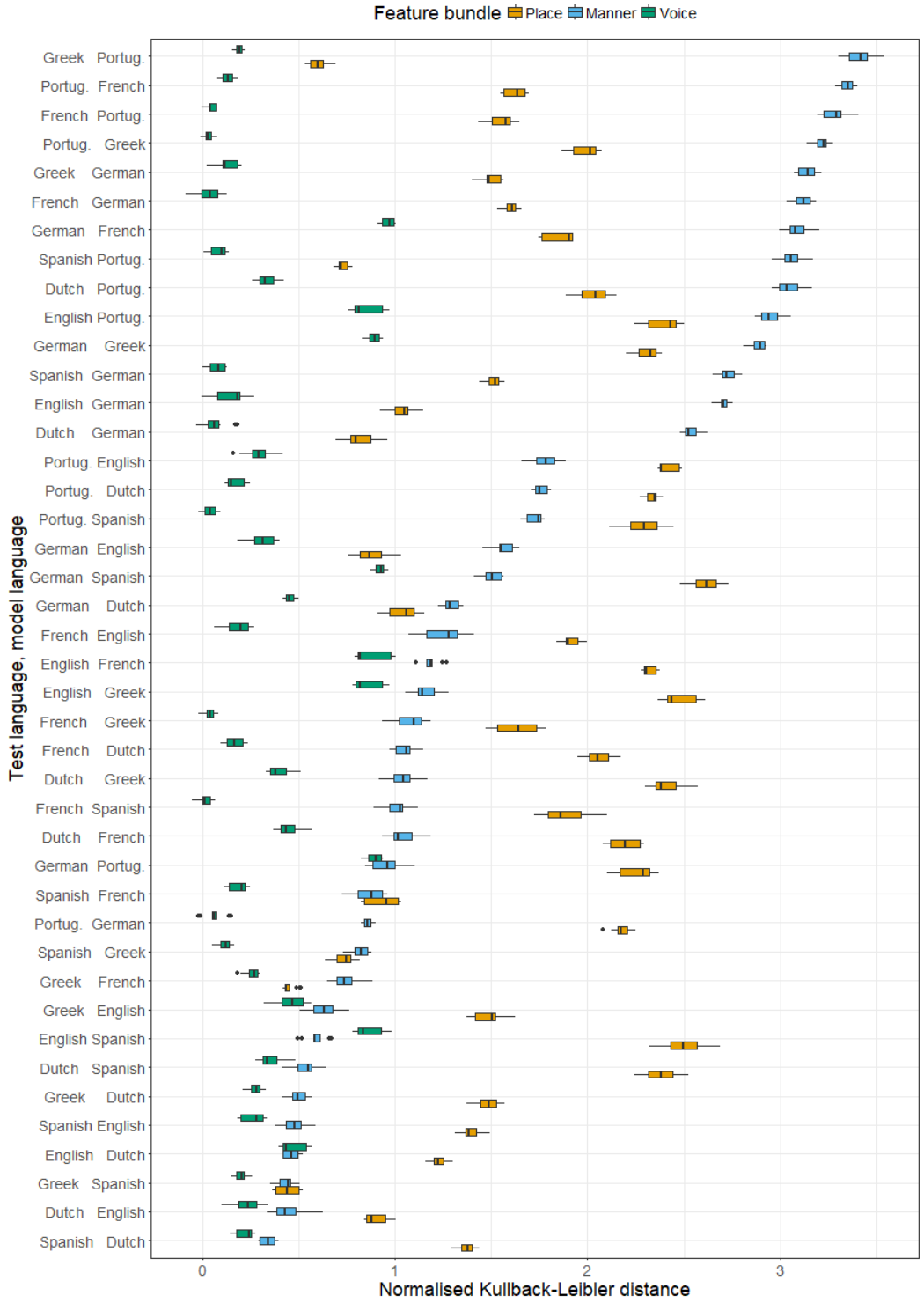


FIGURE 5.21: Kullback-Leibler divergence of language-specific SPE-style feature bundles, ordered by manner

TABLE 5.29: Kullback-Leibler values and error for laryngeal, place and manner for language-specific SPE-style features

Feature	Mean	Range	Mean error	99th percentile error	Maximum error	Percentage error
Manner	1.61	3.25	0.05	0.16	0.20	5%
Place	1.65	2.37	0.06	0.18	0.22	8%
Laryngeal	0.33	1.09	0.05	0.15	0.18	13%

Elements

Firstly, I examined strings of individual elements, with headed and unheaded elements treated separately:

|A|

−+ −−−−+−− ++ −+ −−+ +−−−+ +−+−+
ðə bɪgmɪŋ əv ðə gʊd nju:z əbaʊt

|A̲|

−− −−−−−−− −− −− −−− −−−−− −−+−−
ðə bɪgmɪŋ əv ðə gʊd nju:z əbaʊt

As with the language-specific binary features, individual elements cannot reliably identify the language of the test string with test strings of 900 characters or under. I therefore combined presence with headedness into a single character, such that each composite character has three possible states: absent, unheaded or headed. These states are entirely independent, such that, for example, a pattern involving headed A will not aid in identifying the same pattern in a different language which refers to unheaded A instead. However, this approach does help identify patterns in which the headedness of an element has predictive power for neighbouring unheaded versions, or vice versa.

|A| with |A̲|

–A −−−−A−− AA –A −−A A−−−A A–B–A
ðə bɪgmɪŋ əv ðə gʊd nju:z əbaʊt

Using these bundles as characters, test strings of 800 characters or more are reliably identified. (With test strings of length 700 characters, one of the 28 English test strings was misidentified as Dutch.)

There is a similar range of stability in elements as in static binary features, with most individual elements being similar in stability to, or slightly less stable than, their combination into a single bundle, which had an error rate of 13%. (See Table 5.30). H gives more stable results than combining all elements, perhaps because of the stark division of languages in this sample into voicing and aspirating languages, with no confounds from tone or breathy voice.

TABLE 5.30: Kullback-Leibler values and error for Elements

Element	Mean	Range	Mean error	99^{th} percentile error	Maximum error	Percentage error
H	0.42	1.60	0.04	0.14	0.17	8%
I	0.36	1.41	0.05	0.18	0.25	13%
A	0.30	1.08	0.04	0.15	0.17	14%
ʔ	0.22	0.94	0.04	0.13	0.16	14%
Syllabicity	0.29	0.73	0.04	0.12	0.15	16%
L	0.30	1.53	0.05	0.25	0.32	17%
U	0.18	0.60	0.04	0.13	0.15	22%

As well as varying in their stability, individual elements vary in their predictability (see Table 5.31). For example, |A̲| is approximately 1.5 times more predictable than |A| on average, and |H| 20% more predictable than |I|. This implies that the functional load is different between different elements. An interesting avenue of future research would be to compare these findings between different languages, and look for phonetic or phonological phenomena which correlate to differences in modelled predictability. Likewise, such calculations could be used to compare the implications for information transer of different analytical decisions in element assignment.

5.6.8 Transparent segments

In Subsection 5.6.2 – Subsection 5.6.5, I have examined segments represented as monolithic blocks, and in Subsection 5.6.7 as individual, independent features. An alternative would be to calculate cross-entropy once for each test string, from a representation in which the internal structure of each segment is more transparent. For example, a linear listing of feature values:

−+-++—-+——+++— +−+++—+———-+——
ðə

TABLE 5.31: Predictability by feature

Element	Bits required	Increase over minimum	% increase
ʔ	0.71	-	-
A	1.13	0.43	60%
U	1.19	0.49	69%
L	1.20	0.49	70%
H	1.28	0.57	81%
H	1.37	0.66	94%
I	1.38	0.68	96%
L	1.42	0.72	101%
ʔ	1.53	0.82	117%
Syllabicity	1.53	0.83	117%
U	1.56	0.85	121%
I	1.65	0.94	133%
A	1.73	1.03	146%

However, because the Text Mining Toolkit was designed for orthographic text, it uses a context limit smaller than the number of binary features in a single segment. Representing a segment as a linear list of feature values would therefore mean that, for example, the feature value for voicing in one segment would be too far away to form part of the context for the value of voicing in the next segment; there would be too many intervening values. A possible future avenue of investigation is the efficiency and reliability of extending the context limit of the entropy calculations.

5.6.9 Hypotheses: summary

1. The language of a test string can be reliably identified.

   For all segmental representations, the language of a test string can be reliably identified out of the seven options with under 50 characters. The language can also be identified for feature bundles with fewer features, using longer strings. Test strings comprised of individual features/elements are not identifiable using under 800–900 characters.

2. The minimum required test string length for reliable language identification is consistent.

   The percentage of correctly identified strings by length of string follows an exponential curve for each representation.

3. If the language of the test string can be reliably identified, the Kullback-Leibler divergences between each pair of languages will be consistently ranked.

   Whilst the Kullback-Leibler divergences calculated using each representation did not result in identical rankings (see Figure 5.18), each ranking was internally consistent; the 99th percentile error from 10-fold cross-validation was between 9% – 13% of the range for segmental representations, and between 5% –22% for individual features / elements.

4. The Kullback-Leibler divergence is symmetrical for all language pairs.

   The majority of language pairs show significant asymmetry, though the magnitude of the asymmetry per pair is not consistent across all representations.

5. Languages do not differ in their segmental predictability

   Languages with smaller inventories had consistently more predictable test strings. There was no significant variation in predictability given the language of the model. (See Subsection 5.6.6)

6. Every feature encodes the same amount of information

   For all representations, there is considerable variation in predictability between different features / elements. The alignment between theoretical dependencies and the observed information transmission could be a fruitful avenue for future research.

5.7 Conclusion

Entropy is a measure of the amount of information in a given system, and cross-entropy a measure of the information in a representation of that system. In this chapter, I have shown that the distance between a pair of languages, for a given phonological representation, can be measured using the difference in their cross-entropies, called the Kullback-Leibler divergence.

I have shown this transparently, using a basic unigram calculation applied to IPA characters. This method also replicates Juola's earlier work using orthographic representation.

I have also shown this using Teahan's Text Mining Toolkit, which calculates a more accurate entropy value with less data. It has been applied to representation of texts in the IPA; binary

features, statically mapped to the IPA; language-specific binary features; and elements. The language distances of these representations are correlated, but not uniform. However, this does not mean that they are inaccurate for a given representation. The reliability of findings for each representation was established with cross-validation, finding that the 99th percentile error rate was around 10% of the range of Kullback-Leibler divergences.

Further work is required to establish which factors external to segmental representation affect cross-entropy. However, even a partial measure of entropic distance can also inform investigation into phonological representation. The approach outlined in this chapter can be applied to any system of representation, aiding researchers in reflecting on implications of their theory for information transfer, such as whether all features carry equal quantities of information in real usage.

Chapter 6

ACCDIST

The previous two chapters have described metrics which rely on phonological representations of speech. In this chapter, I use an accent distance metric – ACCDIST – to measure more directly the similarity between audio recordings.

I have established that most existing 'metrics' of language distance used in second language acquisition are insufficient for comparing phonological knowledge. They are mostly either too subjective, based on personal impressions of how similar different linguistic systems are, or too broad, including factors like dissimilarity in the writing system. Similarly, historical linguistic relationships based on cognates do not provide a measure of similarity for unrelated languages, and again are based on factors other than the phonological system.

On the other hand, phonetic comparisons are too specific to compare languages as a whole. The similarity of individual recordings of speakers depends on physical factors such as height, age or gender. This problem of separating out individual variation from variation between speech communities is an important issue for speech recognition, speech synthesis and accent identification. As such, there are several methods for modelling accent distance whilst controlling for these other factors. These methods can equally be applied to second language speakers, giving a baseline for the similarity of pronunciation of speakers from different language backgrounds to which to compare my metrics.

6.1 Language identification techniques

The principal techniques used in spoken language identification (LID) are Phone Recognition and Language Modelling (PRLM) and Gaussian Mixture Models (GMM) (e.g. Zissman, 1996,

Gelly and Gauvain, 2017). In PRLM, the speech is first segmented into phones, then an n-gram probability model is estimated (see Section 5.3). This technique and its successors rely on phonotactics for identification. In GMM, the speech signal – generally processed into a discrete form – is modelled as a combination of latent variables. These components comprise speaker-dependent characteristics (e.g. gender, age), channel-dependent characteristics (e.g. microphone, background), and others, and may not necessarily be specified in advance. In some LID systems (e.g. Gelly and Gauvain, 2017), the language similarity component of the model can be factored out, but this is not a universal feature.

I have chosen to use the ACCDIST system (Huckvale, 2004), as described below, due to its non-proprietary nature, ready availability, small input data requirements, and transparent inner workings.

6.2 ACCDIST

ACCDIST (Huckvale, 2004) is a metric based on the relative similarity of a speaker's realisations of different segments. For example, a northern British English speaker will pronounce the stressed vowels in 'after' and 'cat' with greater similarity than 'after' and 'father', whereas in a southern British English speaker, this pattern would be reversed. ACCDIST has been used to successfully group British English speakers into their respective accent groups, with regional accent groups clustering together (Huckvale, 2004).

There is also a correlation between the similarity of accent of talkers and listeners (as measured by ACCDIST) and their mutual intelligibility (Pinet, Iverson and Huckvale, 2011). This correlation holds for foreign-accented speech as well as regional variation; Pinet, Iverson and Huckvale's experiment used speech samples from Standard Southern British English (SSBE), Irish English, Korean-accented English, bilingual French-English, experienced French-accented English, and inexperienced French-accented English.

There are several advantages to applying the ACCDIST metric to the issue of second-language accented speech, rather than examining mutual intelligibility directly. ACCDIST has the advantage of being extensible to more languages in future, subject to the availability of suitable input data. By contrast, most mutual intelligibility studies only compare two or three languages, and

would need replicating in their entirety in order to compare ten or more languages; additional languages cannot simply be added on.

ACCDIST is a more direct analogue of the phonological distance which I am trying to measure than mutual intelligibility is. Mutual intelligiblity depends on a variety of factors, of which accent distance is only one, such as familiarity (Adank et al., 2009).

6.2.1 Method

ACCDIST is calculated as follows. The same base text is recorded by each speaker. A single idealized transcription consisting of phonemes-in-words (e.g. a/after, a/cat, a/father) is aligned with the recording. This transcription is identical for all speakers, regardless of actual phonetic detail, and is used to locate corresponding segments and compare them between speakers.

Each vowel segment is represented as a set of mel-frequency cepstral coefficients (MFCCs). This involves three transformations of the speech signal. Firstly, a Fourier transform is applied to get a representation of the signal as a function of frequency. Secondly, it is scaled using the Mel scale (Mermelstein, 1976), which represents human perception of pitch, with each equal step of the scale perceived as the same difference in pitch. Thirdly, a logarithm is taken, which permits the separation of fundamental frequency and formant data.[1]

For each of the experiments below, the processing from audio recording to MFCCs was done using the Speech Filing System (Huckvale, 2008). (See Appendix Section B.3 for links.)

The position of vowels in each recording was identified programatically using Analign, based on "a set of phone hidden-Markov models which have been trained on Southern British English" (Huckvale, 2008). However, using a language-specific model for this task is not a significant limitation for my purposes. The first experiment is based on English text, even if vowel quality differs between speakers. Later experiments using non-words require only the alignment of a single CVCV item per file, and English approximations are sufficient for this. (See Subsection B.3.3 for the mapping used.)

The MFCCs for a given segment could be compared directly between speakers, but this would group speakers by personal characteristics such as gender. Instead, with the ACCDIST

[1]The source signal (vocal excitations) is convolved with the filter (vocal tract), resulting in the speech signal. A Fourier transform turns this convolution into multiplication, and taking a logarithm transforms this into a linear addition, which allows the terms to be examined separately.

method, the MFCCs for a given speaker are first compared across their segments. These relative differences can then be compared across speakers, which removes the personal characteristics.

Speaker clustering based on ACCDIST measurements can be performed in several different ways. In the experiments below, I have measured the average correlation between speakers, as this is both accurate in categorising English speakers (Huckvale, 2007) and gives fairly stable results with low numbers of speakers.

6.3 Speech Accent Archive

For my first analysis, I used data taken from the Speech Accent Archive (Weinberger, 2015). This is a database of recordings of a passage of English by speakers from a variety of linguistic backgrounds. The passage contains most of the segments and sequences of General American English, using common vocabulary items:

> "Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."

6.3.1 Input data

I wanted to ensure that the vowels in the sample were as close to speakers' L1 language productions as possible. Pinet, Iverson and Huckvale (2011) found that the closeness of French speakers' vowel spaces to SSBE speakers' increased with increased English experience. For this reason, I chose samples from non-native speakers who had learned English in an academic context, and who had spent less than six months living in an English-speaking country.

As far as possible, samples for each language were chosen from a single region/country to maximise similarity. This obviously does not necessarily form a coherent accent group – e.g. 'south-eastern English' would include both Standard Southern and Norfolk accents – but there are not yet enough samples from L2 speakers in the archive to be more specific.

I examined six male and two female Dutch speakers; five male and two female English speakers; five male and three female French speakers; six male and two female Italian speakers; two male and six female Korean speakers; seven male and one female Polish speakers; and two male and three female Portuguese speakers (see Subsection B.2.1).

Using American English orthography introduced additional variation unrelated to phonetic effects. The differences in pronunciation values of the Latin alphabet across languages can reflect historical accident, rather than contemporary differences. For example, some L2 speakers of all backgrounds failed to apply diphthongisation before a silent ⟨e⟩, producing ⟨snake⟩ as ⟨snack⟩, though [æ] and [eɪ] would not be adapted as a single phoneme in those languages aurally. Some British speakers stumbled over American English vocabulary ('snow peas') or grammar ('meet her Wednesday'), sometimes 'correcting' the passage to British English.

6.3.2 Results

The most homogenous language group – the group closest to their colinguals – were Portuguese, followed by Dutch, Polish, Italian, English, Korean, and finally French. French speakers were the only group to show greater mean similarity to speakers of other languages (Portuguese, Dutch, Italian, Polish) than to their colinguals. (See Table 6.1.)

TABLE 6.1: Mean distance between speakers of different languages

	Dutch	English	French	Italian	Korean	Polish	Portuguese
Dutch	**0.33**						
English	0.39	**0.38**					
French	0.43	0.48	0.44				
Italian	0.38	0.43	0.43	**0.38**			
Korean	0.42	0.48	0.45	0.42	**0.41**		
Polish	0.37	0.40	0.43	0.39	0.42	**0.36**	
Portuguese	0.36	0.44	**0.41**	0.38	0.38	0.37	**0.30**

Are the language groups sufficiently distinguishable from one another that between-language speaker distances are significantly different from within-language speaker distances?

Let δ_{vi} and δ_{vj} be the distance between MFCCs for each speaker i and j respectively, for each pair of vowels v. Let $\Delta_i = \sum_v^V \delta_{vi}$, where V is the total number of vowel pairs, and $\Delta_{ij} = \sum_v^V \delta_{vi}\delta_{vj}$.

The Pearson's correlation coefficient between the two speakers i and j is measured as

$$r_{ij} = \frac{V\Delta_{ij} - \Delta_i\Delta_j}{\sqrt{V\Delta_{ii} - \Delta_i\Delta_i}\sqrt{V\Delta_{jj} - \Delta_j\Delta_j}}$$

Let D_{ij} be the distance between two speakers, equal to $1 - r_{ij}$. Let $\mu_{x,y}$ be the mean of $D_{i=x,j=y}$ where all speakers i speak language x, and all speakers j speak language y.

If x and y are not the same, $\mu_{x,y}$ is the mean language distance between speakers of two different languages (e.g. the mean distance between English speakers and French speakers). Where they are the same, $\mu_{x,y}$ is the mean language distance between colinguals, and written as $\mu_{x,x}$ or $\mu_{y,y}$. (E.g. $\mu_{x,x}$ is the mean distance between two English speakers and $\mu_{y,y}$ is the mean distance between two French speakers.)

If $\mu_{x,y}$ is significantly different from both $\mu_{x,x}$ and $\mu_{y,y}$, then speakers of x and y form distinct groups.

$\mu_{x,y}$ is only significantly different from both $\mu_{x,x}$ and $\mu_{y,y}$ in three cases: English-Portuguese, English-French, and English-Korean (see Figure 6.1 on the next page.)

Other language pairs have a significant difference between $\mu_{x,y}$ and $\mu_{x,x}$ but not $\mu_{y,y}$. This means that cross-language distance is not distinguishable from language-internal variation of the language y (see Figure 6.2).

Where x is Portuguese, $\mu_{x,y}$ is significantly different from $\mu_{x,x}$ (mean distance between Portuguese speakers) for all languages y. But it is only significantly different from $\mu_{y,y}$ where y is English, as mentioned above.

Where x is Dutch, $\mu_{x,y}$ is significantly different from $\mu_{x,x}$ (mean distance between Dutch speakers) where y is English, French, Italian or Korean (but not Polish or Portuguese). But $\mu_{x,y}$ is not significantly different from $\mu_{y,y}$ for any of the four.

Finally, μ(Korean, French) is significantly different from μ(Korean, Korean); μ(Polish, French) is significantly different from μ(Polish, Polish); and μ(Korean, Polish) is significantly different from μ(Polish, Polish).

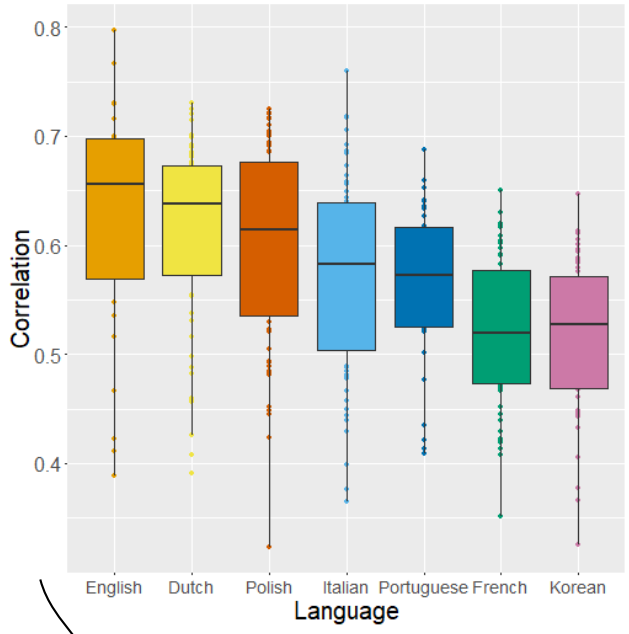
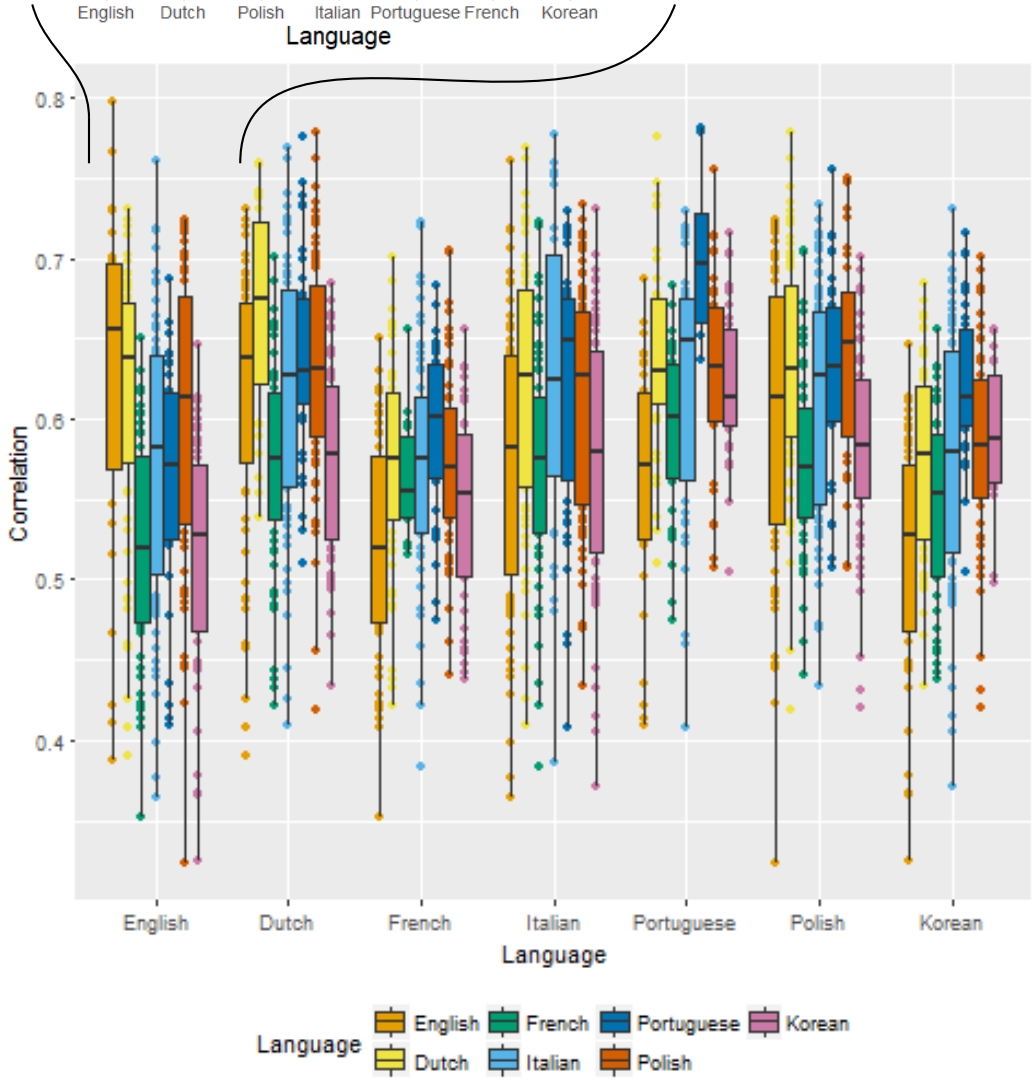


FIGURE 6.1: Left: Correlation r_{ij} between individual English speakers and other individual speakers.

FIGURE 6.2: Below: Correlation r_{ij} between individual speakers, labelled by language background.



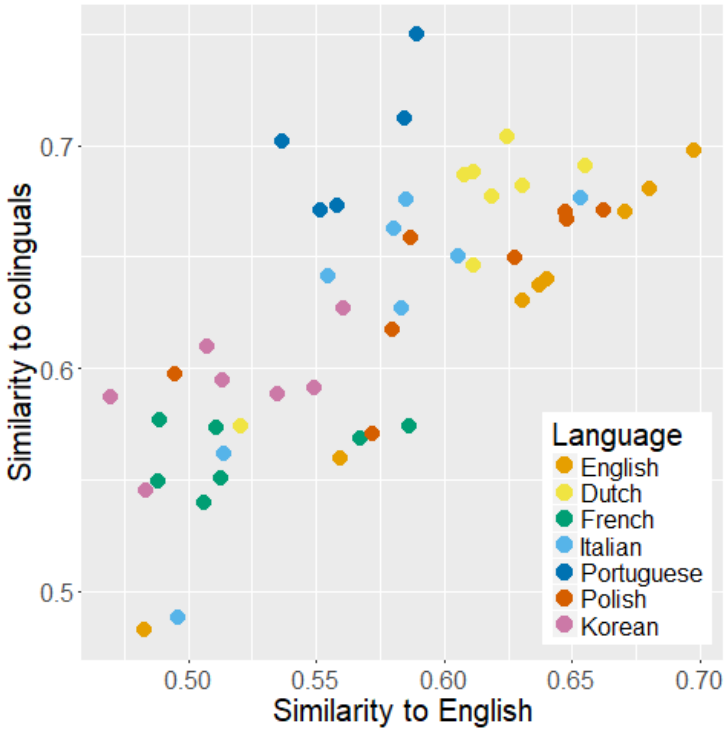


FIGURE 6.3: Correlation of English proficiency with group cohesiveness; each point represents a single speaker

Additional observations

I extended the inventory of analysed segments to include fricatives as well as vowels, since fricatives also have steady-state MFCCs. However, this made the categorisation of speakers by language strictly less accurate.

The average similarity of a speaker to their co-linguals was strongly correlated with the similarity of that speaker to the native English speakers ($\mu_{x,x} \propto \mu_{x,English}; r = 0.76, p < 10^{-8}$; Figure 6.3). It appears that all felicitous pronunciations are alike; each infelicitous pronunciation is infelicitous in its own way.

6.3.3 Conclusions

There were sufficient issues with the audio quality of the samples, and their low number, that I needed to record fresh samples. Given the limitations of using read English, I designed an entirely new set of samples.

6.4 Non-word repetition task

This task was designed to elicit nativised versions of a large cross-section of the vowel space. As far as possible, it is language-neutral; i.e. not biased by education in English.

6.4.1 Methodology

A set of 80 bisyllabic CVCV words were constructed, using 16 vowels and five consonants. Each vowel appears twice in conjunction with each consonant, with the CV syllable in both initial and final position. The five consonants chosen were the most common, cross-linguistically: three plain voiceless stops [p],[t],[k], the alveolar nasal [n] and the voiceless alveolar fricative [s]. The 16 vowels cover the major contrasts of the vowel space: the seven peripheral vowels [i],[e],[ɛ],[a],[ɔ],[o] and [u]; front rounded vowels [y],[œ]; central vowels [ɨ],[ə]; back unrounded vowels [ɯ],[ɤ], nasal [ã], breathy [a̤] and creaky [a̰] vowels.

Each syllable was recorded in isolation with a flat intonation by a trained phonetician. This allowed a variety of different combinations to be generated before a final vowel set was chosen.

Participants were told that "An international department store is expanding into the UK. They want to know how their product names will be pronounced by English-speaking customers". Where possible, all instructions were presented in their native language.

They were presented aurally with a "product name", then visually with a written sentence in their native language with a gap. E.g. "The ____ plates are cheap." For a full list of example sentences, with their translations, please see Subsection B.1.1.

Participants were instructed to read out the sentence with the product name in the gap. They were then asked to repeat the product name again, in isolation. Each participant was given 3 – 9 demonstration items to become comfortable with the task before the 80 test items were presented.

Use of the sentence helped to reduce direct mimicry of the stimuli, and the second repetition was, subjectively, more natural than the first. (I am confident of this judgement regarding the English participants, and also received this as feedback from multiple linguistically aware participants.) This was especially important because the concatenation of different samples to create words from syllables did not result in particularly natural intonation, but rather words that participants variously described as "Chinese", "robotic" or "alien".

13 mel frequency cepstrum coefficients are found for each half of every vowel, capturing the changes in vowel quality present in a diphthong. For a pair of vowels u and v with MFCCs $u_1 ... u_{26}$ and $v_1 ... v_{26}$ their dissimilarity is calculated as:

$$\sqrt{\frac{\sum_{i=1}^{26} u_i - v_i}{26}}$$

i.e. the mean difference between each coefficient.

For each pair of speakers, the correlation between them is calculated as the Pearson Correlation Coefficient of their vowel pair differences. If the same pairs of vowels are similar, the speakers will have high correlation. If one speaker has small differences between pairs of vowels for which the other has large differences, the speakers will have low correlation.

6.4.2 Pilot results

After open recruitment, speakers of the following languages were recorded: Japanese (5), English (4), Spanish (4), Cantonese (1), French (1), Greek (1), and Polish (1). The instructions were translated into Japanese, English, Spanish, French and Greek (see Section B.3).

I shall present here the findings for Japanese, English and Spanish, since those had multiple speakers and hence consistency between co-linguals could be measured.

I applied an Analysis of Variance to the correlation between speakers with the factors of gender identity, age difference and language interaction. There was no significant effect of sharing a gender or of similarity in age, as expected given the design of the ACCDIST calculation. Language interaction was significant ($p < 0.01$), and the effect size was large ($\eta^2 = 0.15$). This is due to a difference between within-language and between-language groupings. In particular, Japanese speakers gave homogenous responses; the only significant difference between interspeaker correlations grouped by language was between Japanese cohesiveness and other pairings. Japanese speakers were most similar to their co-linguals, followed by Spanish speakers to their co-linguals and English to theirs. These is no significant difference between the correlations across language groups (see Figure 6.4).

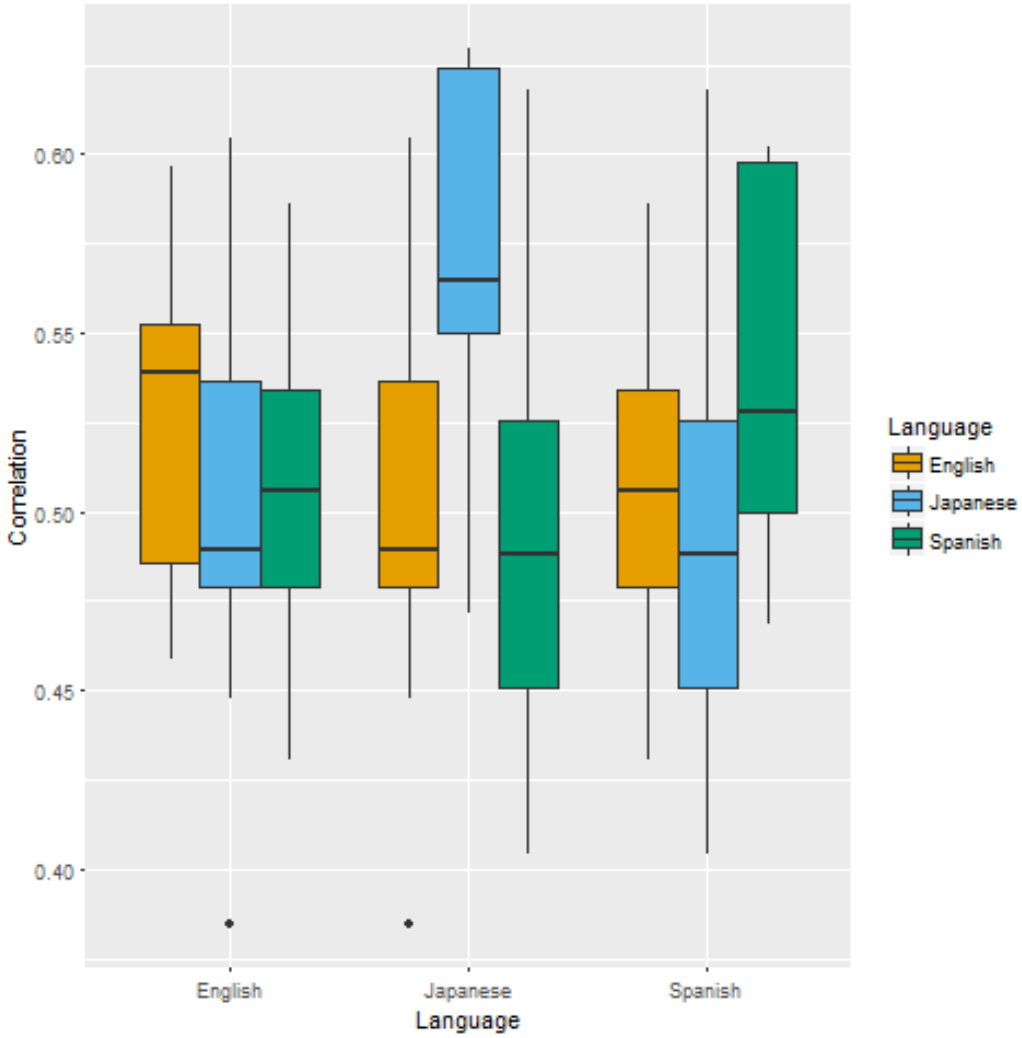


FIGURE 6.4: Correlation between individual speakers, labelled by language background

6.4.3 Alterations following the pilot

English and Spanish had higher correlations within-group than compared to speakers of other languages, but with only four participants, these findings were not significant. I therefore repeated the study with many more participants.

I had the stimuli re-recorded as entire words, since many participants found the concatenated syllables to have unnatural prosody, making it more difficult to perceive them as real lexical items. This was also evident in early trials without examples sentences, in which participants copied both the intonation and the vowel quality fairly exactly, despite instructions to the contrary. The introduction of example sentences, forcing participants to use the trial items in the context of their native language, made a significant difference. The second repetition, in isolation, was subjectively a more nativised version; participants copied their own previous pronunciation in the example sentence.

Participants reported that they found it easier to produce a natural (nativised) version of the trial item in a longer example sentence, and when the item was not sentence initial. The example sentences were modified to fit these criteria, and to be of equivalent length in each language. The new examples had six syllables preceding and six syllables following the test item, such as “I prefer the dark green ____ to the one you’re holding ”or “Me gusta mucho el ____, y es muy barato”. Full examples can be found in Subsection B.1.2. Since the sentences themselves were unimportant, I dispensed with translating the sentences directly, to make the length requirement easier.

Instructions were also repeated verbally for the participants in the second study. In the pilot, participants received written instructions in their native language as part of the consent form, then again screen-by-screen as they became relevant. Summarising the activity verbally between the written form and the start of the experiment reduced problems, but did not entirely eliminate misunderstandings or refusal to follow instructions.

6.4.4 Data

The audio recordings and analysis code described in this section can be found at the link at Section B.3. Participants have agreed to release their recordings into the public domain, along with anonymised demographic data.

The stimuli were presented and audio recorded using Psychopy (Peirce, 2007), which allowed a consistent presentation across languages.[2]

A larger number of speakers was recruited for two languages, English and German, and those speakers were tested in their native countries. Unfortunately, logistical problems prevented the recording of Greek and Spanish speakers in their native countries, so a smaller number of speakers were recorded in the UK.

25 English speakers, six standard Greek speakers and eight global Spanish speakers were recorded in London, UK; and 24 German speakers were recorded in Düsseldorf, Germany by Dr Martin Rönsch. Their demographic data can be found in Subsection B.2.2. Of these participants, I have excluded two English speakers who were not Standard Southern British English speakers and one Spanish speaker who was outside the age range of 18–35, as well as three German and two English speakers whose recordings were unusable due to noise. This leaves 21 English, 21 German, seven Spanish and six Greek speakers. The English speakers were from London or south-east England, and spoke London English or Standard Southern British English. The Spanish speakers were from Aragon, Spain; Santiago, Chile; Mexico City, Mexico; and Buenos Aires, Argentina. The Greek speakers were from Thessalonika, Pagra, Zakynthos, Argos, and Athens in Greece.

In total, there were 4117 usable utterances from 55 speakers, with 283 utterances discarded due to background noise. 7798 sets of MFCCs were able to be calculated from the 8234 vowels. No MFCCs were calculated if the detected vowel length was too short, either inherently or because SFS was unable to align a vowel transcription with the full duration of the vowel.

6.4.5 Results

Nearest neighbour

Using the ACCDIST results, the closest other speaker to each participant is found in Table 6.2. For no language was every single speaker closest to another speaker of that same language.

Looking at these speakers individually (see Figure 6.5), German speaker deu2 is fairly dissimilar to almost all speakers, including Spanish speaker spa1 who is their nearest neighbour. German speaker deu18 is very similar to many other speakers, and is the nearest neighbour of six

[2] Whilst I have made the experimental code available for future use, audio recording with Psychopy is presently unreliable and highly platform dependent.

TABLE 6.2: Nearest neighbour

Their closest match	German	Greek	English	Spanish
Language of speaker				
German	20	0	0	1
Greek	1	2	1	2
English	7	0	14	0
Spanish	0	1	0	6

German speakers, one Greek speaker ell1, and all seven English speakers whose nearest neighbour was not English. Likewise, Greek speaker ell6 was the nearest neighbour to two other Greek speakers and to Spanish speaker spa5; Spanish speaker spa2 was the nearest neighbour to three other Spanish speakers and to Greek speaker ell4; and English speaker eng8 was the nearest neighbour to three other English speakers and to Greek speaker ell3.

Analysis

Applying an Analysis of Variance to the correlation between speakers that was calculated using the ACCDIST method, there was a significant effect of both gender and language interaction. The size of the gender effect was negligible ($\eta^2 = 0.006$), but the language interaction was large ($\eta^2 = 0.146$).

Within-language correlations were significantly larger than between-language correlations, with two exceptions. Firstly, the correlations between Greek and Spanish speakers were statistically indistinguishable from the correlations between Greek colinguals, between Spanish colinguals, or between German colinguals. Secondly, English speakers did not form a homogenous group, with correlation between English speakers being significantly lower than other colingual groups, and than Greek-Spanish.

By contrast to the pilot study, these results show a measurable difference between monophthongal five-vowel systems (Greek, Spanish) and larger vowel systems with diphthongs (German, English). However, German and English are not significantly more correlated than any other pairing. As in the pilot, English is less internally similar than the other three languages, despite much stricter dialectal requirements. This implies that there is less consensus among speakers as to how to adapt non-native vowels to the English vowel system than for the other

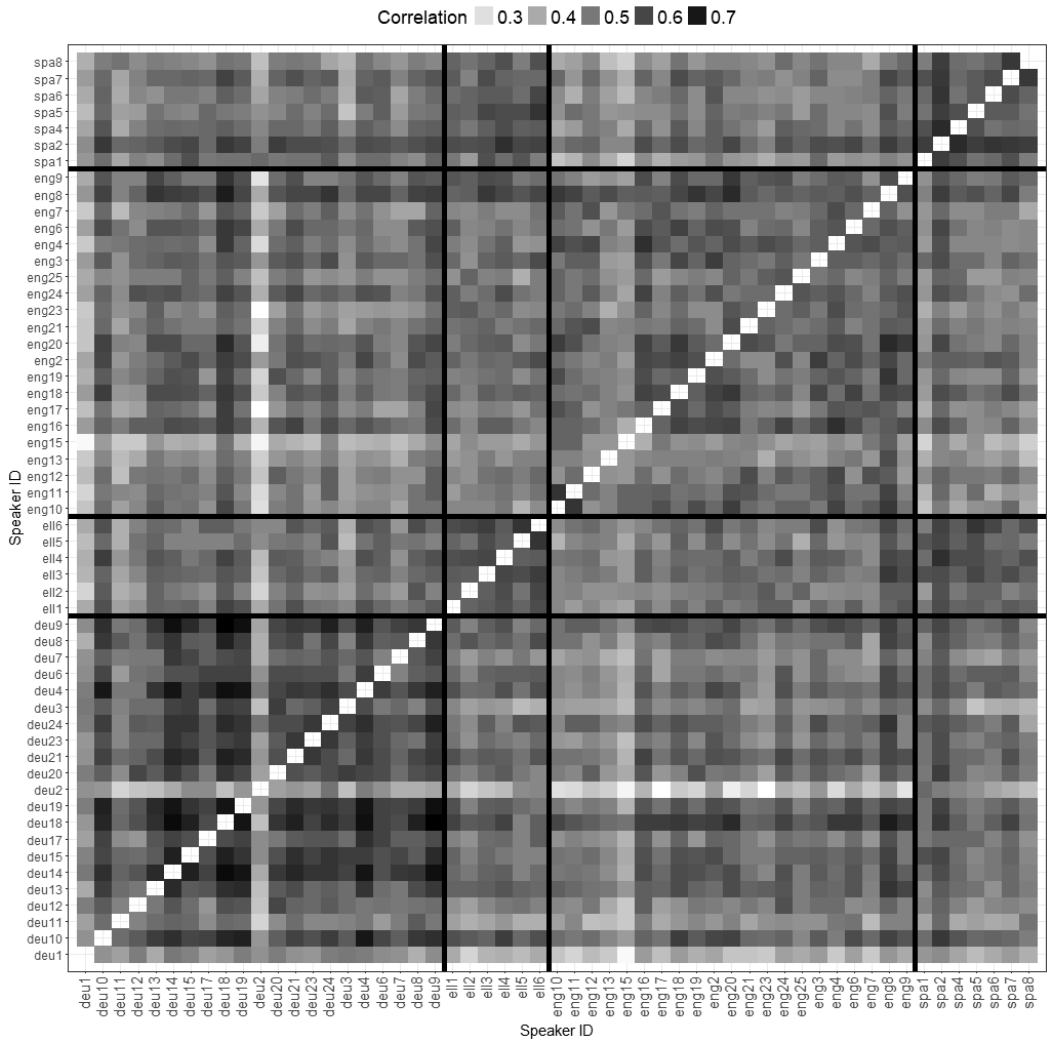


FIGURE 6.5: ACCDIST correlations between individual speakers

TABLE 6.3: Average correlation between speakers by language

Languages		Mean	Standard deviation
Greek	Greek	0.575	0.033
Spanish	Spanish	0.572	0.050
German	German	0.560	0.085
Greek	Spanish	0.554	0.044
English	English	0.522	0.055
German	Spanish	0.503	0.059
German	Greek	0.500	0.064
Greek	English	0.498	0.050
German	English	0.493	0.078
English	Spanish	0.480	0.061

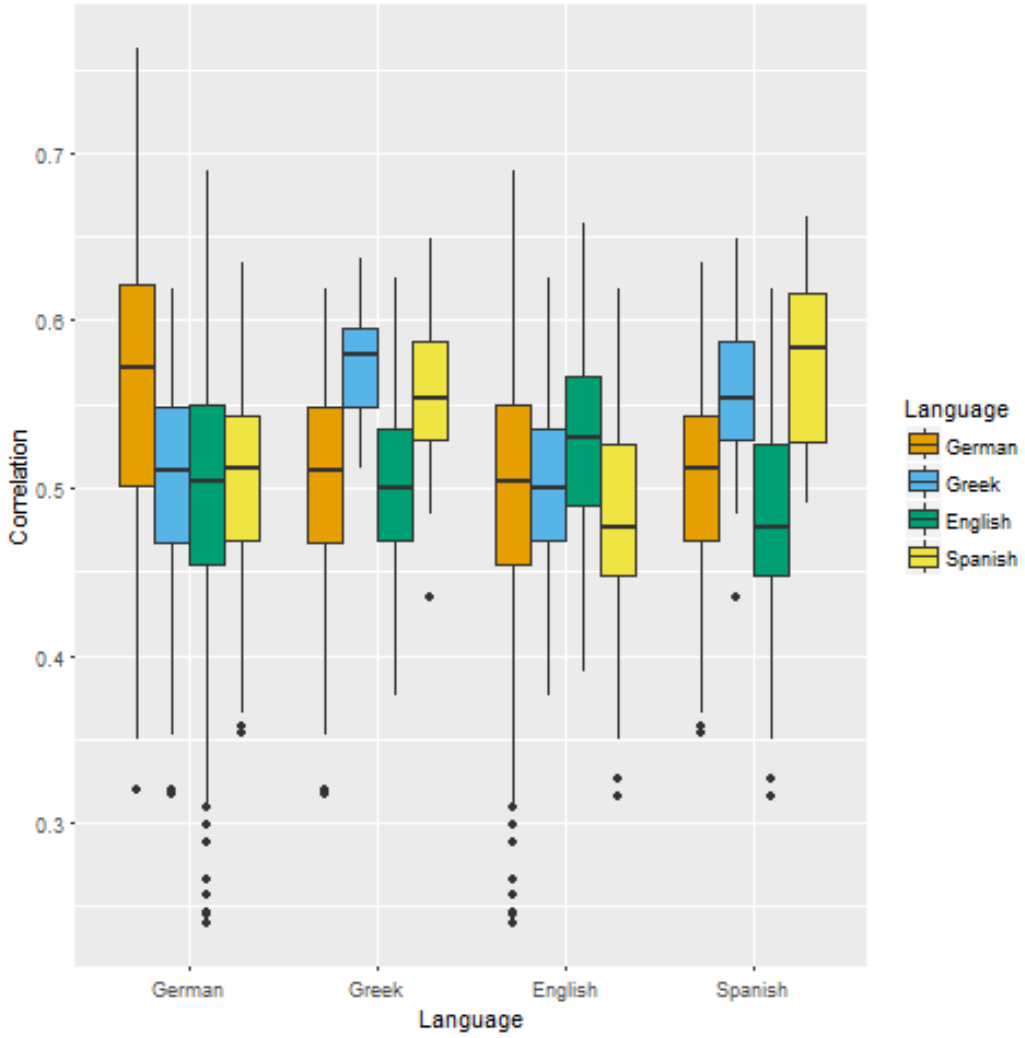


FIGURE 6.6: Correlation between individual speakers, labelled by language background

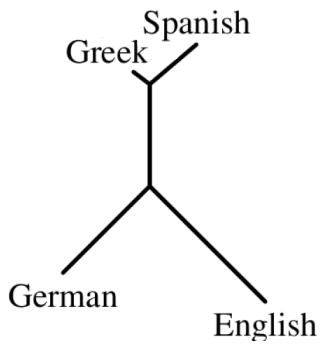


FIGURE 6.7: Visualisation of language distances. (Dereeper et al., 2008, Felsenstein, 1989)

languages. This is an interesting difference when compared to German speakers, who have a similarly wide range of options to choose from, yet are more consistent.

Vowel similarity

To illuminate the origins of these language distances, I shall discuss the observed nativisations by each language group.

Figure 6.8 - Figure 6.11 illustrate the distribution of vowels by speakers of each language. For each of the 160 vowel instances in the stimuli, the mean MFCC values were calculated across all speakers of the same language. The distance between these 'average vowels' was calculated as the sum of squares, as described above. In the following figures, similarity is given as the inverse of the mean distance between average vowels. The label assigned to each production is that of the stimulus.

In all four languages, [a] is produced fairly similarly regardless of whether the stimulus was nasalised, or creaky, breathy or modal voiced. Other notable features include the tense-lax distinction in front mid vowels, which is visible in German and English and completely lacking in Greek and Spanish; the distinction between [i] and [y] in German which is missing from the other languages; and the similarity between [œ] and [ə] in English and German, which is less evident in Greek and Spanish. Not captured in this vowel data, Greek and Spanish speakers produced almost all instances of [œ] with a following rhotic.

6.4.6 Conclusion

The ACCDIST metric can be used to identify the vowel patterns of German speakers in contrast to speakers of languages with five vowel systems, but English speakers are sufficiently diverse in their nativisation strategies that they cannot be identified as a homogenous group, distinct from the other language groups.

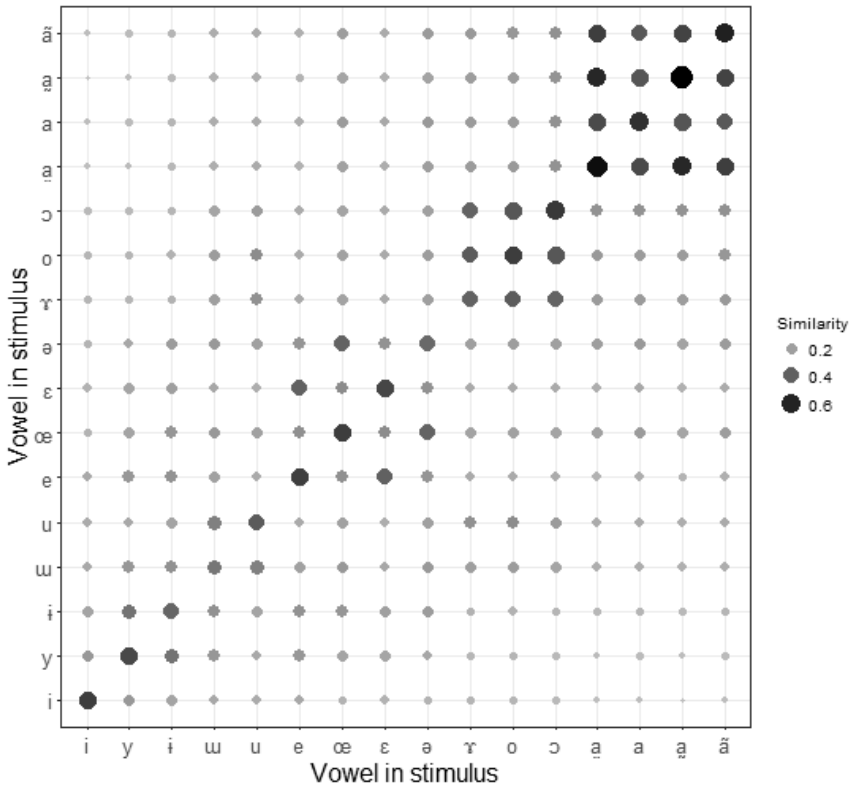


FIGURE 6.8: Mean similarity between vowels produced by German speakers

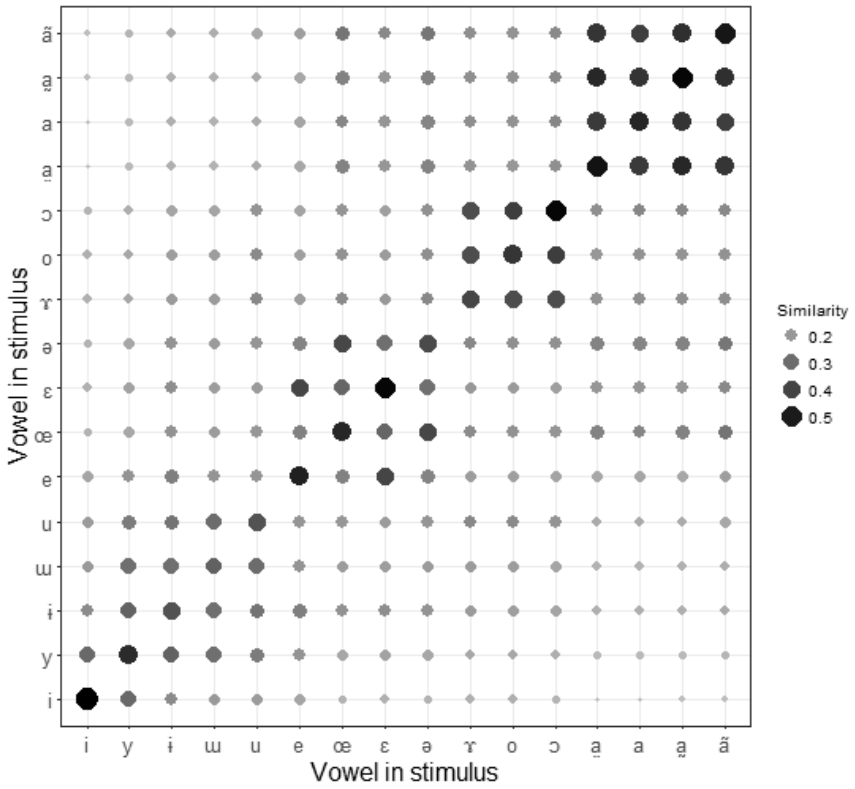


FIGURE 6.9: Mean similarity between vowels produced by English speakers

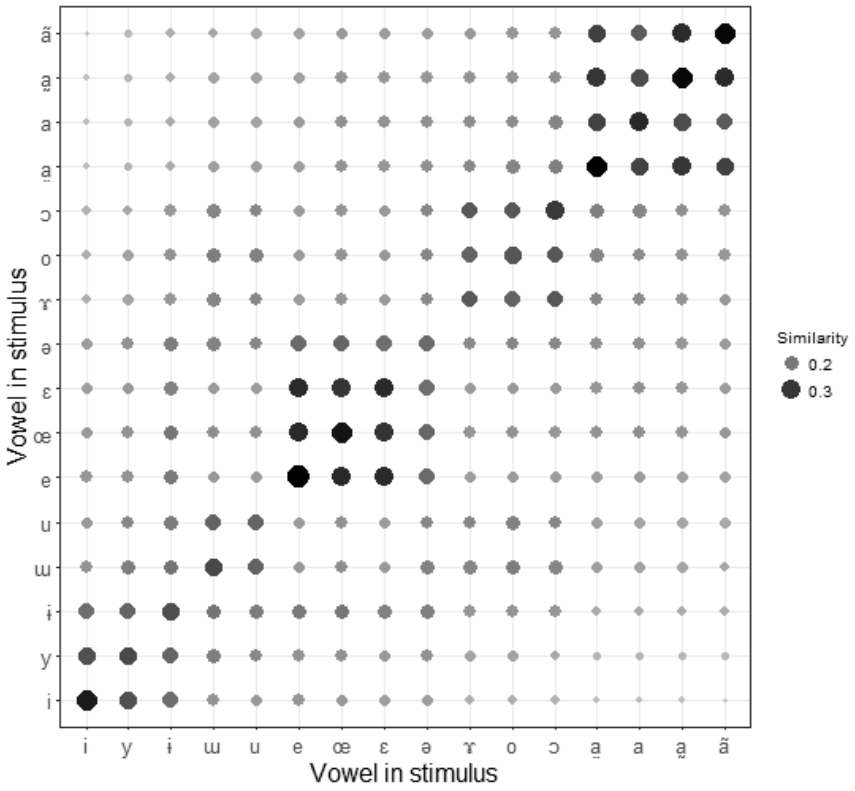


FIGURE 6.10: Mean similarity between vowels produced by Greek speakers

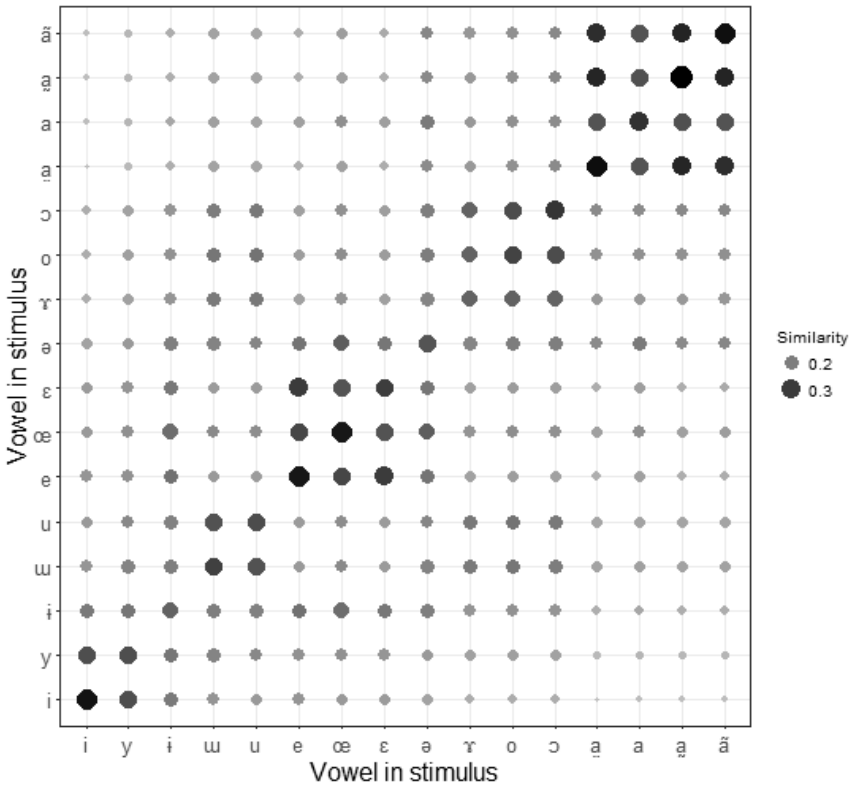


FIGURE 6.11: Mean similarity between vowels produced by Spanish speakers

6.5 Conclusion

Measuring the similarity of speaker's accents in their L2 is possible using ACCDIST, but the results of using read text are too dependent on orthographic effects and on speaker proficiency to give a consistent picture of their L1. Using audio stimuli and non-word adaptation instead removed these effects, but made data acquisition more difficult.

ACCDIST produces the average correlation between individual vowel stimuli across participants of a given language background, and the consistency of adaptation between different speakers. German, Greek and Spanish speakers all adapt vowels predictably depending on their language background, but SSBE English speakers behave more variably. This makes the results of the ACCDIST metric unsuitable as a metric of distance between all vowel systems, since there is no single English system to measure from.

This study could be expanded to use more speakers, both of the existing dialects and of more languages and more dialects, to establish if there is a replicable difference between SSBE English speakers and speakers of other backgrounds in their consistency of adaptation, or if monolingual speakers of prestige dialects of other backgrounds are similarly variable in their treatment of novel loan items.

Chapter 7

Comparison

In this chapter I will compare the three approaches detailed in the preceding chapters. In Section 7.1, I compare the data and analysis required to implement each of the methods. In Section 7.2, I compare the internal consistency of each of the methods. In Section 7.3, I look at the language distances given by each of the methods, how consistent they are between methods, and how they correspond to genetic similarity and linguists' intuitions.

7.1 Comparison of requirements

For all three approaches, it is important that all input data used to calculate the metric is of the same quantity and quality, regardless of which language it is from, in order to produce consistent results. If a metric is to be useful when applied to all languages, as opposed to just Indo-European languages for which we have a wealth of acoustic data, orthographic data and pre-existing analyses, the data requirements must not be too onerous.

The parametric approach and the entropy approach both rely on data which is routinely produced as part of the initial documentation of a language.

Nidaba takes a transcribed lexicon as its input, and provides a suite of tools to aid in producing a phonemically transcribed version. The set of parameters outlined in Section 4.8 includes both inventory and phonotactic characteristics of the lexicon, both expected parts of an initial documentation. However, the diagnostic criteria used must be identical for all languages, which is why it is not sufficient simply to take an existing grammar and assume its analysis is adequate.

The entropy approach requires fewer than a thousand phonemes of training text to produce

a model that is applicable to other languages. To calculate language distance to an already-modelled language (but not vice-versa) requires only a test string, which can be much shorter. I found that a string of length 34 phonemes was sufficient to correctly identify a language from out of seven options, with accuracy increasing asymptotically with length.

The entropy approach can be applied to broad phonetically transcribed texts directly, or programmatically using a statically mapped featural representation (e.g. Hayes, 2008), neither of which require an in-depth analysis of the language of transcription. More consideration of phonological behaviour is required for both language-specific featural representations and element representations. In these latter cases, producing a metric is more time-consuming and difficult, but the results provide more insight into representational theories. If such underlying representations are a more accurate depiction of phonological systems than surface representations, they will also produce a more accurate metric.

Both the parametric approach and the entropy approach rely on constructed records of human speech: on phonological analysis and transcription. They can therefore be applied to historical data, to the reconstruction of a phonological system or the reconstruction of the pronunciation of a written text. By contrast, the ACCDIST approach requires audio recordings of native speakers.

This experimental approach requires targeted recordings: sounds recorded deliberately for this purpose. The data collected may be interesting for other reasons – e.g. relative consistency in loan item adaption – but it is not going to be produced spontaneously, nor collected for any other purpose. There exist alternative spoken language identificiation techniques which do not require particular input data. However, most use training data from all input languages simultaneously, and so produce results relative to the input languages used, rather than an absolute distance. The three approaches I have described in this thesis are applicable to new languages with no alteration; adding new languages does not change the distances measured between pre-existing language pairs.

7.2 Comparison of internal consistency

In this section, I compare the internal consistency and precision of the three approaches. If a metric is accurate, it will consistently produce the same distance when presented with a given

language pair. If it is precise, it will consistently rank two pairs of languages which have very similar distances.

The parameter-based approach relies on a single set of values for each language, so it is not inherently variable. However, inconsistencies may arise since establishing those values is subject to researcher fallibility. Firstly, if a lexicon is unrepresentative of the language it is drawn from, other lexicons may produce different parameter values. This can be mitigated by the inclusion of frequency data, but this is not available for the under-documented languages which are most likely to have short and potentially unrepresentative lexicons, and for which errors are least likely to be caught by peer-review. Secondly, marginal items may be treated inconsistently between languages, being permitted to influence a parameter-value in some cases and not in others. Finally, a user who has specialist knowledge of particular phenomena in one language but not another may selectively deviate from diagnostic criteria. Nidaba contains several tools to mitigate the influence of user variability by automating certain processes, but relying on these to the exclusion of expert knowledge would remove an important verification step.

The resolution of the parameter-based metric is dependent on the number of parameters applicable to a given language pair. The language pair with the smallest number in my sample had 41 applicable parameters, so the metric has a precision of 0.025, and can distinguish between 41 distances. Since no language pairs in my sample are antithetical - something that would be highly unlikely to occur by chance even including thousands of languages - the range of distances observed is 0.06-0.40. This corresponds to approximately 13 distinct categories of language distance. Increasing the number of parameters would increase the precision of the metric.

The entropy-based approach requires transcribed texts to act as exemplars of the language; one to train a model, and one to test against. The accuracy of the metric therefore depends on how representative these texts are of the language as a whole. The results presented here used translations of a single text for all languages to eliminate confounds such as author- or genre-based variations in entropy. In future, it would be good to repeat the calculations using a variety of source texts, to examine the impact this has on entropy-based language distance metrics.

The results were cross-validated, by repeating the same calculation of Kullback-Leibler divergence on multiple sample texts. For all four representational approaches examined, the variation observed between repetitions had a magnitude below 13% of the range of language

distances calculated (see Subsection 5.6.9). Unlike the parameter-based approach, it is therefore not possible to consistently rank up to 41 distinct language distances (which would require a precision of $\pm 1.25\%$), nor ever the 21 language pairs used in the entropy calculations (requiring $< \pm 2.5\%$). Instead, it is possible to consistently divide language pairs into five non-overlapping groups using the entropy approach, regardless of which of the four transcription methods is used. With only seven languages under examination, it is quite possible that there exist language pairs with greater, or even lesser, language distance between them than we have seen here. In that case, the number of non-overlapping groups would increase. However, since Kullback-Leilber divergence has a fixed normalisation, extending the observed values for the metric would not alter the existing values, and the 21 language pairs examined here will never have fully distinguishable distances using this metric with the transcription systems described. It is possible that the precision and reliability of the metric could be improved with different representational choices, or with more advanced entropic calculations.

The ACCDIST approach does not have high internal consistency. As with the entropy-based metrics, altering the source data for a language can alter the resulting language distance. However, the entropy-based metric successfully established a minimum data requirement, above which a language could be reliably identified. This is not the case for the ACCDIST approach, where five of the 21 English speakers were more similar to Greek speakers than to their colinguals.

The ACCDIST metric has a resolution of only three statistically distinct language distance categories: 'colingual', 'similar', and 'dissimilar'. Looking at the six non-colingual language pairs that all three approaches have in common, this is the same resolution as three of the four entropy-based metrics. However, these all include German-English in the 'similar' category along with Greek-Spanish, which ACCDIST does not (see Table 7.1). By contrast, the entropy metric depending on language-specific binary features divides the language pairs not into two, but into three categories: Greek-Spanish is the closest, followed by Greek-English, with German-English having a comparable distance to German-Greek or Spanish-English. Finally, the parameter-based approach sorts all six language pairs into distinct categories: German-English is closest, followed by Greek-German, then Greek-Spanish, Greek-English, Spanish-English, and Spanish-German.

TABLE 7.1: Categorisation of English (Eng.), German, Greek and Spanish (Spa.) by different metrics

ACCDIST	Entropy: IPA	Entropy: static	Entropy: language-specific	Entropy: Elements	Parameters
Greek-Spa.	Greek-Spa. German-Eng.	Greek-Spa. German-Eng.	Greek-Spa.	Greek-Spa. German-Eng.	German-Eng.
			Greek-Eng.		Greek-German
					Greek-Spa.
German-Spa. German-Greek Greek-Eng. German-Eng. Eng.-Spa.	Greek-Eng. German-Greek Eng.-Spa. German-Spa.	Greek-Eng. German-Greek Eng.-Spa. German-Spa.	German-Eng. German-Greek Eng.-Spa.	Greek-Eng. German-Greek Eng.-Spa. German-Spa.	Greek-Eng.
			German-Spa.		Spa.-Eng.
					German-Spa.

7.3 Comparison of language distances

7.3.1 Correlation between metrics

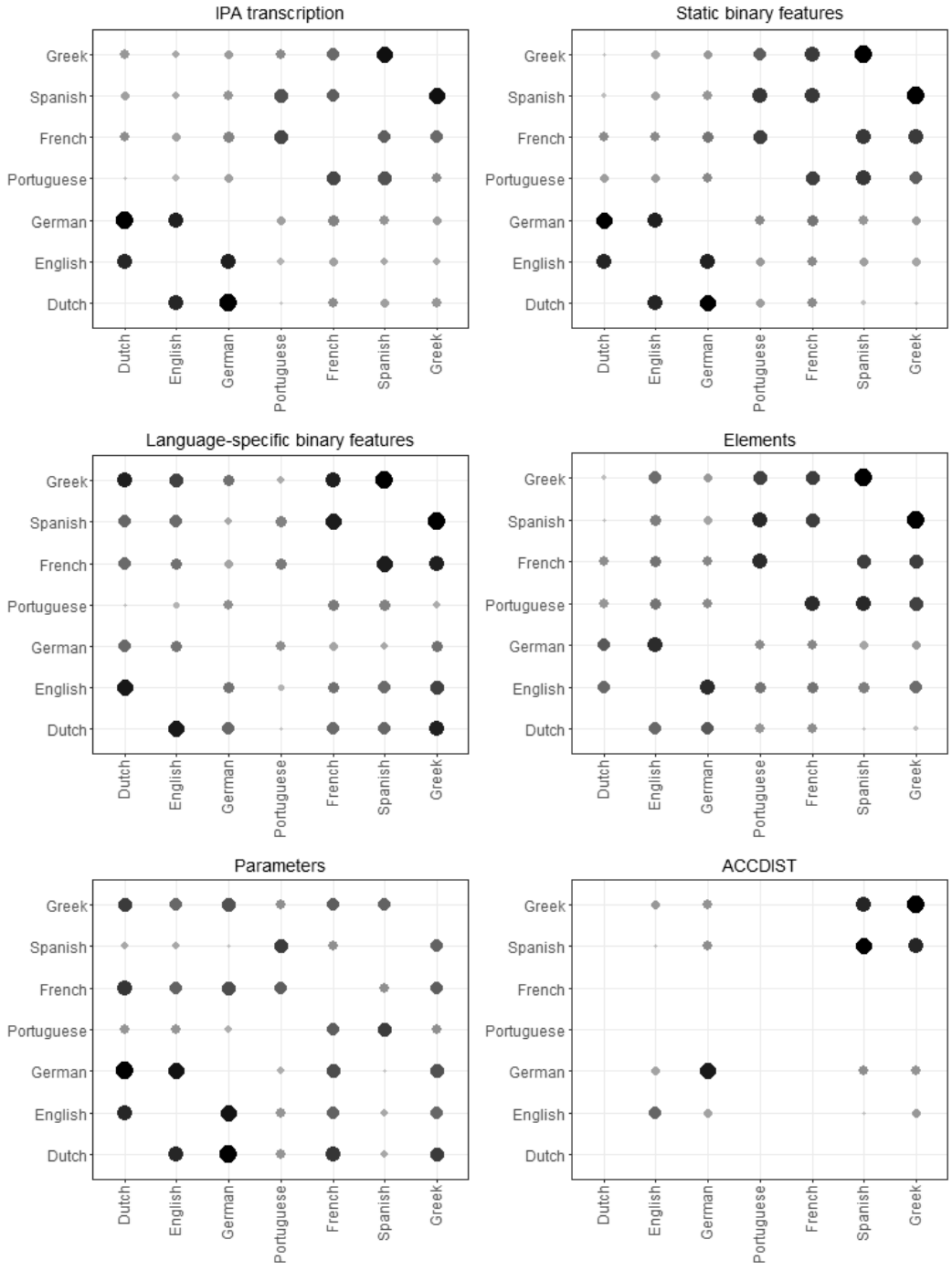
Moving on from internal consistency, we can now ask: how similar are the results of the different metrics to each other?

Table 7.2 shows the Pearson correlation between all six metrics. ACCDIST is included in the table for completeness, but only has six data points to the others' 21, and has been discussed above.

Figure 7.1 comprises six heatmaps showing the relative similarity between languages produced by the parametric Hamming distance, by the mean Kullback-Leibler divergence of each of the four different representational approaches, and by ACCDIST. It includes the 21 language pairs for which the Kullback-Leibler calculations were performed.

The strongest correlation is, unsurprisingly, between the IPA-representation entropy-based metric and the static binary features-representation entropy-based metric. The binary features map directly onto the IPA, and entropy was calculated from abstract segments which therefore closely correspond between the two.

FIGURE 7.1: Similarity between language pairs for each approach, scaled for optimal visualisation. Larger, blacker points are more similar; smaller, greyer points are less similar.



	Parameter	Entropy: IPA	Entropy: static	Entropy: language-specific	Entropy: Elements	ACCDIST
Parameter		0.67	0.55	0.33	0.31	0.10
Entropy: IPA	0.67		0.94	0.46	0.70	0.64
Entropy: static binary features	0.55	0.94		0.38	0.85	0.69
Entropy: language-specific binary features	0.33	0.46	0.38		0.31	0.74
Entropy: Elements	0.31	0.70	0.85	0.31		0.67
ACCDIST	0.10	0.64	0.69	0.74	0.67	

TABLE 7.2: Pearson correlation between all six metrics.

The parameter-based metric is strongly correlated with both of these, as is the element representation entropy-based metric. However, the parameter- and element-based metrics only correlate weakly. The parameter-based metric has stronger similarities between French and Germanic (Dutch, English and German), and between Greek and Germanic, than the other three metrics. The element-representation entropy-based metric has fewer similarities between Germanic languages (see below), and more similarities between Romance languages in comparison to the other three metrics.

The final metric, entropy-based using language-specific binary features, correlates weakly with all the other metrics, excluding ACCDIST. It has weak correlation between German and Dutch/English; strong correlation between French and Spanish/Greek; and strong correlation between Dutch and Greek/French.

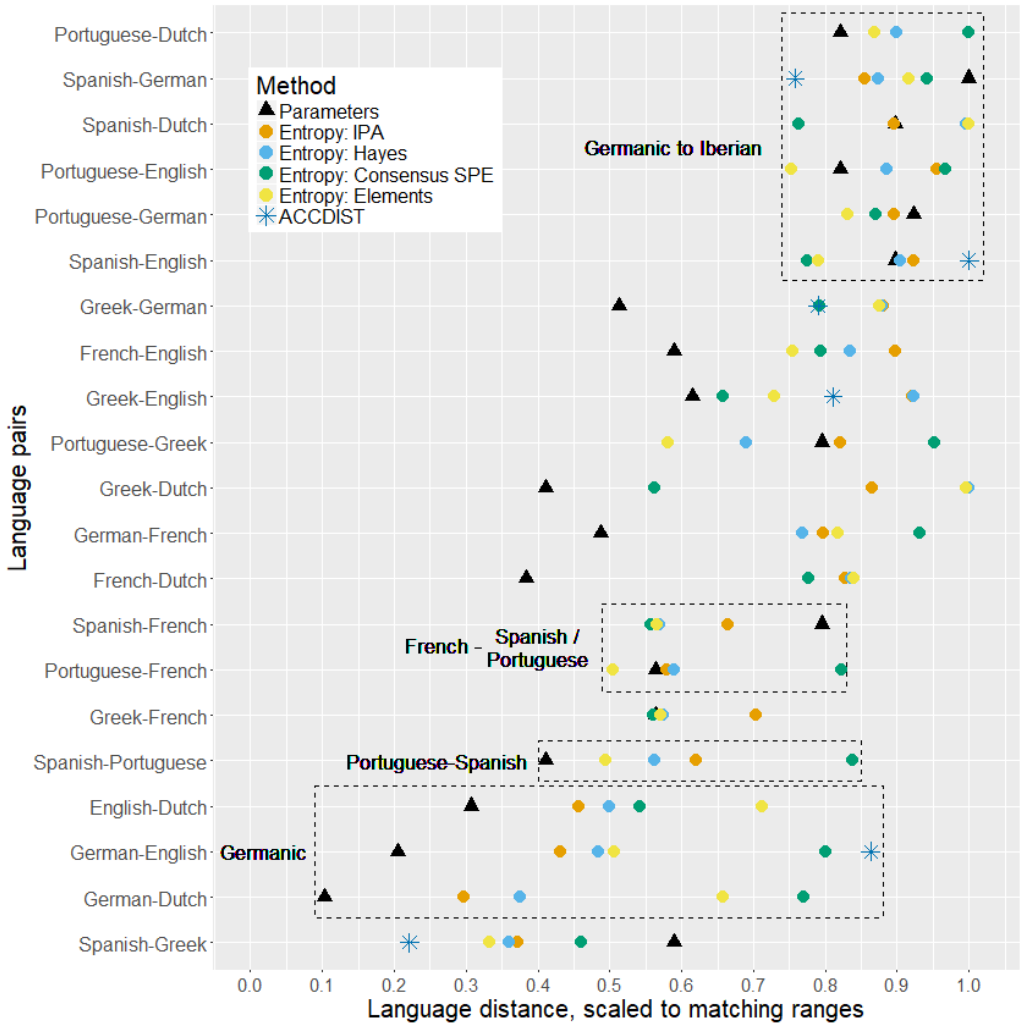
7.3.2 Overview of language-pair distances

Figure 7.2 is an alternative visualisation of language-pair distances. There are three main groupings visible: Germanic pairs, Romance pairs, and Germanic to Iberian.

Greek-Spanish has a small language distance using all metrics, especially the vowel based ACCDIST metric.

The Germanic languages are similar to each other, using the parameter-based metric and the IPA-based and static binary features-based entropy metrics. Dutch is dissimilar to English and German using the element-based entropy metric, due to the relative importance of voicing

FIGURE 7.2: Mean distances between language pairs, using each of the six metrics



contrasts in determining |H| and |L| patterns. German is dissimilar to Dutch and English according to the language-specific binary features-based entropy metric.

The Romance languages all have middling language distances between them, the same or closer than the Germanic languages according to the element and language specific features entropy-based metrics, but more dissimilar according to the parametric and remaining entropy-based metrics. French-Greek also fits this description.

Dutch-Greek has a very large range of language distances. The static features and element entropy-based metrics assign this pair the maximum observed distance, whilst the language-specific featural entropy metric assigns it the same low distance as it assigns Dutch-English, and the parameter-based metric assigns it almost the same distance as Spanish-Greek.

Finally, the Germanic languages are distant from the Iberian Romance languages (Portuguese and Spanish) for all metrics.

7.4 Conclusion

Both the parameter approach and the entropy approach result in a reliable metric. They rely on data gathered in the preliminary stages of language documentation, and thus easily applied to new languages. The parameter-based metric results in greater precision, but overall I find the entropy approach to be superior. It captures all phonotactic patterns present in the data, not just those in a limited set of parameters; the metric corresponds not just to the abstract sense of phonological distance, but the observable consequences for information transfer, both of similarity and of representational choices.

Chapter 8

Conclusion

At the outset of this thesis, I posed the question: Is it possible to derive a meaningful quantitative measure of phonological similarity between individual languages? Such a measure would allow us to address phonological questions that would benefit from quantitative answers, in areas such as second language acquisition, bilingualism and historical linguistics. Non-quantitative, intuitive answers to such questions only take us so far.

Three independent approaches to measuring cross-language phonological distance have been pursued in this thesis: exploiting phonological typological parameters; measuring the cross-entropy of phonologically transcribed texts (i.e. the relative predictability of a transcribed passage in one language given knowledge of some other language); and measuring the phonetic similarity of non-word nativisations by speakers from different language backgrounds.

Firstly, I presented a set of freely accessible online tools to aid in establishing parametric values for syllable structure and phoneme inventory in different languages. The tools are designed to allow researchers to make differing analytical and observational choices and compare the results. I laid out a case study for the use of these tools in analysing the Indo-Aryan language Sylheti. I then applied the tools to 16 languages from four language families, and used correspondence between the resulting parameter values as a measure of phonological distance. This method produces results broadly in accordance with intuition. For example, it groups Germanic languages together, and groups Greek with Spanish. It can distinguish distances to the nearest 2.5%. The tools are designed to be extensible, so that alternative transcription systems or parametric criteria can be incorporated in future, and an alternative metric produced.

Secondly, I applied the computational technique of cross-entropy measurement to texts from seven languages, transcribed in four different ways: a phonemic IPA transcription; with

elements; and with two sets of binary distinctive features in the SPE tradition. This technique results in consistently replicable rankings of phonological similarity for each transcription system, which broadly correlate with the findings of the parameter-based metric. It is sensitive to differences in transcription systems. It can be used to probe the consequences for information transfer of the choices made in devising a representational system. That is, how inclusion or exclusion of certain contrasts affect the amount and predictability of data transferred between speaker and listener. In future, this technique could be extended to more languages; to alternative representations or implementations of these four representations; and to a variety of genres and examples of source texts.

Thirdly, I presented a set of phonetic studies to act as a control for the findings of the other two approaches. Participants from different language backgrounds were presented aurally with non-words covering the vowel space, and asked to nativise them. The accent distance metric ACCDIST was applied to the resulting words. A profile of how each speaker's productions cluster in the vowel space was produced, and ACCDIST measured the similarity of these profiles. Averaging across speakers with a shared native language produced a measure of similarity between language profiles. This technique had mixed success, with English speakers nativising inconsistently, so that there was no coherent language profile to compare with the other three languages. Whilst this is an interesting case study of nativisation behaviour, it is less internally consistent than the two approaches outlined above. A better control in future may come from advances in spoken language identification systems. Many do not model individual languages in such a way that they can be compared, and all are less phonologically transparent than ACCDIST, but systems such as Gelly and Gauvain (2017) are internally consistent.

Both the parameter-based approach and the entropy-based approach deliver a quantitative measure of phonological similarity between individual languages. They are each sensitive to different analytical choices, and require different types and quantities of input data, and so can complement each other. This thesis provides a proof-of-concept for methods which are both internally consistent and falsifiable.

Appendix A

Entropy

A.1 Feature criteria

(1) **sonorant**

[±sonorant] is determined by air pressure: if air flows freely, such that pressure is equalised, a segment is [+sonorant]. If a constriction results in a pressure differential, that segment is [-sonorant]. Vowels, glides, liquids and nasals are [+sonorant], whilst plosives, fricatives, affricates, implosives, and clicks are [-sonorant]. Laryngeals (i.e. [h], [ɦ], [ʔ]) are controversial; Gussenhoven and Jacobs (2005) and Hayes (2008) classify them as [-son], since there is a pressure differential. By contrast, Odden (2005) classifies them as [+son], since spontaneous voicing is precluded on different physical grounds from other [-son] segments, i.e. that the constriction is above the glottis (c.f. Stevens and Keyser, 1989). I shall follow Stevens and Keyser (1989), and treat laryngeals as [-son].

Under this definition, [±sonorant] is not language dependent, despite the behaviour of, for example, the French uvular fricative as a sonorant.

(2) **consonantal**

[-consonantal] is defined as having "greater acoustic energy" than [+consonantal] (Hayes, 2008); this includes vowels and glides, but not liquids, nasals and obstruents. Laryngeals are also [-consonantal], as they have no superlaryngeal constriction. [±consonantal] is not language dependent; the criteria are the same as those used to decide on a transcription.

(3) **continuant**

Sounds involving a full closure in the oral cavity, such that airflow is blocked, are [-continuant]. Plosives, nasals, affricates, implosives, clicks are [-continuant]. Vowels, approximants and fricatives (including [h]) are [+continuant]. Lateral vary, having a central blockage but lateral airflow. Likewise, taps and trills have only a brief closure, and are variably classified. For a full discussion of the issues with [continuant] as a feature, see Mielke (2005).

(4) **voice**

[+voice] sounds are as having vibration of the vocal folds. As with all features, this is specified categorically as a segment property; I am not including phrase-final devoicing or other gradient effects. As an alternative input to the cross-entropy process, it would be possible to record or sample vocal fold vibration and thereby use a continuous or discrete account of the physical effect, without phonological abstraction.

All three sources use this articulatory definition, but then transcribe English with [b, d, g], i.e. symbols specified as [+voice], despite the lack of vocal fold vibration in initial stops in English. I shall follow the 'laryngeal realism' analysis instead (Honeybone, 2005), and assign [+voice] only if there is an active voicing contrast for a segment series in a given language.

(5) **constricted glottis**

[+constricted glottis] sounds are produced with tension in the vocal folds, constricting them. These include glottalised (including ejective), laryngealised or implosive sounds.

(6) **spread glottis**

[+spread glottis] sounds are articulated with spread vocal folds, resulting in audible frication noise. Examples include aspirated obstruents and breathy sonorants, as well as [h, ɦ].

(7) **coronal**

Coronal sounds are articulated with tip or blade. Includes dental, alveolar, alveo-palatal, palato-alveolar, palatal, and retroflex.

(8) **anterior**

[+anterior] sounds are articulated in front of or at the alveolar ridge. This features is only applicable to coronals. Gussenhoven and Jacobs (2013) have anterior as a subnode of coronal; Odden (2005) and Hayes (2008) also extend it to labials, but this does not give any additional contrasts.

(9) **distributed**

[±distributed] is a subfeature of [+coronal]: sounds articulated with the tip of the tongue are [-dist], those with the blade are [+dist]. Dentals and interdentals are [+dist] because the blade contributes. However, this contrast is redundant, given that English - the language with the relevant contrast - also contrasts interdental and apical fricatives using stridency.

(10) **strident**

Stridency is a relative property, with [+strident] segments having more turbulance than their [-strident] counterparts.

(11) **lateral**

[±lateral] is determined by whether air escapes the oral cavity laterally. It is only specified where it is contrastive so most sounds are underspecified.

(12) **nasal**

[±nasal] is determined by whether the velum is raised or lowered, and therefore whether there is airflow through the nasal cavity.

(13) **labial**

[+labial] sounds are articulated with the lips.

(14) **round**

There are no contrasts between [+round] and [-round] labial segments in the languages in my sample, so [±round] is unspecified for all segments.

(15) **back**

As in Section Subsection 4.8.4, I am following Odden (2005) and Gussenhoven and Jacobs (2005) in only including a single parameter [back], rather than both [front] and [back].

This feature applies to vowels and to consonants articulated with the tongue body, i.e. velars, uvulars and pharyngeals, as [high] and [low] do.

[+back] is defined as the bunch of the tongue being relatively back. Back and central vowels are both [+back], as are non-fronted velars and uvulars.

(16) **high**

(17) **low**

[±high] and [±low] are relative properties, based on contrasts between vowels. According to Kostakis (2017), a mid-vowel may be specified as neither high nor low (the traditional representation), or as simultaneously high and low. The choice of representation for a given language depends, as expected, on evidence from synchronic and diachronic processes which refer to these features. In the absence of such evidence for some of the languages in this sample, I have used the traditional specification for all languages, for consistency.

(18) **tense**

In complementary distribution with [ATR] as a feature (i.e. no language has both); labelled tense here as the languages under examination are Indo-European.

A.2 Redundant natural class descriptions

(1) **Spanish**

The feature set is at Table A.1.

All glides are high, so [-lateral, -consonantal] or [-lateral, +high].

All rounded vowels are back. Rhotics are [+sonorant], [+consonantal] or [+anterior] with [-lateral]; labial consonants are [-sonorant], [+consonantal] or [-continuant] with [+labial]; [l, r] are [-continuant] and [-nasal] with [+sonorant]; and non-labial stops are [-sonorant], [+consonant], [-continuant] with [-labial].

(2) **Greek**

The feature set is at Table A.6.

All rounded vowels are back. Labial stops are [-sonorant], [+consonantal] or [-continuant] with [+labial]; and non-labial stops are [-sonorant], [+consonant], [-continuant] with [-labial].

(3) **French**

The feature set is at Table A.4.

Labial stops are [-sonorant], [+consonantal] or [-continuant] with [+labial]; non-labial stops are [-sonorant], [+consonant], [-continuant] with [-labial]; nasal stops are [+consonantal] or [-continuant] with [+nasal]; nasal vowels are [-consonant] or [-high] with [+nasal]; dental fricatives are [-sonorant] or [+continuant] with [+anterior]; fricatives are [-sonorant] or [+continuant] with [-anterior]; and [l, b, d, g] are [+consonant] or [-continuant] or [+voice] with [-nasal].

(4) **Portuguese**

The feature set is at Table A.7.

There are no low front vowels, low rounded vowels or low nasal vowels; [-back] vowels which are [-high] are also [-low], as are [+nasal] vowels and [+labial] vowels. [+tense] vowels which are [+back] are [+labial]. Obstruents [-sonorant] are all [-continuant] and [-nasal]. Labial stops are [-sonorant], [+consonantal] or [-continuant] with [+labial]; non-labial stops are [-sonorant], [+consonant], [-continuant] with [-labial]. Nasal stops are [+consonantal] or [-continuant] with [+nasal]; oral stops are [+consonantal] or [-continuant] with [-nasal]. Voiceless stops are [-continuant] or [-nasal] with [-voice]. Coronal stops are [-labial] or [-nasal] with

[+coronal]. There is only a rounding contrast in the back vowels; [+labial] sounds which are [-consonant] are also [+back], as are [-labial] sounds which are [-consonant]. [ʎ, ɾ] form the class of continuants which are specified for nasality [-nasal], and also for sonorancy [-sonorant].

(5) **German**

The feature set is at Table A.5.

There are no low front vowels or low rounded vowels; [-back] vowels which are [-high] are [-low], as are [+labial] vowels. All labial sonorants are vocalic; [+labial] sounds which are [+sonorant] are [-consonantal]. All labial consonants are non-coronal. Oral stops can be characterised by any two of [-sonorant], [-continuant] and [-nasal], since fricatives are not specified for nasality. All aspirated consonants are non-nasal stops.

(6) **Dutch**

The feature set is at Table A.2.

The inventory contains the following redundancies: There are no low front vowels or low rounded vowels; [-back] vowels which are [-high] are [-low], as are [+labial] vowels. All labial consonants are non-coronal. All voiceless sounds specified as non-nasals are stops, and vice-versa. All labial sonorants are vocalic. Oral stops can be characterised by any two of [-sonorant], [-continuant] and [-nasal], since fricatives are not specified for nasality. Likewise, voiced oral stops can be characterised by [+voice] and either of [-nasal] and [-continuant].

(7) **English**

The feature set is at Table A.3.

If [+high], [+back] $\iff$ [+labial], and [-back] $\iff$ [-labial]. All rounded sonorants are back vowels. All labials are non-coronal. All coronals are non-labial. Fricatives are the only obstruents specified for anteriority Liquids are can be specified as sonorants which are [-nasal] or [+consonant], since nasals are not specified for sonorancy and vowels are not specified for nasality. Nasals are the only stops specified for anteriority. Oral stops can be specified by any two of [-sonorant],[-continuant] and [-nasal].

A.3 Feature sets

TABLE A.1: Minimally-specified binary features - Spanish

Segment name	sonorant	consonant	continuant	nasal	lateral	strident	labial	round	coronal	anterior	distributed	high	low	back	tense	voice	constricted glottis	spread glottis
a	o	-	o	o	o	o	-	o	o	o	o	-	o	+	o	+	o	o
e	o	-	o	o	o	o	-	o	o	o	o	-	o	-	o	+	o	o
i	o	-	o	o	o	o	-	o	o	o	o	+	o	-	o	+	o	o
j	o	-	o	o	-	o	-	o	o	o	o	+	o	-	o	+	o	o
o	o	-	o	o	o	o	+	o	o	o	o	-	o	+	o	+	o	o
u	o	-	o	o	o	o	+	o	o	o	o	+	o	+	o	+	o	o
w	o	-	o	o	-	o	+	o	o	o	o	+	o	+	o	+	o	o
m	o	+	-	+	o	o	o	o	-	o	o	o	o	o	o	+	o	o
n	o	+	-	+	o	o	o	o	+	+	o	o	o	o	o	+	o	o
ɲ	o	+	-	+	o	o	o	o	+	-	o	o	o	o	o	+	o	o
l	+	+	-	-	+	o	o	o	o	+	o	o	o	o	o	+	o	o
ɾ	+	+	+	o	-	o	o	o	o	+	o	o	o	o	o	+	o	o
ʎ	+	+	+	o	+	o	o	o	o	-	o	o	o	o	o	+	o	o
p	-	+	-	o	o	o	+	o	o	o	o	o	o	o	o	-	o	o
g	-	+	-	-	o	o	-	o	-	o	o	o	o	o	o	+	o	o
k	-	+	-	o	o	o	-	o	-	o	o	o	o	o	o	-	o	o
b	-	+	-	-	o	o	+	o	o	o	o	o	o	o	o	+	o	o
t	-	+	-	o	o	o	-	o	+	+	o	o	o	o	o	-	o	o
d	-	+	-	-	o	o	-	o	+	+	o	o	o	o	o	+	o	o
ʧ	-	+	-	o	o	o	-	o	+	-	o	o	o	o	o	-	o	o
f	-	+	+	o	o	+	o	o	-	o	o	o	o	o	o	-	o	o
x	-	+	+	o	o	-	o	o	-	o	o	o	o	o	o	-	o	o
s	-	+	+	o	o	+	o	o	+	+	o	o	o	o	o	-	o	o
θ	-	+	+	o	o	-	o	o	+	+	o	o	o	o	o	-	o	o
z	-	+	+	o	o	o	o	o	+	+	o	o	o	o	o	+	o	o
ʝ	-	+	+	o	o	o	o	o	+	-	o	o	o	o	o	+	o	o
r	+	+	-	-	-	o	o	o	o	+	o	o	o	o	o	+	o	o

TABLE A.2: Minimally-specified binary features - Dutch

Segment name	sonorant	consonant	continuant	nasal	lateral	strident	labial	round	coronal	anterior	distributed	high	low	back	tense	voice	constricted glottis	spread glottis
w	+	-	+	o	o	o	+	o	o	o	o	+	o	+	o	o	o	o
j	+	-	+	o	o	o	-	o	o	o	o	+	o	-	o	o	o	o
ɹ	+	+	+	o	-	o	o	o	o	o	o	o	o	o	o	o	o	o
l	+	+	o	-	+	o	-	o	o	o	o	o	o	o	o	o	o	o
b	-	+	-	-	o	o	+	o	-	o	o	o	o	o	o	+	o	o
n	o	+	-	+	o	o	-	o	+	o	o	o	o	o	o	o	o	o
ŋ	o	+	-	+	o	o	-	o	-	o	o	o	o	o	o	o	o	o
m	o	+	-	+	o	o	+	o	-	o	o	o	o	o	o	o	o	o
d	-	+	-	-	o	o	-	o	+	o	o	o	o	o	o	+	o	o
h	-	-	+	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
ɣ	-	+	+	o	o	o	-	o	-	o	o	o	o	o	o	+	o	o
x	-	+	+	o	o	o	-	o	-	o	o	o	o	o	o	-	o	o
s	-	+	+	o	o	o	-	o	+	o	o	o	o	o	o	-	o	o
z	-	+	+	o	o	o	-	o	+	o	o	o	o	o	o	+	o	o
v	-	+	+	o	o	o	+	o	-	o	o	o	o	o	o	+	o	o
f	-	+	+	o	o	o	+	o	-	o	o	o	o	o	o	-	o	o
t	-	+	-	-	o	o	-	o	+	o	o	o	o	o	o	-	o	o
k	-	+	-	-	o	o	-	o	-	o	o	o	o	o	o	-	o	o
p	-	+	-	-	o	o	+	o	-	o	o	o	o	o	o	-	o	o
u	+	-	+	o	o	o	+	o	o	o	o	+	o	+	+	o	o	o
uː	+	-	+	o	o	o	+	o	o	o	o	+	o	+	+	o	o	o
ɪ	+	-	+	o	o	o	-	o	o	o	o	+	o	-	-	o	o	o
iː	+	-	+	o	o	o	-	o	o	o	o	+	o	-	+	o	o	o
i	+	-	+	o	o	o	-	o	o	o	o	+	o	-	-	o	o	o
y	+	-	+	o	o	o	+	o	o	o	o	+	o	-	-	o	o	o
ʏ	+	-	+	o	o	o	+	o	o	o	o	+	o	-	-	o	o	o
yː	+	-	+	o	o	o	+	o	o	o	o	+	o	-	+	o	o	o
aː	+	-	+	o	o	o	-	o	o	o	o	-	+	+	+	o	o	o
ə	+	-	+	o	o	o	-	o	o	o	o	-	-	+	+	o	o	o
ɑ	+	-	+	o	o	o	-	o	o	o	o	-	+	+	-	o	o	o
ʌ	+	-	+	o	o	o	-	o	o	o	o	-	-	+	-	o	o	o
ɔ	+	-	+	o	o	o	+	o	o	o	o	-	-	+	-	o	o	o
o	+	-	+	o	o	o	+	o	o	o	o	-	-	+	-	o	o	o
oː	+	-	+	o	o	o	+	o	o	o	o	-	-	+	+	o	o	o
ɛ	+	-	+	o	o	o	-	o	o	o	o	-	-	-	-	o	o	o
eː	+	-	+	o	o	o	-	o	o	o	o	-	-	-	+	o	o	o
œ	+	-	+	o	o	o	+	o	o	o	o	-	-	-	-	o	o	o
ø	+	-	+	o	o	o	+	o	o	o	o	-	-	-	+	o	o	o
øː	+	-	+	o	o	o	+	o	o	o	o	-	-	-	+	o	o	o

TABLE A.3: Minimally-specified binary features - English

Segment name	sonorant	consonant	continuant	nasal	lateral	strident	labial	round	coronal	anterior	distributed	high	low	back	tense	voice	constricted glottis	spread glottis
b	-	+	-	-	o	o	+	o	-	o	o	o	o	o	o	o	o	-
p	-	+	-	-	o	o	+	o	-	o	o	o	o	o	o	o	o	+
t	-	+	-	-	o	o	-	o	+	o	o	o	o	o	o	o	o	+
d	-	+	-	-	o	o	-	o	+	o	o	o	o	o	o	o	o	-
k	-	+	-	-	o	o	-	o	-	o	o	o	o	o	o	o	o	+
g	-	+	-	-	o	o	-	o	-	o	o	o	o	o	o	o	o	-
d͡ʒ	-	+	-	-	o	o	-	o	+	o	o	o	o	o	o	o	o	-
t͡ʃ	-	+	-	-	o	o	-	o	+	o	o	o	o	o	o	o	o	+
v	-	+	+	o	o	o	+	o	-	+	o	o	o	o	o	+	o	o
f	-	+	+	o	o	o	+	o	-	+	o	o	o	o	o	-	o	o
ð	-	+	+	o	o	-	-	o	+	+	o	o	o	o	o	+	o	o
θ	-	+	+	o	o	-	-	o	+	+	o	o	o	o	o	-	o	o
s	-	+	+	o	o	+	-	o	+	+	o	o	o	o	o	-	o	o
z	-	+	+	o	o	+	-	o	+	+	o	o	o	o	o	+	o	o
ʃ	-	+	+	o	o	o	-	o	+	-	o	o	o	o	o	-	o	o
ʒ	-	+	+	o	o	o	-	o	+	-	o	o	o	o	o	+	o	o
h	-	-	+	o	o	o	-	o	-	o	o	o	o	o	o	o	o	o
m	o	+	-	+	o	o	+	o	-	+	o	o	o	o	o	o	o	o
n	o	+	-	+	o	o	-	o	+	+	o	o	o	o	o	o	o	o
ŋ	o	+	-	+	o	o	-	o	-	o	o	o	o	o	o	o	o	o
l	+	+	o	-	+	o	-	o	o	o	o	o	o	o	o	o	o	o
ɹ	+	+	+	-	-	o	-	o	o	o	o	o	o	o	o	o	o	o
w	+	-	+	o	o	o	+	o	o	o	o	+	o	+	+	o	o	o
j	+	-	+	o	o	o	-	o	o	-	o	+	o	-	+	o	o	o
iː	+	-	+	o	o	o	-	o	o	o	o	+	o	-	+	o	o	o
uː	+	-	+	o	o	o	+	o	o	o	o	+	o	+	+	o	o	o
e	+	-	+	o	o	o	-	o	o	o	o	-	-	-	+	o	o	o
a	+	-	+	o	o	o	-	o	o	o	o	-	+	+	+	o	o	o
ɛ	+	-	+	o	o	o	-	o	o	o	o	-	-	-	-	o	o	o
ə	+	-	+	o	o	o	-	o	o	o	o	-	-	+	+	o	o	o
ɔ	+	-	+	o	o	o	+	o	o	o	o	-	-	+	+	o	o	o
ɒ	+	-	+	o	o	o	-	o	o	o	o	-	+	+	-	o	o	o
ɑː	+	-	+	o	o	o	-	o	o	o	o	-	+	+	+	o	o	o
æ	+	-	+	o	o	o	-	o	o	o	o	-	+	-	+	o	o	o
ɔː	+	-	+	o	o	o	+	o	o	o	o	-	-	+	+	o	o	o
ɜː	+	-	+	o	o	o	-	o	o	o	o	-	-	-	+	o	o	o
ɪ	+	-	+	o	o	o	-	o	o	o	o	+	o	-	-	o	o	o
ʊ	+	-	+	o	o	o	+	o	o	o	o	+	o	+	-	o	o	o
ʌ	+	-	+	o	o	o	-	o	o	o	o	-	-	+	-	o	o	o

TABLE A.4: Minimally-specified binary features - French

Segment name	sonorant	consonant	continuant	nasal	lateral	strident	labial	round	coronal	anterior	distributed	high	low	back	tense	voice	constricted glottis	spread glottis
i	o	-	o	-	o	o	-	o	o	o	o	+	o	-	o	o	o	o
j	o	-	o	o	o	o	-	o	o	o	o	+	o	-	o	o	o	o
y	o	-	o	o	o	o	+	o	o	o	o	+	o	-	o	o	o	o
ɥ	o	-	o	o	o	o	+	o	o	o	o	+	o	-	o	o	o	o
ɛ	o	-	o	-	o	o	-	o	o	o	o	-	o	-	-	o	o	o
ɛ̃	o	-	o	+	o	o	-	o	o	o	o	-	o	-	-	o	o	o
e	o	-	o	-	o	o	-	o	o	o	o	-	o	-	+	o	o	o
ø	o	-	o	-	o	o	+	o	o	o	o	-	o	-	+	o	o	o
œ	o	-	o	-	o	o	+	o	o	o	o	-	o	-	-	o	o	o
œ̃	o	-	o	+	o	o	+	o	o	o	o	-	o	-	-	o	o	o
u	o	-	o	o	o	o	+	o	o	o	o	+	o	+	o	o	o	o
w	o	-	o	o	o	o	+	o	o	o	o	+	o	+	o	o	o	o
ɑ̃	o	-	o	+	o	o	-	o	o	o	o	-	o	+	o	o	o	o
ɔ	o	-	o	-	o	o	+	o	o	o	o	-	o	+	-	o	o	o
ɔ̃	o	-	o	+	o	o	+	o	o	o	o	-	o	+	-	o	o	o
a	o	-	o	-	o	o	-	o	o	o	o	-	o	+	o	o	o	o
o	o	-	o	-	o	o	+	o	o	o	o	-	o	+	+	o	o	o
õ	o	-	o	+	o	o	+	o	o	o	o	-	o	+	+	o	o	o
ʁ	+	+	+	o	-	o	o	o	-	o	o	o	o	o	o	+	o	o
m	o	+	-	+	o	o	o	o	-	o	o	o	o	o	o	+	o	o
n	o	+	-	+	o	o	o	o	+	+	o	o	o	o	o	+	o	o
ɲ	o	+	-	+	o	o	o	o	+	-	o	o	o	o	o	+	o	o
l	+	+	-	-	+	o	o	o	o	o	o	o	o	o	o	+	o	o
f	-	+	+	o	o	o	o	o	-	o	o	o	o	o	o	-	o	o
s	-	+	+	o	o	o	o	o	+	+	o	o	o	o	o	-	o	o
z	-	+	+	o	o	o	o	o	+	+	o	o	o	o	o	+	o	o
ʃ	-	+	+	o	o	o	o	o	+	-	o	o	o	o	o	-	o	o
ʒ	-	+	+	o	o	o	o	o	+	-	o	o	o	o	o	+	o	o
v	-	+	+	o	o	o	o	o	-	o	o	o	o	o	o	+	o	o
b	-	+	-	-	o	o	+	o	o	o	o	o	o	o	o	+	o	o
d	-	+	-	-	o	o	-	o	+	o	o	o	o	o	o	+	o	o
g	-	+	-	-	o	o	-	o	-	o	o	o	o	o	o	+	o	o
k	-	+	-	o	o	o	-	o	-	o	o	o	o	o	o	-	o	o
p	-	+	-	o	o	o	+	o	o	o	o	o	o	o	o	-	o	o
t	-	+	-	o	o	o	-	o	+	o	o	o	o	o	o	-	o	o

TABLE A.5: Minimally-specified binary features - German

Segment name	sonorant	consonant	continuant	nasal	lateral	strident	labial	round	coronal	anterior	distributed	high	low	back	tense	voice	constricted glottis	spread glottis
j	+	-	+	o	o	o	-	o	o	o	o	+	o	-	+	o	o	o
iː	+	-	+	o	o	o	-	o	o	o	o	+	o	-	+	o	o	o
ɪ	+	-	+	o	o	o	-	o	o	o	o	+	o	-	-	o	o	o
i	+	-	+	o	o	o	-	o	o	o	o	+	o	-	-	o	o	o
yː	+	-	+	o	o	o	+	o	o	o	o	+	o	-	+	o	o	o
y	+	-	+	o	o	o	+	o	o	o	o	+	o	-	-	o	o	o
ʏ	+	-	+	o	o	o	+	o	o	o	o	+	o	-	-	o	o	o
ɛ	+	-	+	o	o	o	-	o	o	o	o	-	-	-	-	o	o	o
e	+	-	+	o	o	o	-	o	o	o	o	-	-	-	+	o	o	o
eː	+	-	+	o	o	o	-	o	o	o	o	-	-	-	+	o	o	o
ɛː	+	-	+	o	o	o	-	o	o	o	o	-	-	-	+	o	o	o
øː	+	-	+	o	o	o	+	o	o	o	o	-	-	-	+	o	o	o
œ	+	-	+	o	o	o	+	o	o	o	o	-	-	-	-	o	o	o
ø	+	-	+	o	o	o	+	o	o	o	o	-	-	-	+	o	o	o
uː	+	-	+	o	o	o	+	o	o	o	o	+	o	+	+	o	o	o
ʊ	+	-	+	o	o	o	+	o	o	o	o	+	o	+	-	o	o	o
u	+	-	+	o	o	o	+	o	o	o	o	+	o	+	+	o	o	o
ə	+	-	+	o	o	o	-	o	o	o	o	-	-	+	+	o	o	o
ɐ	+	-	+	o	o	o	-	o	o	o	o	-	-	+	-	o	o	o
ɔ	+	-	+	o	o	o	+	o	o	o	o	-	-	+	-	o	o	o
oː	+	-	+	o	o	o	+	o	o	o	o	-	-	+	+	o	o	o
o	+	-	+	o	o	o	+	o	o	o	o	-	-	+	-	o	o	o
a	+	-	+	o	o	o	-	o	o	o	o	-	+	+	-	o	o	o
aː	+	-	+	o	o	o	-	o	o	o	o	-	+	+	+	o	o	o
l	+	+	o	-	+	o	-	o	o	o	o	o	o	o	o	o	o	o
ʁ	+	+	+	o	-	o	o	o	o	o	o	o	o	o	o	o	o	o
ɹ	+	+	+	o	-	o	o	o	o	o	o	o	o	o	o	o	o	o
h	-	-	+	o	o	o	o	o	o	o	o	o	o	o	o	o	o	+
t	-	+	-	-	o	o	-	o	+	o	o	o	o	o	o	o	o	+
d	-	+	-	-	o	o	-	o	+	o	o	o	o	o	o	o	o	-
k	-	+	-	-	o	o	-	o	-	o	o	o	o	o	o	o	o	+
g	-	+	-	-	o	o	-	o	-	o	o	o	o	o	o	o	o	-
t͡s	-	+	-	-	o	o	-	o	+	o	o	o	o	o	o	o	o	o
b	-	+	-	-	o	o	+	o	-	o	o	o	o	o	o	o	o	-
p	-	+	-	-	o	o	+	o	-	o	o	o	o	o	o	o	o	+
p͡f	-	+	-	-	o	o	+	o	-	o	o	o	o	o	o	o	o	o
x	-	+	+	o	o	o	-	o	-	o	o	o	o	o	o	-	o	o
s	-	+	+	o	o	o	-	o	+	+	o	o	o	o	o	-	o	o
z	-	+	+	o	o	o	-	o	+	+	o	o	o	o	o	+	o	o
ʃ	-	+	+	o	o	+	-	o	+	-	o	o	o	o	o	-	o	o
v	-	+	+	o	o	o	+	o	-	o	o	o	o	o	o	+	o	o
f	-	+	+	o	o	o	+	o	-	o	o	o	o	o	o	-	o	o
ç	-	+	+	o	o	-	-	o	+	-	o	o	o	o	o	-	o	o
n	o	+	-	+	o	o	-	o	+	o	o	o	o	o	o	o	o	o
ŋ	o	+	-	+	o	o	-	o	-	o	o	o	o	o	o	o	o	o
m	o	+	-	+	o	o	+	o	-	o	o	o	o	o	o	o	o	o

TABLE A.6: Minimally-specified binary features - Greek

Segment name	sonorant	consonant	continuant	nasal	lateral	strident	labial	round	coronal	anterior	distributed	high	low	back	tense	voice	constricted glottis	spread glottis
a	o	-	o	o	o	o	-	o	o	o	o	-	o	+	o	+	o	o
e	o	-	o	o	o	o	-	o	o	o	o	-	o	-	o	+	o	o
i	o	-	o	o	o	o	-	o	o	o	o	+	o	-	o	+	o	o
o	o	-	o	o	o	o	+	o	o	o	o	-	o	+	o	+	o	o
u	o	-	o	o	o	o	+	o	o	o	o	+	o	+	o	+	o	o
w	o	-	o	o	-	o	+	o	o	o	o	+	o	+	o	+	o	o
m	o	+	-	+	o	o	o	o	-	o	o	o	o	o	o	+	o	o
n	o	+	-	+	o	o	o	o	+	o	o	o	o	o	o	+	o	o
p	-	+	-	o	o	o	+	o	o	o	o	o	o	o	o	-	o	o
b	-	+	-	-	o	o	+	o	o	o	o	o	o	o	o	+	o	o
t	-	+	-	o	o	o	-	o	+	o	o	o	o	o	o	-	o	o
d	-	+	-	-	o	o	-	o	+	o	o	o	o	o	o	+	o	o
k	-	+	-	o	o	o	-	o	-	o	o	o	o	o	o	-	o	o
g	-	+	-	-	o	o	-	o	-	o	o	o	o	o	o	+	o	o
f	-	+	+	o	o	+	o	o	-	o	o	o	o	o	o	-	o	o
v	-	+	+	o	o	+	o	o	-	o	o	o	o	o	o	+	o	o
θ	-	+	+	o	o	-	o	o	+	o	o	o	o	o	o	-	o	o
ð	-	+	+	o	o	-	o	o	+	o	o	o	o	o	o	+	o	o
s	-	+	+	o	o	+	o	o	+	o	o	o	o	o	o	-	o	o
z	-	+	+	o	o	+	o	o	+	o	o	o	o	o	o	+	o	o
x	-	+	+	o	o	-	o	o	-	o	o	o	o	o	o	-	o	o
ɣ	-	+	+	o	o	-	o	o	-	o	o	o	o	o	o	+	o	o
l	+	+	-	-	+	o	o	o	o	o	o	o	o	o	o	+	o	o
r	+	+	-	-	-	o	o	o	o	o	o	o	o	o	o	+	o	o

TABLE A.7: Minimally-specified binary features - Portuguese

Segment name	sonorant	consonant	continuant	nasal	lateral	strident	labial	round	coronal	anterior	distributed	high	low	back	tense	voice	constricted glottis	spread glottis
i	o	-	o	-	o	o	o	o	o	o	o	+	o	-	o	o	o	o
ĩ	o	-	o	+	o	o	o	o	o	o	o	+	o	-	o	o	o	o
j̃	o	-	o	+	o	o	o	o	o	o	o	+	o	-	o	o	o	o
j	o	-	o	-	o	o	o	o	o	o	o	+	o	-	o	o	o	o
e	o	-	o	-	o	o	o	o	o	o	o	-	-	-	+	o	o	o
ẽ	o	-	o	+	o	o	o	o	o	o	o	-	-	-	+	o	o	o
ɛ	o	-	o	-	o	o	o	o	o	o	o	-	-	-	-	o	o	o
ɨ	o	-	o	-	o	o	-	o	o	o	o	+	o	+	o	o	o	o
u	o	-	o	-	o	o	+	o	o	o	o	+	o	+	o	o	o	o
w	o	-	o	-	o	o	+	o	o	o	o	+	o	+	o	o	o	o
w̃	o	-	o	+	o	o	+	o	o	o	o	+	o	+	o	o	o	o
ũ	o	-	o	+	o	o	+	o	o	o	o	+	o	+	o	o	o	o
a	o	-	o	-	o	o	-	o	o	o	o	-	+	+	o	o	o	o
ɐ	o	-	o	-	o	o	-	o	o	o	o	-	-	+	o	o	o	o
ɐ̃	o	-	o	+	o	o	-	o	o	o	o	-	-	+	o	o	o	o
ɘ	o	-	o	-	o	o	-	o	o	o	o	-	-	+	-	o	o	o
ɔ	o	-	o	-	o	o	+	o	o	o	o	-	-	+	-	o	o	o
o	o	-	o	-	o	o	+	o	o	o	o	-	-	+	+	o	o	o
õ	o	-	o	+	o	o	+	o	o	o	o	-	-	+	+	o	o	o
m	o	+	-	+	o	o	o	o	-	o	o	o	o	o	o	o	o	o
ɲ	o	+	-	+	o	o	o	o	+	-	o	o	o	o	o	o	o	o
n	o	+	-	+	o	o	o	o	+	+	o	o	o	o	o	o	o	o
g	-	+	-	-	o	o	-	o	-	o	o	o	o	o	o	+	o	o
k	-	+	-	-	o	o	-	o	-	o	o	o	o	o	o	-	o	o
d	-	+	-	-	o	o	-	o	+	o	o	o	o	o	o	+	o	o
t	-	+	-	-	o	o	-	o	+	o	o	o	o	o	o	-	o	o
b	-	+	-	-	o	o	+	o	o	o	o	o	o	o	o	+	o	o
p	-	+	-	-	o	o	+	o	o	o	o	o	o	o	o	-	o	o
f	-	+	+	o	o	o	o	o	-	o	o	o	o	o	o	-	o	o
v	-	+	+	o	o	o	o	o	-	o	o	o	o	o	o	+	o	o
s	-	+	+	o	o	o	o	o	+	+	o	o	o	o	o	-	o	o
ʃ	-	+	+	o	o	o	o	o	+	-	o	o	o	o	o	-	o	o
z	-	+	+	o	o	o	o	o	+	+	o	o	o	o	o	+	o	o
ʒ	-	+	+	o	o	o	o	o	+	-	o	o	o	o	o	+	o	o
l	+	+	-	-	+	o	o	o	o	o	o	o	o	o	o	+	o	o
ʎ	+	+	+	-	+	o	o	o	o	-	o	o	o	o	o	+	o	o
ʀ	+	+	-	-	-	o	o	o	-	o	o	o	o	o	o	+	o	o
ʁ	+	+	-	-	-	o	o	o	-	o	o	o	o	o	o	+	o	o
ɾ	+	+	+	-	-	o	o	o	o	+	o	o	o	o	o	+	o	o

A.4 Element sets

TABLE A.8: Elements - Spanish

Segment	|A|	|A̲|	|I|	|I̲|	|U|	|U̲|	|H|	|H̲|	|L|	|L̲|	|ʔ|	|ʔ̲|	Vowel
j	-	-	+	+	-	-	-	-	-	-	-	-	-
w	-	-	-	-	+	+	-	-	-	-	-	-	-
ɾ	-	-	+	-	-	-	-	-	-	-	-	-	-
l	+	-	+	-	-	-	-	-	-	-	-	-	-
r	+	+	-	-	-	-	-	-	-	-	-	-	-
ʎ	+	-	+	+	-	-	-	-	-	-	-	-	-
θ	-	-	+	-	-	-	+	-	-	-	-	-	-
x	-	-	-	-	+	-	+	-	-	-	-	-	-
f	+	-	-	-	+	+	+	-	-	-	-	-	-
s	+	-	-	-	-	-	+	-	-	-	-	-	-
ʝ	-	-	+	-	-	-	+	-	+	+	-	-	-
z	+	-	-	-	-	-	+	-	+	+	-	-	-
p	-	-	-	-	+	+	-	-	-	-	+	-	-
k	-	-	-	-	+	-	-	-	-	-	+	-	-
ʧ	-	-	+	+	-	-	-	-	-	-	+	-	-
t	+	-	-	-	-	-	-	-	-	-	+	-	-
m	-	-	-	-	+	+	-	-	+	-	+	-	-
ɲ	-	-	+	+	+	-	-	-	+	-	+	-	-
b	-	-	-	-	+	+	-	-	+	+	+	-	-
g	-	-	-	-	+	-	-	-	+	+	+	-	-
n	+	-	-	-	-	-	-	-	+	-	+	-	-
d	+	-	-	-	-	-	-	-	+	+	+	-	-
i	-	-	+	-	-	-	-	-	-	-	-	-	+
u	-	-	-	-	+	-	-	-	-	-	-	-	+
a	+	-	-	-	-	-	-	-	-	-	-	-	+
e	+	-	+	-	-	-	-	-	-	-	-	-	+
o	+	-	-	-	+	-	-	-	-	-	-	-	+

TABLE A.9: Elements - Dutch

Segment	|A|	|A̲|	|I|	|I̲|	|U|	|U̲|	|H|	|H̲|	|L|	|L̲|	|ʔ|	|ʔ̲|	Vowel	Long
w	-	-	-	-	+	+	-	-	-	-	-	-	-	
j	-	-	+	+	-	-	-	-	-	-	-	-	-	
ɪ	+	-	-	-	-	-	-	-	-	-	-	-	-	
l	+	-	+	-	-	-	-	-	-	-	-	-	-	
b	-	-	-	-	+	+	-	-	+	+	+	-	-	
m	-	-	-	-	+	+	-	-	+	-	+	-	-	
ŋ	-	-	-	-	+	-	-	-	+	-	+	-	-	
d	+	-	-	-	-	-	-	-	+	+	+	-	-	
n	+	-	-	-	-	-	-	-	+	-	+	-	-	
h	-	-	-	-	-	-	+	-	-	-	-	-	-	
ɣ	-	-	-	-	+	-	+	+	+	+	-	-	-	
x	-	-	-	-	+	-	+	+	-	-	-	-	-	
s	-	-	+	-	-	-	+	-	-	-	-	-	-	
z	-	-	+	-	-	-	+	-	+	+	-	-	-	
v	+	-	-	-	+	+	+	-	+	+	-	-	-	
f	+	-	-	-	+	+	+	-	-	-	-	-	-	
p	-	-	-	-	+	+	+	-	-	-	+	-	-	
k	-	-	-	-	+	-	+	-	-	-	+	-	-	
t	+	-	-	-	-	-	+	-	-	-	+	-	-	
u	-	-	-	-	+	-	-	-	-	-	-	-	+	✗
uː	-	-	-	-	+	+	-	-	-	-	-	-	+	✓
ɪ	-	-	+	-	-	-	-	-	-	-	-	-	+	✗
iː	-	-	+	+	-	-	-	-	-	-	-	-	+	✓
i	-	-	+	+	-	-	-	-	-	-	-	-	+	✗
y	-	-	+	-	+	-	-	-	-	-	-	-	+	✓
ʏ	-	-	+	-	+	-	-	-	-	-	-	-	+	✗
yː	-	-	+	-	+	-	-	-	-	-	-	-	+	✓
aː	+	+	-	-	-	-	-	-	-	-	-	-	+	✓
ə	+	-	-	-	-	-	-	-	-	-	-	-	+	✗
ɑ	+	+	-	-	-	-	-	-	-	-	-	-	+	✗
ʌ	+	+	-	-	-	-	-	-	-	-	-	-	+	✗
ɔ	+	+	-	-	+	-	-	-	-	-	-	-	+	✗
o	+	-	-	-	+	-	-	-	-	-	-	-	+	✗
oː	+	-	-	-	+	-	-	-	-	-	-	-	+	✓
ɛ	+	-	+	+	-	-	-	-	-	-	-	-	+	✗
eː	+	-	+	+	-	-	-	-	-	-	-	-	+	✓
œ	+	-	+	-	+	-	-	-	-	-	-	-	+	✗
ø	+	-	+	+	+	-	-	-	-	-	-	-	+	✗
øː	+	-	+	+	+	-	-	-	-	-	-	-	+	✓

Table A.10: Elements - English

Segment	|A|	|A̲|	|I|	|I̲|	|U|	|U̲|	|H|	|H̲|	|L|	|L̲|	|ʔ|	|ʔ̲|	Vowel	Long
b	-	-	-	-	+	+	-	-	-	-	+	-	-	
p	-	-	-	-	+	+	+	+	-	-	+	-	-	
t	+	-	-	-	-	-	+	+	-	-	+	-	-	
d	+	-	-	-	-	-	-	-	-	-	+	-	-	
k	-	-	-	-	+	-	+	+	-	-	+	-	-	
g	-	-	-	-	+	-	-	-	-	-	+	-	-	
d͡ʒ	-	-	+	+	-	-	-	-	-	-	+	-	-	
t͡ʃ	-	-	+	+	-	-	+	+	-	-	+	-	-	
v	+	-	-	-	+	+	+	-	-	-	-	-	-	
f	+	-	-	-	+	+	+	+	-	-	-	-	-	
ð	-	-	+	-	-	-	+	-	-	-	-	-	-	
θ	-	-	+	-	-	-	+	+	-	-	-	-	-	
s	+	-	-	-	-	-	+	+	-	-	-	-	-	
z	+	-	-	-	-	-	+	-	-	-	-	-	-	
ʃ	-	-	+	+	-	-	+	+	-	-	-	-	-	
ʒ	-	-	+	+	-	-	+	-	-	-	-	-	-	
h	-	-	-	-	-	-	+	-	-	-	-	-	-	
m	-	-	-	-	+	+	-	-	+	-	+	-	-	
n	+	-	-	-	-	-	-	-	+	-	+	-	-	
ŋ	-	-	-	-	+	-	-	-	+	-	+	-	-	
l	+	-	+	-	-	-	-	-	-	-	-	-	-	
ɹ	+	-	-	-	-	-	-	-	-	-	-	-	-	
w	-	-	-	-	+	+	-	-	-	-	-	-	-	
j	-	-	+	+	-	-	-	-	-	-	-	-	-	
n̩	+	-	-	-	-	-	-	-	+	-	+	-	+	
l̩	+	-	+	-	-	-	-	-	-	-	-	-	+	
iː	-	-	+	+	-	-	-	-	-	-	-	-	+	✓
uː	-	-	-	-	+	+	-	-	-	-	-	-	+	✓
e	+	-	+	+	-	-	-	-	-	-	-	-	+	✗
a	+	+	-	-	-	-	-	-	-	-	-	-	+	✗
ɛ	+	-	+	+	-	-	-	-	-	-	-	-	+	✗
ə	+	-	-	-	-	-	-	-	-	-	-	-	+	✗
ɔ	+	+	-	-	+	-	-	-	-	-	-	-	+	✗
ɒ	+	+	-	-	+	-	-	-	-	-	-	-	+	✗
ɑː	+	+	-	-	-	-	-	-	-	-	-	-	+	✓
æ	+	+	+	-	-	-	-	-	-	-	-	-	+	✗
ɔː	+	+	-	-	+	-	-	-	-	-	-	-	+	✓
ɜː	+	-	-	-	-	-	-	-	-	-	-	-	+	✓
ɪ	-	-	+	-	-	-	-	-	-	-	-	-	+	✗
ʊ	-	-	-	-	+	-	-	-	-	-	-	-	+	✗
ʌ	+	+	-	-	-	-	-	-	-	-	-	-	+	✗
i	-	-	+	+	-	-	-	-	-	-	-	-	+	✗

TABLE A.11: Elements - French

Segment	|A|	|A̲|	|I|	|I̲|	|U|	|U̲|	|H|	|H̲|	|L|	|L̲|	|ʔ|	|ʔ̲|	Vowel
j	-	-	+	+	-	-	-	-	-	-	-	-	-
l	+	-	+	-	-	-	-	-	-	-	-	-	-
w	-	-	-	-	+	+	-	-	-	-	-	-	-
f	+	-	-	-	+	+	+	-	-	-	-	-	-
s	+	-	-	-	-	-	+	-	-	-	-	-	-
ʃ	-	-	+	+	-	-	+	-	-	-	-	-	-
v	+	-	-	-	+	+	+	-	+	+	-	-	-
z	+	-	-	-	-	-	+	-	+	+	-	-	-
ʒ	-	-	+	+	-	-	+	-	+	+	-	-	-
b	-	-	-	-	+	+	-	-	+	+	+	-	-
d	+	-	-	-	-	-	-	-	+	+	+	-	-
k	-	-	-	-	+	-	-	-	-	-	+	-	-
m	-	-	-	-	+	+	-	-	+	-	+	-	-
n	+	-	-	-	-	-	-	-	+	-	+	-	-
ɲ	-	-	+	+	+	-	-	-	+	-	+	-	-
p	-	-	-	-	+	+	-	-	-	-	+	-	-
t	+	-	-	-	-	-	-	-	-	-	+	-	-
a	+	-	-	-	-	-	-	-	-	-	-	-	+
ɔ	+	+	-	-	+	-	-	-	-	-	-	-	+
e	+	-	+	+	-	-	-	-	-	-	-	-	+
ɛ	+	-	+	-	-	-	-	-	-	-	-	-	+
i	-	-	+	-	-	-	-	-	-	-	-	-	+
o	+	-	-	-	+	-	-	-	-	-	-	-	+
u	-	-	-	-	+	-	-	-	-	-	-	-	+
ɡ	-	-	-	-	+	-	-	-	+	+	+	-	-
ɥ	-	-	+	-	+	-	-	-	-	-	-	-	-
ø	+	-	+	+	+	-	-	-	-	-	-	-	+
œ	+	-	+	-	+	-	-	-	-	-	-	-	+
ʁ	+	-	-	-	-	-	-	-	-	-	-	-	-
y	-	-	+	-	+	-	-	-	-	-	-	-	+
œ̃	+	-	+	-	+	-	-	-	+	-	-	-	+
õ	+	-	-	-	+	-	-	-	+	-	-	-	+
ɔ̃	+	+	-	-	+	-	-	-	+	-	-	-	+
ɑ̃	+	-	-	-	-	-	-	-	+	-	-	-	+
ɛ̃	+	-	+	-	-	-	-	-	+	-	-	-	+

Table A.12: Elements - German

Segment	|A|	|A̲|	|I|	|I̲|	|U|	|U̲|	|H|	|H̲|	|L|	|L̲|	|ʔ|	|ʔ̲|	Vowel	Long
b	-	-	-	-	+	+	-	-	-	-	+	-	-	
p	-	-	-	-	+	+	+	+	-	-	+	-	-	
t	+	-	-	-	-	-	+	+	-	-	+	-	-	
d	+	-	-	-	-	-	-	-	-	-	+	-	-	
k	-	-	-	-	+	-	+	+	-	-	+	-	-	
g	-	-	-	-	+	-	-	-	-	-	+	-	-	
p͡f	+	-	-	-	+	+	+	+	-	-	+	-	-	
t͡s	-	-	+	-	-	-	+	+	-	-	+	-	-	
f	+	-	-	-	+	+	+	+	-	-	-	-	-	
v	+	-	-	-	+	+	+	-	+	-	-	-	-	
s	-	-	+	-	-	-	+	+	-	-	-	-	-	
z	-	-	+	-	-	-	+	-	+	-	-	-	-	
ʃ	-	-	+	+	-	-	+	+	-	-	-	-	-	
ç	-	-	+	+	+	-	+	+	-	-	-	-	-	
x	-	-	-	-	+	-	+	+	-	-	-	-	-	
h	-	-	-	-	-	-	+	-	-	-	-	-	-	
m	-	-	-	-	+	+	-	-	+	-	+	-	-	
n	+	-	-	-	-	-	-	-	+	-	+	-	-	
ŋ	-	-	-	-	+	-	-	-	+	-	+	-	-	
l	+	-	+	-	-	-	-	-	-	-	-	-	-	
j	-	-	+	+	-	-	-	-	-	-	-	-	-	
ʁ	+	-	-	-	-	-	-	-	-	-	-	-	-	
ɾ	+	-	-	-	-	-	-	-	-	-	-	-	-	
l̩	+	-	+	-	-	-	-	-	-	-	-	-	+	
iː	-	-	+	+	-	-	-	-	-	-	-	-	+	✓
ɪ	-	-	+	-	-	-	-	-	-	-	-	-	+	✗
i	-	-	+	+	-	-	-	-	-	-	-	-	+	✗
eː	+	-	+	+	-	-	-	-	-	-	-	-	+	✓
ɛː	+	-	+	-	-	-	-	-	-	-	-	-	+	✓
ɛ	+	-	+	+	-	-	-	-	-	-	-	-	+	✗
e	+	-	+	+	-	-	-	-	-	-	-	-	+	✗
yː	-	-	+	-	+	-	-	-	-	-	-	-	+	✓
y	-	-	+	-	+	-	-	-	-	-	-	-	+	✗
ʏ	-	-	+	-	+	-	-	-	-	-	-	-	+	✗
øː	+	-	+	+	+	-	-	-	-	-	-	-	+	✓
œ	+	-	+	-	+	-	-	-	-	-	-	-	+	✗
ø	+	-	+	+	+	-	-	-	-	-	-	-	+	✗
aː	+	+	-	-	-	-	-	-	-	-	-	-	+	✓
a	+	+	-	-	-	-	-	-	-	-	-	-	+	✗
ə	+	-	-	-	-	-	-	-	-	-	-	-	+	✗
ɐ	+	-	-	-	-	-	-	-	-	-	-	-	+	✗
uː	-	-	-	-	+	+	-	-	-	-	-	-	+	✓
ʊ	-	-	-	-	+	-	-	-	-	-	-	-	+	✗
u	-	-	-	-	+	-	-	-	-	-	-	-	+	✗
oː	+	-	-	-	+	-	-	-	-	-	-	-	+	✓
ɔ	+	+	-	-	+	-	-	-	-	-	-	-	+	✗
o	+	-	-	-	+	-	-	-	-	-	-	-	+	✗

TABLE A.13: Elements - Greek

Segment	|A|	|A̲|	|I|	|I̲|	|U|	|U̲|	|H|	|H̲|	|L|	|L̲|	|ʔ|	|ʔ̲|	Vowel
a	+	-	-	-	-	-	-	-	-	-	-	-	+
e	+	-	+	-	-	-	-	-	-	-	-	-	+
i	-	-	+	-	-	-	-	-	-	-	-	-	+
o	+	-	-	-	+	-	-	-	-	-	-	-	+
u	-	-	-	-	+	-	-	-	-	-	-	-	+
w	-	-	-	-	+	+	-	-	-	-	-	-	-
m	-	-	-	-	+	+	-	-	+	-	+	-	-
n	+	-	-	-	-	-	-	-	+	-	+	-	-
p	-	-	-	-	+	+	-	-	-	-	+	-	-
b	-	-	-	-	+	+	-	-	+	+	+	-	-
t	+	-	-	-	-	-	-	-	-	-	+	-	-
d	+	-	-	-	-	-	-	-	+	+	+	-	-
k	-	-	-	-	+	-	-	-	-	-	+	-	-
g	-	-	-	-	+	-	-	-	+	+	+	-	-
f	+	-	-	-	+	+	+	-	-	-	-	-	-
v	+	-	-	-	+	+	+	-	+	+	-	-	-
θ	-	-	+	-	-	-	+	-	-	-	-	-	-
ð	-	-	+	-	-	-	+	-	+	+	-	-	-
s	+	-	-	-	-	-	+	-	-	-	-	-	-
z	+	-	-	-	-	-	+	-	+	+	-	-	-
x	-	-	-	-	+	-	+	-	-	-	-	-	-
ɣ	-	-	-	-	+	-	+	-	+	+	-	-	-
l	+	-	+	-	-	-	-	-	-	-	-	-	-
r	+	+	-	-	-	-	-	-	-	-	-	-	-

Table A.14: Elements - Portuguese

Segment	|A|	|A̲|	|I|	|I̲|	|U|	|U̲|	|H|	|H̲|	|L|	|L̲|	|ʔ|	|ʔ̲|	Vowel
a	+	+	-	-	-	-	-	-	-	-	-	-	+
ɐ	+	-	-	-	-	-	-	-	-	-	-	-	+
ɐ̃	+	-	-	-	-	-	-	-	+	-	-	-	+
b	-	-	-	-	+	+	-	-	+	+	+	-	-
ɔ	+	+	-	-	+	-	-	-	-	-	-	-	+
d	+	-	-	-	-	-	-	-	+	+	+	-	-
e	+	-	+	+	-	-	-	-	-	-	-	-	+
ẽ	+	-	+	+	-	-	-	-	+	-	-	-	+
ɛ	+	-	+	-	-	-	-	-	-	-	-	-	+
ɘ	+	-	-	-	-	-	-	-	-	-	-	-	+
f	+	-	-	-	+	+	+	-	-	-	-	-	-
g	-	-	-	-	+	-	-	-	+	+	+	-	-
i	-	-	+	+	-	-	-	-	-	-	-	-	+
ĩ	-	-	+	-	-	-	-	-	+	-	-	-	+
ɨ	-	-	+	-	-	-	-	-	-	-	-	-	+
j̃	-	-	+	+	-	-	-	-	+	-	-	-	-
j	-	-	+	+	-	-	-	-	-	-	-	-	-
k	-	-	-	-	+	-	-	-	-	-	+	-	-
l	+	-	+	-	-	-	-	-	-	-	-	-	-
m	-	-	-	-	+	+	-	-	+	-	+	-	-
n	+	-	-	-	-	-	-	-	+	-	+	-	-
ɲ	-	-	+	+	+	-	-	-	+	-	+	-	-
o	+	-	-	-	+	-	-	-	-	-	-	-	+
õ	+	-	-	-	+	-	-	-	+	-	-	-	+
p	-	-	-	-	+	+	-	-	-	-	+	-	-
ʁ	+	-	-	-	-	-	-	-	-	-	-	-	-
ʁ̞	+	-	-	-	-	-	-	-	-	-	-	-	-
ɾ	-	-	+	-	-	-	-	-	-	-	-	-	-
s	+	-	-	-	-	-	+	-	-	-	-	-	-
ʃ	-	-	+	+	-	-	+	-	-	-	-	-	-
t	+	-	-	-	-	-	-	-	-	-	+	-	-
u	-	-	-	-	+	-	-	-	-	-	-	-	+
v	+	-	-	-	+	+	+	-	+	+	-	-	-
w	-	-	-	-	+	+	-	-	-	-	-	-	-
w̃	-	-	-	-	+	+	-	-	+	-	-	-	-
ʎ	+	-	+	+	-	-	-	-	-	-	-	-	-
z	+	-	-	-	-	-	+	-	+	+	-	-	-
ʒ	-	-	+	+	-	-	+	-	+	+	-	-	-
ũ	-	-	-	-	+	-	-	-	+	-	-	-	+

A.5 Training and test texts

A.5.1 English - Mark 1

 Transliteration to IPA based on CELEX (Baayen, Piepenbrock and Rijn, 1993).

ðə bɪgɪnɪŋ əv ðə gʊd nju:z əbaʊt ʤi:zəs ðə mɪsaɪə ðə sʌn əv gɒd æz ɪt ɪz ɹɪtn̩ ɪn aɪzaɪə ðə pɹɒfɪt aɪ wɪl sɛnd maɪ mɛsɪnʤə əhɛd əv ju: hu: wɪl pɹɪpɛə jɔ: weɪ ə vɔɪs əv wʌn kɔ:lɪŋ ɪn ðə wɪldənəs pɹɪpɛə ðə weɪ fə ðə lɔ:d meɪk stɹeɪt pɑ:ðz fɔ: hɪm ænd səʊ ʤɒn ðə bæptɪst əpɪəd ɪn ðə wɪldənɪs pɹi:ʧɪŋ ə bæptɪzəm əv ɹɪpɛntəns fə ðə fəgɪvnɪs əv sɪnz ðə həʊl ʤju:di:ən kʌntɪsaɪd ən ɔ:l ðə pi:pl̩ əv ʤɔɹu:sələm wɛnt aʊt tə hɪm kənfɛsɪŋ ðɛə sɪnz ðeɪ wɜ: bæptaɪzd baɪ hɪm ɪn ðə ʤɔ:dən ɹɪvə ʤɒn wɔ: kləʊðɪŋ meɪd əv kæml̩z hɛə wɪð ə leðə bɛlt əɹaʊnd hɪz weɪst ænd hi: ɛt ləʊkəsts ænd waɪld hʌnɪ ænd ðɪs wəz hɪz mɛsɪʤ ɑ:ftə mi: kʌmz ðə wʌn mɔ: paʊəfʊl ðən aɪ ðə stɹæps əv hu:z sændl̩z aɪ æm nɒt wɜ:ðɪ tə stu:p daʊn ænd ʌntaɪ aɪ bæptaɪz ju: wɪð wɔ:tə bʌt hi: wɪl bæptaɪz ju: wɪð ðə həʊlɪ spɹɪɹt ət ðæt taɪm ʤi:zəs keɪm fɹɒm næzæɹəθ ɪn gæləli: ən wəz bæptaɪzd baɪ ʤɒn ɪn ðə ʤɔ:dən ʤʌst əz ʤi:zəs wəz kʌmɪŋ ʌp aʊt əv ðə wɔ:tə hi: sɔ: hɛvn̩ bi:ɪŋ tɔ:n əʊpən ən ðə spɹɪɹt dɪsendɪŋ ɒn hɪm laɪk ə dəʊv ænd ə vɔɪs keɪm fɹɒm hɛvn̩ ju: ɑ: maɪ sʌn hu:m aɪ lʌv wɪð ju: aɪ æm wɛl pli:zd ət wʌns ðə spɹɪɹt sɛnt hɪm aʊt ɪntə ðə wɪldənɪs ænd hi: wəz ɪn ðə wɪldənɪs fɔ:tɪ deɪz bi:ɪŋ tɛmptɪd baɪ seɪtən hi: wəz wɪð ðə waɪld ænɪml̩z ænd eɪnʤəlz ətɛndɪd hɪm ɑ:ftə ʤɒn wəz pʊt ɪn pɹɪzn̩ ʤi:zəs wɛnt ɪntə gæləli: pɹəkleɪmɪŋ ðə gʊd nju:z əv gɒd ðə taɪm hæz kʌm hi: sɛd ðə kɪŋdəm əv gɒd hæz kʌm nɪə ɹɪpɛnt ænd bɪli:v ðə gʊd nju:z æz ʤi:zəs wɔ:kt bɪsaɪd ðə si: əv gæləli: hi: sɔ: saɪmən ən hɪz bɹʌðə ændɹu: kɑ:stɪŋ ə nɛt ɪntə ðə leɪk fə ðeɪ wɜ: fɪʃəmən kʌm fʊləʊ mi: ʤi:zəs sɛd ænd aɪ wɪl sɛnd ju: aʊt tə fɪʃ fɔ: pi:pl̩ æt wʌns ðeɪ lɛft ðɛə nɛts ən fʊləʊd hɪm wɛn hi: hæd gɒn ə lɪtl̩ fɑ:ðə hi: sɔ: ʤɛɪmz sʌn əv zɛbədi: ænd hɪz bɹʌðə ʤɒn ɪn ə bəʊt pɹɪpɛəɹɪŋ ðɛə nɛts wɪðaʊt dɪleɪ hi: kɔ:ld ðəm æn ðeɪ lɛft ðɛə fɑ:ðə zɛbədi: ɪn ðə bəʊt wɪð ðə haɪəd mɛn ən fʊləʊd hɪm ðeɪ wɛnt tə kəpɜ:ɹniəm ən wɛn ðə sæbəθ keɪm ʤi:zəs wɛnt ɪntə ðə sɪnəgɒg æn bɪgæn tə ti:ʧ ðə pi:pl̩ wɜ: əmeɪzd æt hɪz ti:ʧɪŋ bɪkɒz hi: tɔ:t ðɛm æz wʌn hu: hæd ɔ:θɒɹətɪ nɒt əz ðə ti:ʧəz əv ðə lɔ: ʤʌst ðɛn ə mæn ɪn ðɛə sɪnəgɒg hu: wɒz pəzɛst baɪ ən ɪmpjʊə spɹɪɹt kɪaɪd aʊt wɒt du: ju: wɒnt wɪð ʌs ʤi:zəs əv næzæɹəθ hæv ju: kʌm tə dɪstɪɔɪ ʌs aɪ nəʊ hu: ju: ɑ: ðə həʊlɪ wʌn əv gɒd bi: kwaɪət sɛd ʤi:zəs stɜ:nlɪ kʌm aʊt əv hɪm ði: ɪmpjʊə spɹɪɹt ʃʊk ðə

mæn vaɪələntlɪ ən keɪm aʊt əv hɪm wɪð ə ʃɹiːk ðə piːpl̩ wɜː ɔːl səʊ əmeɪzd ðæt ðeɪ ɑːskt iːfʃ ʌðə wɒt ɪz ðɪs ə njuː tiːfʃɪŋ ænd wɪð ɔːθɒɹətɪ hiː iːvn̩ gɪvz ɔːdəz tuː ɪmpjʊə spɪɹɪts ən ðeɪ əbeɪ hɪm njuːz əbaʊt hɪm spɹɛd kwɪklɪ əʊvə ðə həʊl ɹiːʤən əv gæləliː əz suːn æz ðeɪ lɛft ðə sɪməgɒg ðeɪ wɛnt wɪð ʤeɪmz ən ʤɒn tə ðə həʊm əv saɪmən ænd ændɹuː saɪmənz mʌðəɪm lɔː wəz ɪn bɛd wɪð ə fiːvə ænd ðeɪ ɪmiːdjətlɪ təʊld ʤiːzəs əbaʊt hɜː səʊ hiː wɛnt tə hɜː tʊk hɜː hænd ænd hɛlpt hɜː ʌp ðə fiːvə lɛft hɜː ænd ʃiː bɪgæn tə weɪt ɒn ðɛm ðæt iːvn̩ɪŋ ɑːftə sʌnsɛt ðə piːpl̩ bɹɔːt tə ʤiːzəs ɔːl ðə sɪk ən diːmənpəzɛst ðə həʊl taʊn gæðəd æt ðə dɔː ænd ʤiːzəs hiːld mɛnɪ huː hæd vɛəɹɪəs dɪziːzɪz hiː ɔːlsəʊ dɹəʊv aʊt mɛnɪ diːmənz bʌt hiː wʊd nɒt lɛt ðə diːmənz spiːk bɪkɒz ðeɪ njuː huː hiː wɒz vɛɹɪ ɜːlɪ ɪn ðə mɔːnɪŋ waɪl ɪt wəz stɪl dɑːk ʤiːzəs gɒt ʌp lɛft ðə haʊz wɛnt ɒf tuː ə sɒlɪtəɹɪ pleɪs wɛə hiː pɹeɪd saɪmən ən hɪz kəmpænjənz wɛnt tə lʊk fə hɪm æn wɛn ðeɪ faʊnd hɪm ðeɪ ɪkskleɪmd ɛvɹɪwʌn ɪz lʊkɪŋ fɔː juː ʤiːzəs ɹɪplaɪd lɛt ʌs gəʊ sʌmwɛə ɛls tə ðə nɪəbaɪ vɪlɪʤɪz səʊ aɪ kæn pɹiːfʃ ðɛə ɔːlsəʊ ðæt ɪz waɪ aɪ hæv kʌm səʊ hiː tɹævl̩d θɹuːaʊt gæləliː pɹiːfʃɪŋ ɪn ðɛə sɪməgɒgz æn dɹaɪvɪŋ aʊt diːmənz ə mæn wɪð lɛpɹəsɪ keɪm tə hɪm ænd bɛgd hɪm ɒn hɪz niːz ɪf juː ɑː wɪlɪŋ juː kæn meɪk miː kliːn ʤiːzəs wɒz ɪndɪgnənt hiː ɹiːfʃt aʊt hɪz hænd ænd tʌfʃt ðə mæn aɪ æm wɪlɪŋ hiː sɛd biː kliːn ɪmiːdjətlɪ ðə lɛpɹəsɪ lɛft hɪm ənd hiː wəz klɛnzd ʤiːzəs sɛnt hɪm əweɪ ət wʌns wɪð ə stɹɒŋ wɔːnɪŋ siː ðət juː dəʊnt tɛl ðɪs tə ɛnɪwʌn bʌt gəʊ ʃəʊ jɔːsɛlf tə ðə pɹiːst ænd ɒfə ðə sækɹɪfaɪsɪz ðæt məʊzɪz kəmɑːndɪd fɔː jɔː klɛnzɪŋ æz ə tɛstɪmənɪ tuː ðɛm ɪnstɛd hiː wɛnt aʊt ænd bɪgæn tuː tɔːk fɹiːlɪ spɹɛdɪŋ ðə njuːz æz ə ɹɪzʌlt ʤiːzəs kʊd nəʊ lɒŋgə ɛntə ə taʊn əʊpn̩lɪ bʌt steɪd aʊtsaɪd ɪn ləʊnlɪ pleɪsɪz jɛt ðə piːpl̩ stɪl keɪm tə hɪm fɹɒm ɛvɹɪwɛə

A.5.2 Dutch - Mark 1

A.5.3 French - Mark 1

A.5.4 German - Mark 1

Bibeltext der Schlacter Copyright © 2000 Genfer Bibelgesellschaft. Transliteration to IPA based on CELEX (Baayen, Piepenbrock and Rijn, 1993).

anfaŋ dɛs eːvaŋgeːliːʊms fɔn jezʊs kɹɪstʊs deːm zoːn gɔtəs viː gəʃɹiːbən ʃteːt ɪn deːn pɹoːfeːtən ɛseːə ɪx zɛndə mainən boːtən foːɹ dainəm angəzɪçt heːɹ deːɹ dainən veːk foːɹ diːɹ bəɹaitən vɪɹt diː ʃtɪmə ainəs ɹuːfədeːn ɛɹtøːnt ɪn deːɹ vyːstə bəɹaitət deːn veːk dɛs hɛʁn maxt zainə p͡faːdə eːbən zoː bəgan johanəs ɪn deːɹ vyːstə tauftə ʊnt fɛɹkʏndɪxtə ainə taufə deːɹ bysə f͡suːɹ fɛɪgeːboŋ deːɹ zʏndən ʊnt ɛs gɪŋ f͡suː iːm hɪnaus das ganf͡sə lant judɛa ʊnt diː bəvoːnəɹ fɔn jeːɹuːzalɛm ʊnt ɛs vʏɹdən fɔn iːm alə ɪm jɔɹdan gətauft diː iːɹə zʏndən bəkantən johanəs aːbəɹ vaːɹ bəklaidət mɪt kameːlhaːɹən ʊnt tɹuːk ainən leːdəɹnən gʏʁtˌ ʊm zainə lɛndən ʊnt eːɹ aːs hɔyʃɹɛkən ʊnt vɪldən hoːnɪx ʊnt eːɹ fɛɹkʏndɪxtə ʊnt ʃpɹaːx ɛs kɔmt ainəɹ naːx miːɹ deːɹ ʃtaɹkəɹ ɪst als ɪx ʊnt ɪx bɪn nɪxt vʏɹdɪx iːm gəbʏkt zainən ʃuːɹiːmən f͡suː loːzən ɪx haːbə ɔyx mɪt vasəɹ gətauft eːɹ aːbəɹ vɪɹt ɔyx mɪt hailɪgəm gaist taufən ʊnt ɛs gəʃaː ɪn jeːnən taːgən das jezʊs fɔn naːf͡saʁɛt ɪn galilɛja kaːm ʊnt zɪx fɔn johanəs ɪm jɔɹdan taufən liːs ʊnt zoːglaix als eːɹ aus deːm vasəɹ ʃtiːk zaː eːɹ deːn hɪməl f͡sɛɹɪsən ʊnt deːn gaist viː ainə taubə auf iːn hɛɹapʃtaigən ʊnt ainə ʃtɪmə ɛɹtøːntə aus deːm hɪməl duː bɪst main gəliːptəɹ zoːn an deːm ɪx voːlgəfalən haːbə ʊnt zoːglaix tɹaipt iːn deːɹ gaist ɪn diː vyːstə hɪnaus ʊnt eːɹ vaːɹ fɪʁtsɪç taːgə dɔɹt ɪn deːɹ vyːstə ʊnt vʏɹdə fɔn deːm zaːtan fɛɪzuːxt ʊnt eːɹ vaːɹ bai deːn vɪldən tiːɹən ʊnt diː ɛŋəl diːntən iːm naːxdeːm aːbəɹ johanəs gəfaŋən gənɔmən vɔɹdən vaːɹ kaːm jezʊs naːx galilɛja ʊnt fɛɹkʏndɪxtə das eːvaŋgeːliːʊm fɔm ɹaix gɔtəs ʊnt ʃpɹaːx diː f͡sait ɪst ɛɹfʏlt ʊnt das ɹaix gɔtəs ɪst naːə tuːt bysə ʊnt glaupt an das eːvaŋgeːliːʊm als eːɹ aːbəɹ am zeː fɔn galilɛja ɛntlaŋgɪŋ zaː eːɹ siːmɔn ʊnt dɛsən bɹuːdəɹ andɹeːas diː vaɹfən das nɛf͡s aus ɪm zeː dɛn ziː vɛːɹən fɪʃəɹ ʊnt jezʊs ʃpɹaːx f͡suː iːnən fɔlkt miːɹ naːx ʊnt ɪx vɪl ɔyx f͡suː mɛnʃənfɪʃəɹn maxən daː fɛɹliːsən ziː zoːglaix iːɹə nɛf͡sə ʊnt fɔlktən iːm naːx ʊnt als eːɹ fɔn dɔɹt ain veːnɪx vaitəɹgɪŋ zaː eːɹ jakoːbʊs deːn zoːn dɛs f͡sɛbɛdɛjʊs ʊnt zainən bɹuːdəɹ johanəs diː aux ɪm ʃɪf vɛːɹən ʊnt diː nɛf͡sə flɪktən ʊnt zoːglaix bəɹiːf eːɹ ziː ʊnt ziː liːsən iːɹən fɛːtəɹ f͡sɛbɛdɛjʊs zamt deːn taːgəløːnəɹn ɪm ʃɪf ʊnt fɔlktən iːm naːx ʊnt ziː bəgaːbən zɪx naːx kaːpəɹnʊm ʊnt eːɹ gɪŋ am zabat zoːglaix ɪn diː zynagoːgə ʊnt leːɹtə ʊnt ziː ɛɹʃtauntən yːbɛ zainə leːɹə dɛn eːɹ leːɹtə ziː viː ainəɹ deːɹ fɔlmaxt hat ʊnt nɪxt viː diː ʃɹɪftgəleːɹtən ʊnt ɛs vaːɹ ɪn iːɹəɹ zynagoːgə ain mɛnʃ mɪt ainəm ʊnɹainən gaist deːɹ ʃɹiː ʊnt ʃpɹaːx las ap vas haːbən viːɹ mɪt diːɹ f͡suː tuːn jezʊs duː naf͡saɹeːnəɹ bɪst duː gəkɔmən ʊm ʊns f͡suː fɛɹdɛɹbən ɪx vais veːɹ duː bɪst deːɹ hailɪgə gɔtəs aːbəɹ jezʊs bəfaːl iːm ʊnt ʃpɹaːx fɛɹʃtʊmə

ʊnt faːɾə aus fɔn iːm daː fsɛɾtə iːn deːɾ ʊnɾainə gaist hɪn ʊnt heːɾ ʃiː mɪt lautəɾ ʃtɪmə ʊnt fuːɾ fɔn iːm aus ʊnt ziː ɛɾʃtauntən alə zoːdas ziː zɪx ʊntəɾainandəɾ fɾaːktən ʊnt ʃpɾɛːxən vas ɪst das vas fyɾ ainə nɔyə leːɾə ɪst diːs mɪt fɔlmaxt gəbiːtət eːɾ aux deːn ʊnɾainən gaistəɾn ʊnt ziː gəhɔɾxən iːm ʊnt das gəʁʏçt fɔn iːm fɛɾbɾaitətə zɪx zoːglaix ɪn das ganfsə ʊmliːgəndə gəbiːt fɔn galilɛja ʊnt zoːglaix fɛɾliːsən ziː diː zynagoːgə ʊnt gɪŋən mɪt jakoːbɔs ʊnt johanəs ɪn das haus dɛs siːmɔn ʊnt andɾeːas siːmɔns ʃviːgəɾmʊtəɾ aːbəɾ laːk kɾaŋk am fiːbəɾ daniːdə ʊnt zoːglaix zaːktən ziː iːm fɔn iːɾ ʊnt eːɾ tɾaːt hɪnfsuː ɛɾgɾɪf iːɾə hant ʊnt ɾɪxtətə ziː auf ʊnt das fiːbəɾ fɛɾliːs ziː zoːglaix ʊnt ziː diːntə iːnən als ɛs aːbəɾ aːbənt gəvɔɾdən ʊnt diː zɔnə ʊntəɾgəgaŋən vaːɾ bɾɛxtən ziː alə kɾaŋkən ʊnt bəzɛsənən fsuː iːm ʊnt diː ganfsə ʃtat vaːɾ foːɾ deːɾ tyɾ fɛɾzaməlt ʊnt eːɾ hailtə fiːlə diː an manxəɾlai kɾaŋkhaitən lɪtən ʊnt tɾiːp fiːlə dɛmoːnɛn aus ʊnt liːs diː dɛmoːnɛn nɪxt ɾeːdən dɛn ziː kantən iːn ʊnt am mɔɾgən als ɛs nɔx zeːɾ dʊŋkəl vaːɾ ʃtant eːɾ auf gɪŋ hɪnaus an ainən ainzaːmən ɔɾt ʊnt beːtətə dɔɾt ʊnt ɛs fɔlktən iːm siːmɔn ʊnt diː vɛlxə bai iːm vɛːɾən ʊnt als ziː iːn gəfʊndən hɛtən ʃpɾɛːxən ziː fsuː iːm jeːdəɾman zuːxt dɪx ʊnt eːɾ ʃpɾɪxt fsuː iːnən last ʊns ɪn diː ʊmliːgəndən ɔɾtə geːən daːmɪt ɪx aux dɔɾt fɛɾkʏndɪgə dɛn daːfsuː bɪn ɪx gəkɔmən ʊnt eːɾ fɛɾkʏndɪxtə ɪn iːɾən zynagoːgən ɪn ganfs galilɛja ʊnt tɾiːp diː dɛmoːnɛn aus ʊnt ɛs kaːm ain aʊszɛtsɪgəɾ fsuː iːm baːt iːn fiːl foːɾ iːm auf diː kniː ʊnt ʃpɾaːx fsuː iːm vɛn duː vɪlst kanst duː mɪx ɾainɪgən daː ɛɾbaɾmtə zɪx jezʊs yːbɐ iːn ʃtɾɛktə diː hant aus ɾyɾtə iːn an ʊnt ʃpɾaːx fsuː iːm ɪx vɪl zai gəɾainɪxt ʊnt vɛːɾənt eːɾ ɾeːdətə vɪx deːɾ auszafs zoːglaix fɔn iːm ʊnt eːɾ vʏɾdə ɾain ʊnt eːɾ ɛɾmaːntə iːn ɛɾnstlɪx ʊnt ʃɪktə iːn zoːglaix fɔɾt ʊnt ʃpɾaːx fsuː iːm haːp axt zaːgə niːmant ɛtvas zɔndəɾn geː hɪn fsaigə dɪx deːm pɾiːstəɾ ʊnt ɔpfəɾə fyɾ dainə ɾainɪgʊŋ vas moːsə bəfoːlən hat iːnən tsʊm fsɔyknɪs eːɾ aːbəɾ gɪŋ ʊnt fɪŋ an ɛs fiːlfax fsuː fɛɾkʏndɪgən ʊnt bɾaitətə diː zaxə ybəɾal aus zoːdas jezʊs nɪxt meːɾ œfəntlɪx ɪn ainə ʃtat hiːnaingeːən kœntə zɔndəɾn eːɾ vaːɾ dɾausən an ainzaːmən ɔɾtən ʊnt ziː kɛːmən fɔn alən zaitən fsuː iːm

A.5.5 Greek - Mark 1

Today's Greek Version (Society) and Het Nederlands Bijbelgenootschap, 1996). Transliteration to IPA based on GreekLex (Ktori, Heuven and Pitchford, 2008).

aiti eine i arxi tu xarmosinu minimatos ɣia ton iisoi xristo ton iio tu θeoi sta vivlia ton profiton eine ɣrammeno stelno ton aɣɣelioforo mu prin apo sena ɣia na proetimasi to ðromo su mia foni vrodofonazi stin erimo etimaste to ðromo ɣia ton kirio isioste ta monopatia na perasi simfona

m afta parusiastike o ioannis o opoios vaftize stin erimo ke kiritte na metanoisun i anθropi ke na vaftistoin ɣia na siɣxoriθoin i amarties tus piɣenan s afton oli i katiki tis iuðaias ki i ierosolimites ki olus tus vaftize ston potamo iorðani kaθos omoloɣoisan tis amarties tuso ioannis foroise roixo apo trixes kamilas ke ðermatini zoni sti mesi tu etroɣe akriðes ke meli apo ayriomelisses sto kiriɣma tu tonize erxete istera apo mena aftos pu eine pio isxiros ke pu eɣo ðen eime aksios na skipso ke na liso to luri apo ta ipoðimata tu eɣo sas vaftisa me nero ekeinos omos θa sas vaftisi me ayio pneima ekeines tis meres irθe o iisois apo ti nazaret tis galilaias ke vaftistike ston iorðani apo ton ioanniki amesos eno evɣene apo to nero eiðe n anoiɣun i uranoi ke to pneima san peristeri na katevaini pano tu tote mia foni akoistike apo ta urania esi eise o ayapimenos mu iios esi eise o eklektos mu amesos to pneima oðiɣei ton iisoi ekso stin erimo ekei stin erimo emine sarada meres ki adimetopise tus pirasmois tu satana zoise mazi me ta θiria ke ayɣeli ton ipiretoisan meta ti sillipsi tu ioanni o iisois irθe sti galilaia ke kiritte to xarmosino minima ɣia ti vasileia tu θeoi sibliroθike eleɣe o kaθorismenos keros ki eftase i vasileia tu θeoi metanoeite ke pisteiete sto xarmosino afto minima kaθos o iisois perpatoise stin oxθi tis limnis tis galilaias eiðe to simona ke ton anðrea aðerfo tu simona na rixnun ta ðixtia sti limni ɣiati itan psaraðes akoluθiste me tus eipe o iisois ke θa sas kano psaraðes anθropon ekeini amesos afisan ta ðixtia ke ton akoloiθisan afoi proxorise liɣo pio pera o iisois eiðe ton iakovo ɣio tu zeveðaiu ke ton aðerfo tu ton ioanni na taktopioin ki aftoi ta ðixtia mesa sto psarokaiko ke tus kalese amesos aitoi afisan tote ton patera tus to zeveðaio sto psarokaiko me tus misθotois ke ton akoloiθisan erxode stin kapernaoim ki amesos to savvato o iisois bike sti sinaɣoɣi ke ðiðaske oi anθropi emenan kataplikti apo ti ðiðaskalia tu ɣiati tus ðiðaske me afθedia ki oxi opos ðiðaskan i ɣrammateis ekei sti sinaɣoɣi tus itan kapios pu katexotan apo ðemoniko pneima aitos kraiɣase leɣodase ti ðulia exis esi m emas iisoi nazarine irθes na mas afanisis se ksero pios eise eise o eklektos tu θeoi o iisois epitimise to ðemoniko pneima ke tu eipe papse na milas ke vɣes ap afton to ðemoniko pneima afoi sidarakse ton anθropo ke fonakse me ðinati foni vɣike ap aftonoli tote kirieitikan apo ðeos ke sizitoisan metaksi tus ti simainun ola afta pia eine i kenoiria afti ðiðaskalia me afθedia ðiatazi akomi ke ta ðemonika pneimata ke ton ipakoine ki amesos kikloforise i fimi tu padoi s oli tin perioxi tis galilaias molis vɣikan apo ti sinaɣoɣi irθan sto spiti tu simona ke tu anðrea me ton iakovo ke ton ioanniamesos lene ston iisoi ɣia tin peθera tu simona pu itan sto krevati me pireto o iisois tin plisiase tin epiase apo to xeri ke ti sikose o piretos tote tin afise amesos ki afti

tus ipiretoise kata to ðilino otan eðise o ilios toi eferan olus tus arrostus ke tus ðemonismenus ki oli i katiki tis polis eixan mazeftei brosta stin porta o iisois θerapefse pollois pu ipeferan apo ðiafores arrosties ki evɣale polla ðemonia ðen ta afine omos na miloin ɣiati ton anaɣnorizan oti eine o messias to proi poli prin akoma feksi o iisois vɣike ekso ke piɣe s ena erimiko meros ki ekei prosefxotanton anazitisan omos o simon ki i sidrofoi tu ton vrikan ke tu lene oli se zitoin ekeinos tus lei pame sta ɣitonika xoria ɣia na kirikso ki ekei afti eine i apostoli mu kiritte lipon stis sinaɣoɣes tus s oli ti galilaia ki evɣaze ta ðemonia erxete ston iisoi enas lepros ke pesmenos sta ɣonata ton parakaloise leɣodas ean θelis exis ti ðinami na me kaθarisis apo ti lepra o iisois ton splaxnistike aplose to xeri tu ton aɣɣikse ke tu eipe θelo na kaθaristeis apo ti lepra molis ta eipe afta amesos efiɣe ap afton i lepra ke kaθaristike ke sinoðeiodas ton ekso o iisois tu milise se tono afstiro ke tu eipe prosekse min pis tipota se kanenan piɣene omos na ðeiksis ton eafto su ston ierea ke prosfere ɣia ton kaθarismo su oti exi kaθorisi o moisis ɣia na tus apoðeiksis oti θerapeitikes aitos omos vɣike ki arxise na ðialalei ta pada ke na ðiaðiði to ɣeɣonos etsi pu o iisois ðen boroise pia na bi fanera se kapia poli alla emene ekso se erimika meri ostoso erxotan s afton o kosmos apo padoi

A.5.6 Portuguese - Mark 1

A.5.7 Spanish - Mark 1

el prinθipjo del ebanxeljo de xesukristo el ixo de djos komo esta eskrito en el profeta isaias e aki embio mi mensaxero delante de ti kjen preparara tu kamino boθ del ke proklama en el desjerto preparen el kamino del seɲor endereθen sus sendas asi xwan el bautista apareθjo en el desjerto predikando el bautizmo del arepentimjento para perdon de pekados i salia a el toda la probinθja de xuea i todos los de xerusalen i eran bautiθados poɾ el en el ɾio xordan komfesando sus pekados xwan estaba bestido de pelo de kameʎo i kon un θinto de kwero a la θintura i komia

langostas i mjel silbestre i predikaba diθjendo bjene tras de mi el ke es mas poderoso ke jo a kjen no soi digno de desatar agaʧado la korea de su kalθado jo les e bautiθado en agwa pero el les bautiθara en el espiritu santo akonteθjo en akeʎos dias ke xesus bino de naθared de galilea i fwe bautiθado por xwan en el xordan i en segida mjentras subia del agwa bjo ke los θjelos se abrian i ke el espiritu desθendia sobre el komo paloma i bino una boθ dezde el θjelo tu eres mi ixo amado en ti tengo komplaθenθja en segida el espiritu lo impulso al desjerto i estubo en el desjerto kwarenta dias sjendo tentado por satanas estaba kon las fjeras i los anxeles le serbian despwes ke xwan fwe enkarθelado xesus se fwe a galilea predikando el ebanxeljo de djos i diθjendo el tjempo se a kumplido i el reino de djos se a aθerkado arepjentanse i krean en el ebanxeljo i pasando xunto al mar de galilea bjo a simon i a andres ermano de simon eʧando la red en el mar porke eran peskadores xesus les dixo bengan en pos de mi i los are peskadores de ombres i de immedjato dexaron sus redes i lo sigjeron al ir un poko mas adelante bjo a xakobo ixo de θebedeo i a su ermano xwan eʎos estaban en su barka areglando las redes en segida los ʎamo i eʎos dexando a su padre θebedeo en la barka xunto kon los xornaleros se fweron en pos de el entraron en kapernaum i en segida entrando el en la sinagoga los sabados enseɲaba i se asombraran de su enseɲanθa porke les enseɲaba komo kjen tjene autoridad i no komo los eskribas i en ese momento un ombre kon espiritu immundo estaba en la sinagoga de eʎos i esklamo diθjendo ke tjenes kon nosotros xesus de naθared as benido para destrwirnos jo se kjen eres el santo de djos xesus le reprendjo diθjendo kaʎate i sal de el i el espiritu immundo lo sakudjo kon bjolenθja klao a gran boθ i saljo de el todos se marabiʎaon de modo ke diskutian entre si diθjendo ke es esto una nweba doktrina kon autoridad aun a los espiritus immundos el manda i lo obedeθen i pronto se estendjo su fama por todas partes en toda la rexjon alrededor de galilea en segida kwando saljeron de la sinagoga fweron kon xakobo i xwan a la kasa de simon i andres la swegra de simon estaba en kama kon fjebre i de immedjato le ablaron de eʎa el se aθerko a eʎa la tomo de la mano i la lebanto i le dexo la fjebre i eʎa komenθo a serbirles al atardeθer kwando se puso el sol le traian todos los emfermos i los endemonjados toda la θjudad estaba reunida a la pwerta i el sano a muʧos ke padeθian de dibersas emfermedades i eʧo fwera muʧos demonjos i no permitia a los demonjos ablar porke lo konoθian abjendose lebantado mui de madrugada todabia de noʧe xesus saljo i se fwe a un lugar desjerto i aʎi oraba simon i sus kompaɲeros fweron en buska de el lo enkontraron i le dixeron todos te buskan el les respondjo bamos a

otra parte a los pweblos beθinos para ke predike tambjen aʎi porke para esto e benido i fwe predikando en las sinagogas de eʎos en toda galilea i eʃʃando fwera los demonjos i bino a el un leproso implorandole i de rodiʎas le dixo si kjeres pwedes limpjarme xesus mobido a kompasjon estendjo la mano lo toko i le dixo kjero se limpjo i al instante desapareθjo la lepra de el i kedo limpjo en segida lo despidjo despwes de amoestarlo i le dixo mira no digas nada a nadje mas bjen be mwestrate al saθerdote i ofreθe lo ke mando moises en kwanto a tu purifikaθjon para testimonjo a eʎos pero kwando saljo el komenθo a proklamar i a difundir muʃʃo el eʃʃo de modo ke xesus ja no podia entrar abjertamente en ninguna θjudad sino ke se kedaba afwera en lugares depoblados i benian a el de todas partes

Appendix B

ACCDIST materials

B.1 Example sentences used in non-word nativisation

B.1.1 Sentences used in pilot

English

The ___ plates are cheap

I want a ___ picture frame

I like the ___ pillow

I've bought a lovely ___ light

I prefer the ___

What do you think of the ___?

Do you like the ___?

How about the ___?

The ___ is comfy

Do you want a ___?

I like the ___

The ___ is pretty

The ___ are cheap

I want another ___

I chose the ___

Do you have the ___ in blue?

Is this ___ what you wanted?

Let's try the ___

That ___ would look good in my room

Do you have a larger ___?

I like this ___ chair

The ___ is the right size

Do you have any ___ left?

The ___ is nice

I prefer the red ___

Spanish

Los platos __ son baratos

Quiero un marco de fotos __

Me gusta la almohada __

He comprado una preciosa lámpara __

Prefiero la __

¿Qué opinas del __ ?

¿Te gusta el __?

¿Qué te parece __?

El __ es cómodo

Quieres un __?

Me gusta la __

La __ es bonita

El __ es barato

Me gustaría otra __

Escogí el __

¿Tienen la __ en azul?

¿Es este __ lo que quería?

Probamos la __

Esa __ se vería bien en mi habitación

¿Tienen una __ más grande?

Me gusta esta __ silla.

El __ es de tamaño adecuado

Te quedan alguno __?

La __ es agradable

Prefiero el __ rojo

Japanese

__ 皿はすごく安いです

__ ピクチャーフレームを頂きたい

__ 枕が好きです

素敵な____ライトを買ってきました

__ の方が好きです

__ はどう思いますか?

__ はいかがですか?

__ はどうですか?

__ は気持ち良い

__ がほしいですか?

__ も好きです

__ が可愛いです

__ もとても安いです

もう一つの__ が頂きたい

__ を選びます

青色の__ がありますか?

欲しかったのはこの__ ですか?

__ を試してみよう

その__ は私の部屋と似合う

もっと大きい__ がありますか?

この__ 椅子が好きです

__ のサイズがちょうどいい

まだ__ がありますか?

__ がいいです

赤い__ の方が好きです

B.1.2 Sentences used in full study

Each sentence has six syllables preceding and six syllables following the non-word. The participants were given sentences in random order, and heard each sentence approximately twice

over the course of the study.

Spanish

Preferiría un ___ de madera blanda.
Me gustaría el ___ de madera dura.
Este, ¿tienen algún ___ de madera dura?
Estoy buscando un ___ mucho más barato.
Quiero comprar un ___ mucho más pequeño.
Este, ¿tienen algún ___ mucho más pequeño?
Estoy buscando un ___, que es más pequeño.
Quiero comprar un ___ un poco más alto.
Este, ¿tienen algún ___ un poco más alto?
Estoy buscando un ___ un poco más corto.
Estoy buscando un ___ un poco más grande.
Quiero comprar un ___ un poco más grande.
Este, ¿tienen algún ___ un poco más grande?
Quiero un nuevo ___ verde azulado.
No me gusta este ___, es demasiado grande.
No me gusta este ___, no es muy bonito.
Me gustaría el ___, pero es muy caro.
Quiero comprar un ___, pero es muy caro.
Me gusta mucho el ___, y es muy barato.
Me gusta mucho el ___. ¿Y tú qué opinas?
No me gusta este ___. ¿Y tú qué opinas?
Compraré un nuevo ___. El viejo se rompió.
Preferiría un ___. Es minimalista.
He encontrado un ___. Es muy agradable.
Me gusta mucho el ___. Es muy agradable.
Me gustaría el ___. Es muy agradable.
He encontrado un ___. Sé que quieres otro.
¿Que opinas de la ___? Es grande y azul.
¿Que opinas de la ___? Es pequeña y gris.
¿Que opinas de la ___ de madera dura?

Greek

Πόσο έχει ένα ___; Αυτό στη βιτρίνα.
Θα ήθελα το μπλε ___ που είναι στο ράφι.
Θα ήθελα το γκρι ___ που είναι στο ράφι.
Θα ήθελα ένα ___ σε άλλο μέγεθος.
Θα ήθελα ένα ___ λίγο πιο μεγάλο.
Κλίνω προς το μαύρο ___, τι λέτε και εσείς;
Κλίνω προς το άσπρο ___, τι λέτε και εσείς;
Έχετε καθόλου ___; Μου έχει τελειώσει.
Θέλω να αγοράσω ___. Το έχετε εσείς;
Πόσο κάνει ένα ___; Θα πάρω μερικά.
Μ' αρέσει πολύ το ___. Το έχετε μήπως;
Έχετε πιο φθηνό ___ ; Δε διαθέτω τόσα.
Μ' ενδιαφέρει ένα ___. Μπορώ να κοιτάξω;
Θέλω κι άλλο ένα ___. Έχουν μείνει άλλα;
Τελικά το άλλο ___ μ' άρεσε πιο πολύ.
Έχετε αυτό το ___ σε άλλο μέγεθος;
Έχετε αυτό το ___ σε άλλα χρώματα;
Έχετε πιο μικρό ___ για να δοκιμάσω;
Αυτό το μαύρο ___ σου πηγαίνει πολύ.
Υπάρχει πιο φθηνό ___; Μήπως σε έκπτωση;

Ωραίο αυτό το ___. Τι τιμή έχει;

Μου δίνετε ένα ___; Άσπρο αν υπάρχει.

Έμειναν καθόλου ___; Ψάχνω και δε βρίσκω.

Το προηγούμενο ___ έστρωνε ωραία.

Θα προτείνατε το ___; Ή μήπως κάτι άλλο;

English

I prefer the navy ___ to the dark purple one.

I want another blue ___ to go with my old one.

I prefer the smaller ___ to the really big one.

I really like the oak ___, or maybe the walnut.

Do you have a little ___ in dark blue or purple?

I prefer the bigger ___ to the really small one.

I prefer the bigger ___, over there on the left.

I'd like to buy a new ___, mine is getting too old.

I reckon a smaller ___ would fit in the kitchen.

I prefer the orange ___ to the bright yellow one.

I reckon a purple ___ would look good in my room.

I reckon an oval ___ would work well in the hall.

Do you think the navy ___ suits me at all, or not?

That's a really pretty ___. I think I'll buy one.

I prefer the larger ___ to the one you're holding.

I prefer the smaller ___, over there on the right.

I really like the red ___, do you want to buy one?

I reckon a narrow ___ would look good in the hall.

I think I prefer the ___, which one do you prefer?

Do you have a smaller ___? This one is a bit large.

Do you have a larger ___? This one is a bit small.

I'm not sure about this ___, is there a bigger one?

I'm not sure about this ___, is there a yellow one?

Do you have a larger ___ in light blue or turquoise?

I reckon a turquoise ___ would go with the bathroom.

I'm not sure about this ___, is there a smaller one?

There's a problem with my ___, I need to replace it.

I don't know if a big ___ would look good in my room.

I need a small one. This ___ is about the right size.

I really like the silk ___, but the cotton's cheaper.

I reckon a dark blue ___ would match the living room.

Could you help me find a ___? I'd like a chestnut one.

I don't know if the big ___ would fit in the bathroom.

I prefer the dark green ___ to the one you're holding.

I don't know if a pink ___ would look good in the hall.

I'm not sure about this ___, is there a light blue one?

It's quite expensive, but this ___ is really beautiful.

Could you help me reach the ___ on the shelf over there?

I don't know if a beige ___ would look good in the hall.

Could you help me reach the ___? It's on the upper shelf.

I really like the narrow ___, but the wide one's nice too.

I'd quite like a dark red ___ , like the one he's holding.

Could you pass me a blue ___? There are some on that shelf.

I really like the yellow ___, but the green one's nice too.

Could you pass me a square ___? There are some on that shelf.

I'd quite like a light grey ___ , like the one she's holding.

German

Hättest du gerne ein ___ oder was anderes?

Wir alle mögen das ___ das angeberisch ist.

Ich habe ein nettes ___ das auch so gelb ist.

Ich habe ein schönes ___ das angeberisch ist.

Ich habe ein tolles ___ das auch so grün ist.

Ich hätte gerne ein ___ das angeberisch ist.

Ich mag das hölzerne ___ das angeberisch ist.

Wir alle möchten ein ___ das angeberisch ist.

Wir alle mögen das ___ das auch so teuer ist.

Also, mögt ihr dieses ___ das angeberisch ist?

Die hat so ein tolles ___ das angeberisch ist.

Er hat ein sehr rotes ___ das angeberisch ist.

Hättest du gerne ein ___ oder eher doch nicht?

Wir alle mögen das ___ mit dem man bauen kann.

Wir haben ein grosses ___ das angeberisch ist.

Er mag am liebsten das ___ das angeberisch ist.

Er mag am liebsten das ___ das eben so rot ist.

Ich habe ein altes ___ mit dem man backen kann.

Ich habe ein neues ___ mit dem man gucken kann.

Ich hätte gerne ein ___ das auch so teuer ist.

Ich mag das hölzerne ___ das auch so teuer ist.

Sie hat ein hellblaues ___ das angeberisch ist.

Wir alle lieben das ___ das auch so billig ist.

Wir alle lieben das ___ mit dem man malen kann.

Wir alle wollen ein ___ das auch so sauber ist.

Wir alle wollen ein ___ mit dem man malen kann.

Also, mögt ihr dieses ___ das auch so teuer ist?

Ich habe ein nettes ___ das so ähnlich aussieht.

Ich hätte gerne ein ___ das auch so billig ist.

Ich hätte gerne ein ___ mit dem man bauen kann.

Ich hätte gerne ein ___ mit dem man malen kann.

Ich mag das hölzerne ___ das auch so billig ist.

Ich mag das hölzerne ___ mit dem man bauen kann.

Ich mag das hölzerne ___ mit dem man malen kann.

Wir alle lieben das ___ das so ähnlich aussieht.

Wir alle möchten ein ___ das auch so putzig ist.

Wir alle möchten ein ___ mit dem man bauen kann.

Wir alle wollen ein ___ das so ähnlich aussieht.

Wir haben ein grosses ___ das auch so teuer ist.

Also, mögt ihr dieses ___ das auch so billig ist?

Also, mögt ihr dieses ___ mit dem man bauen kann?

Also, mögt ihr dieses ___ mit dem man malen kann?

Die hat so ein tolles ___ mit dem man bauen kann.

Er hat ein sehr pinkes ___ das auch so teuer ist.

Er mag am liebsten das ___ das auch so teuer ist.

Glaubst du du magst so ein ___ oder was anderes?

Ich hätte gerne ein ___ das so ähnlich aussieht.

Ich mag das hölzerne ___ das so ähnlich aussieht.

Sie hat ein hellrotes ___ das eben so schwer ist.

Wir alle mögen das ___ mit dem man zeichnen kann.

Wir haben ein grosses ___ das auch so billig ist.

Wir haben ein grosses ___ mit dem man bauen kann.

Wir haben ein grosses ___ mit dem man malen kann.

Die hat doch ein tolles ___ das auch so teuer ist.

Er hat ein sehr gelbes ___ das auch so billig ist.

Er hat ein sehr rundes ___ mit dem man malen kann.

Er hat ein sehr weißes ___ mit dem man bauen kann.

Er mag am liebsten das ___ mit dem man bauen kann.

Er mag am liebsten das ___ mit dem man malen kann.

Sie hat ein hellblaues ___ das auch so leidig ist.

Sie hat ein hellgrünes ___ mit dem man bauen kann.

Sie hat ein hellgrünes ___ mit dem man malen kann.

Sie hat ein hellrotes ___ das so ähnlich aussieht.

Wir haben ein grosses ___ das so ähnlich aussieht.

Die hat auch ein tolles ___ mit dem man malen kann.

Die hat doch ein tolles ___ das auch so billig ist.

Er hat ein sehr blaues ___ das so ähnlich aussieht.

Er mag am liebsten das ___ das so ähnlich aussieht.

Ich habe ein schönes ___ mit dem man zeichnen kann.

Ich hätte gerne ein ___ mit dem man zeichnen kann.

Ich mag das hölzerne ___ mit dem man zeichnen kann.

Wann möchtest du dieses ___ haben? Morgen um eins?

Wir alle möchten ein ___ mit dem man zeichnen kann.

Also, mögt ihr dieses ___ mit dem man zeichnen kann?

Die hat auch ein tolles ___ das so ähnlich aussieht.

Wir haben ein grosses ___ mit dem man zeichnen kann.

Er hat ein sehr langes ___ mit dem man zeichnen kann.

Er mag am liebsten das ___ mit dem man zeichnen kann.

Glaubst du du magst so ein ___ oder eher doch nicht?

Sie hat ein hellgelbes ___ mit dem man zeichnen kann.

Die hat doch ein tolles ___ mit dem man zeichnen kann.

Ich würde auch gern ein ___ haben. Das wär schon nett..

TABLE B.1: Speaker IDs from Speech Accent Archive, accent.gmu.edu

Language	Sex	Id
Dutch	m	1
Dutch	m	2
Dutch	m	3
Dutch	f	8
Dutch	m	10
Dutch	f	39
Dutch	m	40
Dutch	m	43
English	m	13
English	f	306
English	m	365
English	m	368
English	m	465
English	f	487
English	m	496
French	f	1
French	m	13
French	m	21
French	m	39
French	m	43
French	m	46
French	f	53
French	f	60
Italian	m	2
Italian	f	4
Italian	m	7
Italian	m	8
Italian	m	11
Italian	m	19
Italian	m	26
Italian	f	29
Korean	f	2
Korean	f	3
Korean	f	6
Korean	m	11
Korean	f	16
Korean	f	22
Korean	m	44
Korean	f	46
Polish	m	5
Polish	m	7
Polish	m	8
Polish	f	15
Polish	m	22
Polish	m	23
Polish	m	25
Polish	m	27
Portuguese	f	11
Portuguese	m	20
Portuguese	f	27
Portuguese	m	29
Portuguese	f	39

B.2 Participant data

B.2.1 Speech Accent Archive

B.2.2 ACCDIST participants

TABLE B.2: English speakers

ID	Age	Sex	Place of birth	Native dialect	Other languages
2	26	M	London	London English	French - school
4	18	F	London	SSBE	French A-level
					German - basic
6	23	F	Leeds	SSBE	Italian - intermediate
7	19	M	London	London English	German C1
					French GCSE
8	19	M	London	London English	French - A level
					German - GCSE
					Spanish - beginner
9	19	F	London	English	Spanish - beginner
10	33	M	London	SSBE	French - GCSE
11	29	M	London	SSBE	French - A level
12	28	F	Portsmouth	SSBE	French - GCSE
13	28	M	Colchester	Essex English	None
15	25	M	Winchester	S. British English	None
16	22	F	Southampton	S. British English	French - A level
17	22	M	Hastings	S. British English	None
18	32	F	Leamington Spa	S. British English	Spanish - A level
					French - beginner
					German - beginner
19		M	-	-	
20		M	-	-	
21		M	-	-	
23	21	F	London	London English	French - GCSE
24	22	F	Milton Keynes	S. British English	French - GCSE
					Mandarin Chinese - GCSE
25	30	F	London	London English	German - beginner (school)

TABLE B.3: Spanish speakers

ID	Age	Sex	Place of birth	Native dialect	Other languages
1	32	F	Zaragoza, Aragon, Spain	Spanish (Aragon accent)	English - fluent
2	33	F	Zaragoza, Aragon, Spain	Spanish	English - fluent Catalan - intermediate French - beginner
4	27	F	Santiago, RM, Chile	Santiago Chilean Spanish	English - fluent
5	33	M	Santiago, RM, Chile	Santiago Chilean Spanish	English - fluent
6	22	F	Mexico City	Mexican Spanish	English - fluent
7		F			
8		F			

TABLE B.4: German speakers

ID	Age	Sex	Other languages spoken	Living abroad
1	21	F	English, French, Swedish	New Zealand (1 Year)
2	22	F	English	-
3	21	M	English	-
4	20	F	English, French, Italian, Spanish, Latin, Serbian	-
6	55	F	English	-
7	18	M	English	-
8	28	M	English, Japanese, French, Chinese	-
9	21	F	English	-
10	22	F	English, Japanese, French, Afrikaans, Korean	South Africa (4 Months) Japan (1 year)
11	24	F	English, French, Arabic	-
12	27	M	English	-
13	22	F	Italian, English	-
14	27	F	English, French, Russian, Turkish	-
15	24	F	English	-
17	24	F	Spanish, English	-
18	20	F	English, Portuguese, French, Spanish	Portugal
19	21	F	English, French	-
20	22	M	English, Spanish	-
21	24	M	English, French	-
23	34	M	English	London (1.5 years)
24	19	M	English, Spanish, Japanese, French	Argentina (1 Year)

TABLE B.5: Greek speakers

ID	Age	Sex	Place of birth	Native dialect	Other languages
1	32	F	Zakynthos	Southern Greek	English - fluent French - intermediate Italian - intermediate
2	31	F	Thessalonika	Northern Greek	English - fluent French - beginner
3	33	M	Pagra	Central Greek	English
4	23	F	Athens	Greek	English - proficient French - C1 Arabic - intermediate
5	19	F	Argos	Mainland Greek	English - advanced Hindi - beginner
6	35	M	Athens	Common Modern Greek (UNE)	English - advanced Spanish - advanced German - intermediate Turkish - intermediate

B.3 Scripts

B.3.1 Experimental files

The Psychopy files, audio files and (translated) instructions are available at

https://figshare.com/projects/Measuring_language_distance_-_non-word_adaptation/28506

B.3.2 Analysis code

The following commands run the Speech Filing System programs (Huckvale, 2008) required to process the audiofiles into MFCC inputs for ACCDIST.

1. Create SFS file per audiofile

```
hed -n [filename]
```

2. Link audiofile to SFS file

```
slink -iSP  -tWAV  -r  [audio filename]  [SFS filename]
```

3. Add word to SFS file

```
anload -T  [word]  [SFS filename]
```

4. Find the annotation; don't add silence; load non-English pronunciations from file; transcribe in SAMPA as default.

```
antrans -iAN^anload  -w  -x+[orthography to transcription file]  [SFS
    filename]
```

5. Find the transcription; it's in ARPA format; align it.

```
analign -iAN^antrans  -A  [SFS filename]
```

6. Now that the alignment is finished, change each phoneme to use original, not ARPA, transcription, and to have its context - the word it came from - as well

```
anload -t  word  -h  [new transcription] [SFS filename]
```

7. Calculate MFCCs

```
remove -aco  [SFS filename]
mfcc -H -n12 -e -l100 -h6000  [SFS filename]
```

8. Output language, gender, speaker ID, of vowel utterance with its two sets of MFCCs.

```
acntanal -A [language] -G [gender] -S [speaker] -v -2 -iC0^mfcc [SFS
    filename]
```

The Python code for calculating the ACCDIST metric from the MFCC file is also at https://figshare.com/projects/Measuring_language_distance_-_non-word_adaptation/28506.

B.3.3 SAMPA transcriptions of stimuli

Word in IPA	SAMPA transcription
kəne	k @ n e
kɛpa	k e p A:
kɑ̰nɛ	k A: n e
kɑ̃pɑ̰	k A: p A:
kapo	k A: p O
kɑ̰tu	k A: t u
kɔnɑ̰	k O n A:
ketɔ	k e t O
kʊ̈no	k U n O
kity	k i: t i:
kɨnə	k I n @
kœta	k @ t A:
konã	k O n A:
kupɑ̰	k u p A:
kuunʊ̈	k U n U
kyna	k i: n A:
nəti	n @ t i:
nɛtə	n e t @
nɑ̰pʊ̈	n A: p U
nãtã	n A: t A:
nasœ	n A: s @
nɑ̰py	n A: p i:
nɔtɑ̰	n O t A:
nepœ	n e p @
nʊ̈sy	n U s i:
nisɛ	n i: s e
nɨtɑ̰	n I t A:
nœpə	n @ p @
nopɨ	n O p I
nusɨ	n u s I
nuto	n U t O
nytɛ	n i: t e
pəka	p @ k A:
pɛso	p e s O
pɑ̰sa	p A: s A:
pɑ̃ni	p A: n i:
patɨ	p A: t I
pɑ̰sɑ̰	p A: s A:
pɔsɔ	p O s O
pese	p e s e

Word in IPA	SAMPA transcription
pʊ̈kɔ	p U k O
pinɨ	p i: n I
pɨsʊ̈	p I s U
pœsã	p @ s A:
potɯ	p O t U
putœ	p u t @
puusə	p U s @
pysu	p i: s u
səpɛ	s @ p e
sɛnɔ	s e n O
sɑ̰tʊ̈	s A: t U
saky	s A: k i:
sãki	s A: k i:
sɑ̰kʊ̈	s A: k U
sɔko	s O k O
sekɨ	s e k I
sʊ̈te	s U t e
sikã	s i: k A:
sɨku	s I k u
sœkœ	s @ k @
sokɑ̰	s O k A:
suke	s u k e
suukɛ	s U k e
sykɯ	s i: k U
təsi	t @ s i:
tɛkɑ̰	t e k A:
tɑ̰kə	t A: k @
tãsɯ	t A: s U
tanu	t A: n u
tɑ̰nɑ̰	t A: n A:
tɔpɔ	t O p O
tenɯ	t e n U
tʊ̈pã	t U p A:
tipi	t i: p i:
tɨpe	t I p e
tœnœ	t @ n @
tosɑ̰	t O s A:
tuny	t u n i:
tɯpɯ	t U p U
typu	t i: p u

B.4 Statistical data

TABLE B.6: ACCDIST Correlations between individual speakers: German with German, Greek

	German																				Greek					
	1	10	11	12	13	14	15	17	18	19	2	20	21	23	24	3	4	6	7	8	9	1	2	3	4	5	6
deu1	-	45	42	49	40	50	48	42	44	49	45	49	45	44	49	50	49	46	45	39	46	40	32	38	39	37	40
deu10	45	-	52	56	62	65	60	61	67	68	44	55	61	58	62	56	70	60	55	63	65	58	56	54	62	53	58
deu11	42	52	-	48	52	54	53	46	53	52	32	46	52	47	47	48	50	43	49	56	50	42	39	41	44	39	39
deu12	49	56	48	-	50	57	52	44	58	57	35	60	54	53	56	52	50	51	50	51	52	45	42	47	50	51	48
deu13	40	62	52	50	-	66	56	57	65	66	36	56	59	53	55	52	64	52	50	59	63	54	54	55	57	54	53
deu14	50	65	54	57	66	-	68	61	73	71	40	63	67	64	64	61	71	58	63	64	72	53	50	54	55	48	52
deu15	48	60	53	52	56	68	-	58	68	64	46	58	65	61	64	62	62	58	60	56	66	54	48	53	50	48	48
deu17	42	61	46	44	57	61	58	-	59	61	46	50	56	57	57	52	65	61	55	55	59	48	45	52	53	48	55
deu18	44	67	53	58	65	73	68	59	-	71	36	64	68	61	66	62	72	61	59	68	76	61	56	57	61	48	55
deu19	49	68	52	57	66	71	64	61	71	-	44	60	66	63	64	61	71	61	60	62	72	58	48	55	56	52	50
deu2	45	44	32	35	36	40	46	46	36	44	-	42	44	42	51	36	49	45	39	39	41	40	32	35	37	46	47
deu20	49	55	46	60	56	63	58	50	64	60	42	-	58	56	59	62	60	59	52	58	60	51	48	53	56	50	47
deu21	45	61	52	54	59	67	65	56	68	66	44	58	-	63	59	53	63	59	54	60	63	58	53	56	55	51	57
deu23	44	58	47	53	53	64	61	57	61	63	42	56	63	-	63	55	62	58	53	51	62	51	45	52	50	48	53
deu24	49	62	47	56	55	64	64	57	66	64	51	59	59	63	-	61	65	58	57	60	68	50	46	55	48	57	56
deu3	50	56	48	52	52	61	62	52	62	61	36	62	53	55	61	-	59	50	53	48	58	47	42	44	43	37	39
deu4	49	70	50	50	64	71	62	65	72	71	49	60	63	62	65	59	-	62	57	63	68	58	52	53	58	51	58
deu6	46	60	43	51	52	58	58	61	61	61	45	59	59	58	58	50	62	-	48	55	57	53	47	54	55	55	56
deu7	45	55	49	50	50	63	60	55	59	60	39	52	54	53	57	53	57	48	-	56	61	49	43	48	45	50	44
deu8	39	63	56	51	59	64	56	55	68	62	39	58	60	51	60	48	63	55	56	-	63	52	57	53	55	51	59
deu9	46	65	50	52	63	72	66	59	76	72	41	60	63	62	68	58	68	57	61	63	-	56	48	55	58	48	53

TABLE B.7: ACCDIST Correlations between individual speakers: German with English, Spanish

	English																					Spanish						
	2	3	4	6	7	8	9	10	11	12	13	15	16	17	18	19	20	21	23	24	25	1	2	4	5	6	7	8
deu1	41	43	34	42	34	44	40	35	32	37	34	25	43	36	43	39	35	35	34	42	40	44	46	41	37	40	43	39
deu10	60	56	49	59	53	61	57	52	50	52	47	43	54	51	62	56	61	53	52	54	47	52	63	57	53	54	53	49
deu11	44	45	50	51	37	48	48	42	47	36	37	34	49	40	44	47	48	41	45	46	45	41	53	40	42	40	41	51
deu12	49	52	52	46	47	54	48	47	46	50	40	34	55	42	45	50	47	50	47	58	49	48	54	48	43	47	48	46
deu13	52	54	53	57	48	64	60	55	53	51	50	44	55	54	59	54	59	54	55	57	49	46	56	52	51	45	53	49
deu14	56	56	52	57	49	62	55	50	51	50	46	39	59	53	59	58	59	51	47	58	49	52	59	53	48	50	53	48
deu15	59	55	53	56	52	61	50	51	50	52	47	40	58	51	58	57	53	49	49	54	54	58	60	52	47	50	54	47
deu17	54	51	46	52	44	58	50	45	48	46	44	38	56	45	54	44	53	49	45	52	50	50	53	53	53	48	53	45
deu18	62	57	63	63	57	69	64	60	57	55	52	51	62	62	63	61	66	54	55	61	55	51	63	55	50	55	61	51
deu19	59	56	54	60	54	60	58	50	47	50	48	38	57	51	58	52	57	51	52	57	48	51	61	52	49	52	56	50
deu2	43	44	31	40	34	40	29	30	31	35	33	26	38	24	34	32	27	32	24	39	36	53	49	47	45	41	39	36
deu20	50	56	52	50	42	57	52	49	47	55	44	40	57	44	48	58	51	48	42	56	49	50	62	53	47	48	54	49
deu21	58	58	55	59	52	60	58	51	51	50	47	38	61	51	57	53	57	53	51	60	52	51	59	53	52	49	58	50
deu23	57	53	49	53	49	60	48	44	43	47	44	36	51	48	55	47	52	46	43	53	44	51	59	50	48	52	54	49
deu24	55	59	52	53	55	64	48	47	55	55	49	47	59	54	58	55	53	49	46	55	46	56	58	55	52	49	56	41
deu3	50	45	46	50	46	53	51	43	42	42	41	37	53	46	49	47	52	41	42	52	44	46	53	46	35	39	40	38
deu4	60	56	52	53	51	60	57	47	45	48	45	38	57	49	55	52	57	48	43	58	49	52	59	55	49	56	54	51
deu6	52	53	48	46	45	60	53	47	46	49	44	39	54	41	50	50	49	50	44	57	45	52	58	52	55	55	56	47
deu7	46	53	46	46	41	54	48	41	44	45	39	36	51	41	49	49	47	47	43	48	45	48	56	46	43	41	48	44
deu8	59	51	51	56	41	61	57	51	51	45	49	40	50	48	55	54	57	51	45	50	49	46	58	54	50	49	54	49
deu9	56	58	60	58	54	62	56	51	55	55	51	46	59	60	59	55	60	54	51	57	52	54	60	55	51	53	57	47

TABLE B.8: ACCDIST Correlations between individual speakers: Greek, Spanish with German, Greek

	German																					Greek					
	1	10	11	12	13	14	15	17	18	19	2	20	21	23	24	3	4	6	7	8	9	1	2	3	4	5	6
ell1	40	58	42	45	54	53	54	48	61	58	40	51	58	51	50	47	58	53	49	52	56	-	54	56	58	54	59
ell2	32	56	39	42	54	50	48	45	56	48	32	48	53	45	46	42	52	47	43	57	48	54	-	55	59	51	60
ell3	38	54	41	47	55	54	53	52	57	55	35	53	56	52	55	44	53	54	48	53	55	56	55	-	60	59	62
ell4	39	62	44	50	57	55	50	53	61	56	37	56	55	50	48	43	58	55	45	55	58	58	59	60	-	55	56
ell5	37	53	39	51	54	48	48	48	48	52	46	50	51	48	57	37	51	55	50	51	48	54	51	59	55	-	64
ell6	40	58	39	48	53	52	48	55	55	50	47	47	57	53	56	39	58	56	44	59	53	59	60	62	56	64	-
spa1	44	52	41	48	46	52	58	50	51	51	53	50	51	51	56	46	52	52	48	46	54	50	43	58	52	57	49
spa2	46	63	53	54	56	59	60	53	63	61	49	62	59	59	58	53	59	58	56	58	60	59	58	59	65	60	61
spa4	41	57	40	48	52	53	52	53	55	52	47	53	53	50	55	46	55	52	46	54	55	54	50	55	59	55	56
spa5	37	53	42	43	51	48	47	53	50	49	45	47	52	48	52	35	49	55	43	50	51	55	54	60	59	59	64
spa6	40	54	40	47	45	50	50	48	55	52	41	48	49	52	49	39	56	55	41	49	53	52	54	56	56	52	60
spa7	43	53	41	48	53	53	54	53	61	56	39	54	58	54	56	40	54	56	48	54	57	56	48	59	54	53	53
spa8	39	49	51	46	49	48	47	45	51	50	36	49	50	49	41	38	51	47	44	49	47	54	51	54	58	49	58

TABLE B.9: ACCDIST Correlations between individual speakers: Greek, Spanish with English, Spanish

	English																				Spanish					
	1	10	11	12	13	14	15	17	18	19	2	20	21	23	24	3	4	6	7	8	9	1	2	3	4	5	6
ell1	40	58	42	45	54	53	54	48	61	58	40	51	58	51	50	47	58	53	49	52	56	-	54	56	58	54	59
ell2	32	56	39	42	54	50	48	45	56	48	32	48	53	45	46	42	52	47	43	57	48	54	-	55	59	51	60
ell3	38	54	41	47	55	54	53	52	57	55	35	53	56	52	55	44	53	54	48	53	55	56	55	-	60	59	62
ell4	39	62	44	50	57	55	50	53	61	56	37	56	55	50	48	43	58	55	45	55	58	58	59	60	-	55	56
ell5	37	53	39	51	54	48	48	48	48	52	46	50	51	48	57	37	51	55	50	51	48	54	51	59	55	-	64
ell6	40	58	39	48	53	52	48	55	55	50	47	47	57	53	56	39	58	56	44	59	53	59	60	62	56	64	-
spa1	44	52	41	48	46	52	58	50	51	51	53	50	51	51	56	46	52	52	48	46	54	50	43	58	52	57	49
spa2	46	63	53	54	56	59	60	53	63	61	49	62	59	59	58	53	59	58	56	58	60	59	58	59	65	60	61
spa4	41	57	40	48	52	53	52	53	55	52	47	53	53	50	55	46	55	52	46	54	55	54	50	55	59	55	56
spa5	37	53	42	43	51	48	47	53	50	49	45	47	52	48	52	35	49	55	43	50	51	55	54	60	59	59	64
spa6	40	54	40	47	45	50	50	48	55	52	41	48	49	52	49	39	56	55	41	49	53	52	54	56	56	52	60
spa7	43	53	41	48	53	53	54	53	61	56	39	54	58	54	56	40	54	56	48	54	57	56	48	59	54	53	53
spa8	39	49	51	46	49	48	47	45	51	50	36	49	50	49	41	38	51	47	44	49	47	54	51	54	58	49	58

TABLE B.10: ACCDIST Correlations between individual speakers: English with German, Greek

	German																					Greek					
	1	10	11	12	13	14	15	17	18	19	2	20	21	23	24	3	4	6	7	8	9	1	2	3	4	5	6
eng10	35	52	42	47	55	50	51	45	60	50	30	49	51	44	47	43	47	47	41	51	51	47	47	47	50	38	46
eng11	32	50	47	46	53	51	50	48	57	47	31	47	51	43	55	42	45	46	44	51	55	43	48	49	51	47	51
eng12	37	52	36	50	51	50	52	46	55	50	35	55	50	47	55	42	48	49	45	45	55	47	45	51	52	49	53
eng13	34	47	37	40	50	46	47	44	52	48	33	44	47	44	49	41	45	44	39	49	51	43	45	49	44	48	48
eng15	25	43	34	34	44	39	40	38	51	38	26	40	38	36	47	37	38	39	36	40	46	41	46	43	43	40	44
eng16	43	54	49	55	55	59	58	56	62	57	38	57	61	51	59	53	57	54	51	50	59	52	47	55	55	50	47
eng17	36	51	40	42	54	53	51	45	62	51	24	44	51	48	54	46	49	41	41	48	60	46	45	49	48	44	51
eng18	43	62	44	45	59	59	58	54	63	58	34	48	57	55	58	49	55	50	49	55	59	53	48	55	53	55	54
eng19	39	56	47	50	54	58	57	44	61	52	32	58	53	47	55	47	52	50	49	54	55	48	49	51	52	46	51
eng2	41	60	44	49	52	56	59	54	62	59	43	50	58	57	55	50	60	52	46	59	56	54	50	53	52	55	59
eng20	35	61	48	47	59	59	53	53	66	57	27	51	57	52	53	52	57	49	47	57	60	49	53	55	56	46	50
eng21	35	53	41	50	54	51	49	49	54	51	32	48	53	46	49	41	48	50	47	51	54	51	52	54	49	51	53
eng23	34	52	45	47	55	47	49	45	55	52	24	42	51	43	46	42	43	44	43	45	51	55	56	51	53	46	47
eng24	42	54	46	58	57	58	54	52	61	57	39	56	60	53	55	52	58	57	48	50	57	55	49	52	59	51	49
eng25	40	47	45	49	49	49	54	50	55	48	36	49	52	44	46	44	49	45	45	49	52	45	54	48	49	39	49
eng3	43	56	45	52	54	56	55	51	57	56	44	56	58	53	59	45	56	53	53	51	58	53	45	54	56	53	59
eng4	34	49	50	52	53	52	53	46	63	54	31	52	55	49	52	46	52	48	46	51	60	49	46	51	55	44	45
eng6	42	59	51	46	57	57	56	52	63	60	40	50	59	53	53	50	53	46	46	56	58	55	47	48	51	47	50
eng7	34	53	37	47	48	49	52	44	57	54	34	42	52	49	55	46	51	45	41	41	54	47	46	49	46	46	51
eng8	44	61	48	54	64	62	61	58	69	60	40	57	60	60	64	53	60	60	54	61	62	54	55	63	60	57	60
eng9	40	57	48	48	60	55	50	50	64	58	29	52	58	48	48	51	57	53	48	57	56	57	55	59	58	50	52

TABLE B.11: ACCDIST Correlations between individual speakers: English with English, Spanish

	English																				Spanish						
	10	11	12	13	15	16	17	18	19	2	20	21	23	24	25	3	4	6	7	8	9	1	2	4	5	6	7	8
eng10	-	64	45	48	54	54	58	52	59	53	62	53	50	54	54	50	61	52	56	63	51	35	49	40	45	46	53	40
eng11	64	-	52	48	54	54	55	51	52	50	58	57	52	50	50	55	58	51	51	61	48	38	53	43	52	39	44	42
eng12	45	52	-	45	44	52	45	52	49	47	51	59	48	56	52	55	60	45	58	56	43	46	54	50	50	49	53	47
eng13	48	48	45	-	44	47	48	48	48	49	48	47	39	48	45	43	48	50	44	59	45	39	47	48	44	42	45	36
eng15	54	54	44	44	-	39	50	45	46	43	47	48	45	41	43	42	46	43	50	55	42	32	47	40	45	37	39	33
eng16	54	54	52	47	39	-	54	61	59	60	61	50	54	61	53	58	65	59	52	56	59	46	59	53	46	47	54	46
eng17	58	55	45	48	50	54	-	59	53	59	56	49	47	48	43	50	57	56	57	55	48	39	53	45	46	49	47	42
eng18	52	51	52	48	45	61	59	-	55	61	63	54	52	55	47	60	52	61	51	60	59	43	59	56	51	53	59	51
eng19	59	52	49	48	46	59	53	55	-	58	56	52	46	51	49	54	57	54	49	64	53	44	53	47	46	46	54	44
eng2	53	50	47	49	43	60	59	61	58	-	52	46	46	58	47	62	54	59	54	60	53	50	55	55	54	57	56	48
eng20	62	58	51	48	47	61	56	63	56	52	-	59	57	50	54	54	59	58	51	66	63	44	60	51	49	47	53	48
eng21	53	57	59	47	48	50	49	54	52	46	59	-	54	52	50	52	56	48	51	56	55	43	52	48	50	47	48	47
eng23	50	52	48	39	45	54	47	52	46	46	57	54	-	52	55	50	59	52	51	53	57	39	48	43	48	45	49	45
eng24	54	50	56	48	41	61	48	55	51	58	50	52	52	-	57	55	57	53	48	56	58	44	58	55	50	46	45	52
eng25	54	50	52	45	43	53	43	47	49	47	54	50	55	57	-	50	53	47	49	54	49	43	54	53	43	43	48	47
eng3	50	55	55	43	42	58	50	60	54	62	54	52	50	55	50	-	55	55	53	56	52	53	59	52	52	47	54	51
eng4	61	58	60	48	46	65	57	52	57	54	59	56	59	57	53	55	-	57	59	55	56	43	58	47	47	47	46	47
eng6	52	51	45	50	43	59	56	61	54	59	58	48	52	53	47	55	57	-	53	58	55	45	60	51	47	48	46	47
eng7	56	51	58	44	50	52	57	51	49	54	51	51	51	48	49	53	59	53	-	54	45	44	53	46	46	49	49	40
eng8	63	61	56	59	55	56	55	60	64	60	66	56	53	56	54	56	55	58	54	-	58	48	62	56	58	54	60	49
eng9	51	48	43	45	42	59	48	59	53	53	63	55	57	58	49	52	56	55	45	58	-	45	57	53	49	50	56	53

TABLE B.12: Student's t-test between colingual correlation and cross-linguistic correlation

Colingual language	Other language	t	Degrees of freedom	p-value	p-value x 42	95% confidence interval		Mean colingual correlation	Mean cross-linguistic correlation
Dutch	Korean	7.85	108.91	3.1E-12	1.3E-10	0.07	0.12	0.67	0.58
Dutch	French	7.69	110.00	6.7E-12	2.8E-10	0.07	0.12	0.67	0.57
Portuguese	English	8.88	50.54	6.8E-12	2.9E-10	0.11	0.17	0.70	0.56
Portuguese	French	7.79	42.04	1.1E-09	4.6E-08	0.08	0.14	0.70	0.59
Korean	English	6.38	93.07	6.9E-09	2.9E-07	0.05	0.10	0.59	0.52
English	Korean	6.24	95.50	1.2E-08	4.9E-07	0.07	0.14	0.62	0.52
English	French	5.98	93.25	4.2E-08	1.8E-06	0.07	0.14	0.62	0.52
Polish	French	5.32	107.06	5.7E-07	2.4E-05	0.04	0.09	0.64	0.57
Portuguese	Korean	5.72	38.10	1.4E-06	5.8E-05	0.05	0.11	0.70	0.62
Portuguese	Italian	4.92	55.46	8.1E-06	3.4E-04	0.05	0.12	0.70	0.62
Portuguese	Polish	4.98	43.72	1.1E-05	4.5E-04	0.04	0.10	0.70	0.63
Polish	Korean	4.49	108.98	1.8E-05	7.6E-04	0.03	0.08	0.64	0.58
Dutch	English	4.22	116.07	4.8E-05	2.0E-03	0.03	0.09	0.67	0.61
Korean	French	4.16	87.28	7.5E-05	3.2E-03	0.02	0.07	0.59	0.55
Portuguese	Dutch	4.23	46.34	1.1E-04	4.5E-03	0.03	0.09	0.70	0.64
Dutch	Italian	3.69	116.05	3.4E-04	1.4E-02	0.02	0.08	0.67	0.62
French	English	3.68	85.35	4.0E-04	1.7E-02	0.02	0.06	0.56	0.52
English	Portuguese	3.37	92.95	1.1E-03	4.6E-02	0.02	0.10	0.62	0.56
Italian	French	3.05	99.44	3.0E-03	0.12	0.02	0.08	0.62	0.57
Korean	Portuguese	-3.05	70.81	3.2E-03	0.14	-0.05	-0.01	0.59	0.62
Italian	English	2.93	112.81	4.1E-03	0.17	0.02	0.09	0.62	0.57
English	Italian	2.90	108.67	4.5E-03	0.19	0.02	0.09	0.62	0.57
Dutch	Polish	2.81	117.88	0.01	0.24	0.01	0.06	0.67	0.63
French	Portuguese	-2.79	56.47	0.01	0.30	-0.05	-0.01	0.56	0.59
Italian	Korean	2.62	106.63	0.01	0.42	0.01	0.08	0.62	0.58
Polish	English	2.44	115.10	0.02	0.68	0.01	0.06	0.64	0.60
Polish	Italian	2.30	117.94	0.02	0.98	0.00	0.06	0.64	0.61
Dutch	Portuguese	2.23	87.31	0.03	1.19	0.00	0.06	0.67	0.64
French	Korean	1.55	78.63	0.13	5.30	0.00	0.04	0.56	0.55
Korean	Dutch	1.54	96.00	0.13	5.33	0.00	0.04	0.59	0.58
French	Polish	-1.41	92.30	0.16	6.85	-0.03	0.01	0.56	0.57
English	Polish	1.22	108.05	0.22	9.42	-0.01	0.06	0.62	0.60
Korean	Italian	1.11	87.75	0.27	11.42	-0.01	0.04	0.59	0.58
French	Dutch	-1.09	87.10	0.28	11.71	-0.03	0.01	0.56	0.57
French	Italian	-1.03	84.01	0.30	12.76	-0.03	0.01	0.56	0.57
Italian	Polish	0.92	101.77	0.36	15.01	-0.02	0.05	0.62	0.61
English	Dutch	0.82	104.65	0.41	17.32	-0.02	0.05	0.62	0.61
Korean	Polish	0.81	95.91	0.42	17.64	-0.01	0.03	0.59	0.58
Polish	Portuguese	0.57	90.99	0.57	23.89	-0.02	0.03	0.64	0.63
Polish	Dutch	0.37	117.60	0.71	29.84	-0.02	0.03	0.64	0.63
Italian	Dutch	0.32	109.41	0.75	31.50	-0.03	0.04	0.62	0.62
Italian	Portuguese	0.23	92.40	0.82	34.31	-0.03	0.04	0.62	0.62

Bibliography

Adank, Patti et al. (2009). 'Comprehension of familiar and unfamiliar native accents under adverse listening conditions.' In: *Journal of Experimental Psychology: Human Perception and Performance* 35.2, p. 520.

Ahukanna, Joshua G. W., Nancy J. Lund and J. Ronald Gentile (1981). 'Inter-and Intra-Lingual Interference Effects In Learning a Third Language'. In: *The Modern Language Journal* 65.3, pp. 281–287.

Baayen, R Harald, Richard Piepenbrock and H van Rijn (1993). *CELEX*. URL: `http://celex.mpi.nl` (visited on 13/08/2013).

Backley, Phillip (2011). *Introduction to Element Theory*. Edinburgh University Press.

Baertsch, Karen S. (2002). 'An optimality theoretic approach to syllable structure: The split margin hierarchy'. English. PhD thesis, p. 254. ISBN: 9780493697871.

Ball, Martin J. (1984). 'Phonetics for Phonology'. In: *Welsh phonology: Selected readings*. University of Wales Press. Chap. 1, pp. 5–39.

Ball, Martin John and Glyn E Jones (1984). *Welsh phonology: Selected readings*. University of Wales Press.

Bardel, Camilla and Christina Lindqvist (2006). 'The role of proficiency and psychotypology in lexical cross-linguistic influence. A study of a multilingual learner of Italian L3'. In: *Atti del VI Congresso di Studi dell' Associazione Italiana di Linguistica Applicata, Napoli*, pp. 9–10.

Bartelt, Guillermo (1989). 'Language Shift among Arizona Yaquis'. In: *Anthropos*, pp. 239–243.

Bell, Alan (1978). 'Syllabic consonants'. In: *Universals of human languages*. Ed. by Joseph H. Greenberg. Vol. 2. Stanford, pp. 153–201.

Bent, Tessa and Ann R Bradlow (2003). 'The interlanguage speech intelligibility benefit'. In: *The Journal of the Acoustical Society of America* 114.3, pp. 1600–1610.

Benware, Wilbur A (1986). *Phonetics and phonology of modern German: An introduction*. Georgetown University Press Washington, DC.

Berent, Iris, Katherine Harder and Tracy Lennertz (2011). 'Phonological Universals in Early Childhood: Evidence from Sonority Restrictions'. In: *Language Acquisition* 18.4, pp. 281–293.

Bialystok, Ellen (2009). 'Bilingualism: The good, the bad, and the indifferent'. In: *Bilingualism: Language and cognition* 12.1, pp. 3–11.

Bialystok, Ellen, Gigi Luk and Ernest Kwan (2005). 'Bilingualism, biliteracy, and learning to read: Interactions among languages and writing systems'. In: *Scientific studies of reading* 9.1, pp. 43–61.

Bialystok, Ellen, Michelle M Martin and Mythili Viswanathan (2005). 'Bilingualism across the lifespan: The rise and fall of inhibitory control'. In: *International Journal of Bilingualism* 9.1, pp. 103–119.

Blevins, Juliette (1995). 'Syllable in Phonological Theory'. In: *The Handbook of Phonological Theory*. Ed. by John A. Goldsmith. Blackwell. Chap. 6, pp. 206–244.

Bolozky, Shmuel (2006). 'A note on initial consonant clusters in Israeli Hebrew'. In: *Hebrew Studies* 47.1, pp. 227–235.

Brugmann, Karl and Hermann Osthoff (1878). *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*.

Brysbaert, Marc, Matthias Buchmeier et al. (2011). 'The word frequency effect'. In: *Experimental psychology*.

Brysbaert, Marc and Boris New (2009). 'Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English'. In: *Behavior research methods* 41.4, pp. 977–990.

Burnage, G. (1990). *CELEX: A guide for users*. URL: `https://catalog.ldc.upenn.edu/docs/LDC96L14/` (visited on 15/04/2015).

Calabrese, Andrea (1998). 'Metaphony revisited'. In: *Rivista di linguistica* 10, pp. 7–68.

Cambridge Dictionary (2015). *Cambridge dictionaries online*. URL: `https://dictionary.cambridge.org/dictionary/english`.

Campbell, Lyle (1993). 'On proposed universals of grammatical borrowing'. In: *Historical linguistics 1989*. Ed. by Henk Aertsen and Robert J. Jeffers, pp. 91–109.

Campbell, Lyle (1998). *Historical linguistics: an introduction*. MIT Press.

Cenoz, Jasone (2001). 'The effect of linguistic distance, L2 status and age on cross-linguistic influence in third language acquisition'. In: *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives*. Ed. by Jasone Cenoz, Britta Hufeisen and Ulrike Jessner. Vol. 31. Multilingual Matters. Chap. 1, pp. 8–20.

Cenoz, Jasone, Britta Hufeisen and Ulrike Jessner, eds. (2001). *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives*. Vol. 31. Multilingual Matters.

Chitoran, Ioana (2002). 'A perception-production study of Romanian diphthongs and glide-vowel sequences'. In: *Journal of the International Phonetic Association* 32.2, pp. 203–222.

Cleary, J. G. and W. J. Teahan (1997). 'Unbounded Length Contexts for PPM'. In: *The Computer Journal* 40.2-3, pp. 67–75.

Clements, George N (1990). 'The role of the sonority cycle in core syllabification'. In: *Papers in laboratory phonology* 1, pp. 283–333.

Clements, George N (2009). 'Does sonority have a phonetic basis'. In: *Contemporary views on architecture and representations in phonological theory* 48, pp. 165–175.

Corder, Stephen Pit (1979). 'Language distance and the magnitude of the language learning task'. In: *Studies in Second Language Acquisition* 2.01, pp. 27–36.

Crothers, John (1978). 'Typology and Universals of Vowel Systems'. In: *Universals of human languages*. Ed. by Joseph H. Greenberg. Vol. 2. Stanford, pp. 93–152.

Crowley, Terry and Claire Bowern (2010). *An introduction to historical linguistics*. Oxford University Press.

Cruttenden, Alan (2014). *Gimson's pronunciation of English*. Routledge.

Cyran, Eugeniusz (2011). 'Laryngeal realism and laryngeal relativism: Two voicing systems in Polish?' In: *Studies in Polish Linguistics* 6.1, pp. 45–80.

Dam, Mark van (2004). 'Word final coda typology'. In: *Journal of Universal Language* 119, p. 148.

Dasgupta, Probal (2003). 'Bangla'. In: *The Indo-Aryan Languages*. Routledge. Chap. 9, pp. 351–390.

Davidson, Lisa and Kevin Roon (2008). 'Durational correlates for differentiating consonant sequences in Russian'. In: *Journal of the International Phonetic Association* 38.2, pp. 137–165.

Dell, François (1995). 'Consonant clusters and phonological syllables in French'. In: *Lingua* 95.1-3, pp. 5–26.

Dereeper, Alexis et al. (2008). 'Phylogeny. fr: robust phylogenetic analysis for the non-specialist'. In: *Nucleic acids research* 36.suppl_2, W465–W469.

Donohue, Mark et al. (2013). *World phonotactics database*. URL: `http://phonotactics.anu.edu.au` (visited on 09/02/2015).

D'Souza, Vijay A. (2015). 'Towards a phonology of Hrusso Aka'. MA thesis. Trinity, University of Oxford.

Duchon, Andrew et al. (2013). 'EsPal: One-stop shopping for Spanish word properties'. In: *Behavior research methods*, pp. 1–13.

Duñabeitia, Jon Andoni et al. (2010). 'SYLLABARIUM: An online application for deriving complete statistics for Basque and Spanish orthographic syllables'. In: *Behavior Research Methods* 42.1, pp. 118–125. ISSN: 1554-3528.

Eden, S Elizabeth (in press). 'Does Sylheti have consonant clusters?' In: *Language documentation and description*.

Ellis, N. C. et al. (2001). *Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh*. URL: `www.bangor.ac.uk/canolfanbedwyr/ceg.php.en`.

Felsenstein, J (1989). 'PHYLIP-Phylogeny Inference Package (Version 3.2)'. In: *Cladistics* 8, pp. 164–166.

Féry, Caroline (1991). 'German schwa in prosodic morphology'. In: *Zeitschrift für Sprachwissenschaft* 10.1, pp. 65–85.

Fishman, Joshua A (1977). '"Standard" versus "Dialect" in Bilingual Education: An Old Problem in a New Context'. In: *The Modern Language Journal* 61.7, pp. 315–325.

Flynn, Suzanne, Claire Foley and Inna Vinnitskaya (2004). 'The cumulative-enhancement model for language acquisition: Comparing adults' and children's patterns of development in first, second and third language acquisition of relative clauses'. In: *International Journal of Multilingualism* 1.1, pp. 3–16.

Gelly, G and J. L Gauvain (2017). 'Spoken Language Identification using LSTM-based Angular Proximity'. In: *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*.

Goad, Heather (2012). 'sC clusters are (almost always) coda-initial'. In: *The Linguistic Review* 29.3, pp. 335–373.

Goldsmith, John A (1976). *Autosegmental phonology*. Vol. 159. Indiana University Linguistics Club Bloomington.

Goldsmith, John A., ed. (1995). *The Handbook of Phonological Theory*. Blackwell.

Gomes, Inês and São Luís Castro (2003). 'Porlex, a lexical database in European Portuguese'. In: *Psychologica* (72), pp. 91–108.

Gordon, Matthew (2002). 'A factorial typology of quantity-insensitive stress'. In: *Natural Language & Linguistic Theory* 20.3, pp. 491–552.

Gordon, Matthew and Peter Ladefoged (2001). 'Phonation types: a cross-linguistic overview'. In: *Journal of Phonetics* 29.4, pp. 383–406.

Goswami, Arpita (2013). 'Simplification of CC Sequence of Loan Words in Sylheti Bangla'. In: *Language in India* 13.6.

Gray, Russell D, Simon J Greenhill and Robert M Ross (2007). 'The pleasures and perils of Darwinizing culture (with phylogenies)'. In: *To appear in Biological Theory* 2, p. 4.

Green, Antony Dubach (2003). 'Extrasyllabic consonants and onset well-formedness'. In: *The Syllable in Optimality Theory*. Ed. by Caroline Féry and Ruben van de Vijver. Cambridge University Press. Chap. 9, pp. 238–253.

Green, David W, Jenny Crinion and Cathy J Price (2007). 'Exploring cross-linguistic vocabulary effects on brain structures using voxel-based morphometry'. In: *Bilingualism: Language and Cognition* 10.2, pp. 189–199.

Greenberg, Joseph H (1966). *Language Universals*. Mouton, The Hague.

Greenberg, Joseph H., ed. (1978). *Universals of human languages*. Vol. 2. Stanford.

Gussenhoven, Carlos and Haike Jacobs (2005). *Understanding Phonology* (*Understanding Language*). Arnold.

Gussenhoven, Carlos and Haike Jacobs (2013). *Understanding phonology*. Routledge.

Gussmann, E. (2007). *The Phonology of Polish*. Oxford linguistics. OUP Oxford. ISBN: 9780199267477.

Hannahs, Stephen J (2013). *The phonology of Welsh*. Oxford University Press.

Harris, James W (1983). 'Syllable structure and stress in Spanish. a nonlinear analysis'. In: *Linguistic Inquiry Monographs* 8, pp. 1–158.

Harris, James W. (1984). 'Theories of phonological representation and nasal consonants in Spanish'. In: *Papers from the XIIth Linguistic Symposium on Romance Languages, University Park, April 1[illegible]3, 1982*. Current Issues in Linguistic Theory. John Benjamins Publishing Company. Chap. 10, pp. 153–168. ISBN: 9789027280176.

Harris, James W. and Ellen M. Kaisse (1999). 'Palatal vowels, glides and obstruents in Argentinian Spanish'. In: *Phonology* 16.2, pp. 117–190.

Harris, John (1994). *English Sound Structure*. Blackwell.

Harris, John (2006). 'The Phonology of Being Understood: Further Arguments against Sonority'. English. In: *Lingua: International Review of General Linguistics* 116.10, pp. 1483–94. ISSN: 0024-3841.

Hayes, Bruce (1995). *Metrical stress theory: Principles and case studies*. University of Chicago Press.

Hayes, Bruce (2008). *Introductory phonology*. Blackwell.

Heggarty, Paul, April McMahon and Robert McMahon (2005). 'From phonetic similarity to dialect classification: a principled approach'. In: *Perspectives on Variation*, pp. 43–91.

Henton, Caroline, Peter Ladefoged and Ian Maddieson (1992). 'Stops in the world's languages'. In: *Phonetica* 49.2, pp. 65–101.

Heuven, Walter J. B. van et al. (2014). 'SUBTLEX-UK: A new and improved word frequency database for British English'. In: *The Quarterly Journal of Experimental Psychology* 67.6, pp. 1176–1190.

Honeybone, Patrick (2005). 'Diachronic evidence in segmental phonology: the case of obstruent laryngeal specifications'. In: *The internal organization of phonological segments*. Vol. 77. Studies in Generative Grammar [SGG]. Walter de Gruyter, pp. 317–51. ISBN: 9783110890402.

Howell, Peter et al. (2017). 'Identification of fluency and word-finding difficulty in samples of children with diverse language backgrounds'. In: *International Journal of Language & Communication Disorders* 52.5, pp. 595–611. ISSN: 1460-6984.

Huckvale, Mark (2004). 'ACCDIST: a metric for comparing speakers' accents'. In:

Huckvale, Mark (2007). 'Hierarchical clustering of speakers into accents with the ACCDIST metric'. In: *16th International Congress of Phonetic Sciences*, pp. 1821–1824.

Huckvale, Mark (2008). *Speech filing system*. `http://www.phon.ucl.ac.uk/resources/sfs`.

Hughes, A., P. Trudgill and D. Watt (2013). *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles, Fifth Edition*. The English Language Series. Taylor & Francis. ISBN: 9781134663880.

Hunter, Georgia G and Eunice V Pike (1969). 'The phonology and tone sandhi of Molinos Mixtec'. In: *Linguistics* 7.47, pp. 24–40.

Hyman, Larry M (2008). 'Universals in phonology'. In: *The linguistic review* 25.1-2, pp. 83–137.

Interagency Language Roundtable (2015). (Visited on 22/02/2015).

International Phonetic Association (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Iosad, Pavel (2017). 'The phonologisation of redundancy: length and quality in Welsh vowels'. In: *Phonology* 34.1, pp. 121–162.

Itô, Junko (1989). 'A prosodic theory of epenthesis'. In: *Natural Language & Linguistic Theory* 7.2, pp. 217–259.

Iverson, Gregory K and Joseph C Salmons (2008). 'Germanic aspiration: phonetic enhancement and language contact'. In: *Sprachwissenschaft* 33.3, pp. 257–278.

Jakobson, Roman (1962). *Selected writings*. Mouton, The Hague.

Jones, Glyn E. (1984). 'The Distinctive Vowels and Consonants of Welsh'. In: *Welsh phonology: Selected readings*. University of Wales Press. Chap. 2, pp. 40–64.

Juola, Patrick (1998). 'Cross-entropy and linguistic typology'. In: *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pp. 141–149.

Kaushanskaya, Margarita and Viorica Marian (2009). 'The bilingual advantage in novel word learning'. In: *Psychonomic Bulletin & Review* 16.4, pp. 705–710.

Kaye, Jonathan (1990). ''Coda'licensing'. In: *Phonology* (7), pp. 301–330.

Kaye, Jonathan (1992). 'Do you believe in magic? The story of s+ C sequences'. In: *Working Papers in Linguistics and Phonetics* 2, pp. 293–313.

Kenstowicz, Michael (1972). 'Lithuanian Phonology'. In: 2.

Kessler, Brett (1995). 'Computational dialectology in Irish Gaelic'. In: *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*. Morgan Kaufmann Publishers Inc., pp. 60–66.

Keuleers, Emmanuel, Marc Brysbaert and Boris New (2010). 'SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles'. In: *Behavior research methods* 42.3, pp. 643–650.

Kiparsky, Paul (1979). 'Metrical structure assignment is cyclic'. In: *Linguistic inquiry* 10.3, pp. 421–441.

Kirk, Neil W. et al. (2014). 'No evidence for reduced Simon cost in elderly bilinguals and bidialectals'. In: *Journal of Cognitive Psychology* 26.6, pp. 640–648.

Knuth, Donald E (1973). *The Art of Computer Programming: Volume 3: Sorting and Searching*. Addison Wesley, Reading, MA.

Koehn, Philipp (2005). 'Europarl: A parallel corpus for statistical machine translation'. In: *MT summit*. Vol. 5, pp. 79–86.

Kostakis, Andrew (2017). 'Toward a Typology of Mid Vowels'. Manchester Phonology Meeting.

Ktori, Maria, Walter J B van Heuven and Nicola J Pitchford (2008). 'GreekLex: A lexical database of Modern Greek'. In: *Behavior Research Methods* 40.3, pp. 773–783.

Ladefoged, Peter and Ian Maddieson (1990). 'Vowels of the world's languages'. In: *Journal of Phonetics* 18.2, pp. 93–122.

Lass, Roger (1984). 'Vowel system universals and typology: prologue to theory'. In: *Phonology* 1, pp. 75–111.

Laufer, Batia and Stig Eliasson (1993). 'What Causes Avoidance in L2 Learning: L1-L2 Difference, L1-L2 Similarity, or L2 Complexity?' In: *Studies in Second Language Acquisition* 15.1, pp. 35–48.

Lees, Robert B (1953). 'The basis of glottochronology'. In: *Language*, pp. 113–127.

Lewis, M Paul and F Gary (2017). 'Ethnologue: Languages of the World, Twentieth edition'. In: ed. by Gary F. Simons and Charles D. Fennig.

List, Johann-Mattis (2017). 'Establishing a cross-linguistic database of phonetic notation systems'. 47th Poznań Linguistic Meeting.

Lombardi, Linda (2003). 'Second language data and constraints on Manner: explaining substitutions for the English interdentals'. In: *Second Language Research* 19.3, pp. 225–250.

Longobardi, Giuseppe and Cristina Guardiano (2009). 'Evidence for syntax as a signal of historical relatedness'. In: *Lingua* 119.11, pp. 1679–1706.

Longobardi, Giuseppe, Cristina Guardiano, Alessio Boattini et al. (2012). 'Phylogenetic reconstruction and syntactic parameters. Quantitative experiments on Indo-European'. In: *14th Diachronic Generative Syntax Conference*.

Longobardi, Giuseppe, Cristina Guardiano, Giuseppina Silvestri et al. (2013). 'Toward a syntactic phylogeny of modern Indo-European languages'. In: *Journal of Historical Linguistics* 3.1, pp. 122–152.

Luo, Lin et al. (2013). 'Bilingualism interacts with domain in a working memory task: Evidence from aging.' In: *Psychology and aging* 28.1, p. 28.

MacWhinney, Brian (2000). *The CHILDES Project: Tools for Analyzing Talk*.

Maddieson, Ian (1984). *Patterns of sounds*. Cambridge university press.

Maddieson, Ian (2011). 'Vowel quality inventories'. In: *The World Atlas of Language Structures Online. Max Planck Digital Library.*

Mahanta, Shakuntala (2008). *Directionality and locality in vowel harmony: With special reference to vowel harmony in Assamese*. Netherlands Graduate School of Linguistics.

Major, Roy C (2008). 'Transfer in second language phonology'. In: *Phonology and second language acquisition*. Vol. 36. Studies in Bilingualism. John Benjamins Publishing. Chap. 3, pp. 63–94.

Mandera, Paweł et al. (2014). 'Subtlex-pl: subtitle-based word frequency estimates for Polish'. In: *Behavior research methods*, pp. 1–13.

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

Masica, Colin P (1991). *The Indo-Aryan Languages*. Cambridge University Press.

Mateus, Maria Helena and Ernesto D'Andrade (1998). 'The Syllable Structure in European Portuguese'. In: *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada* 14, pp. 13–32.

Mateus, Maria Helena and Ernesto d'Andrade (2000). *The phonology of Portuguese*. Oxford University Press.

McMahon, April and Robert McMahon (2005). *Language classification by numbers*. Oxford University Press.

Mermelstein, Paul (1976). 'Distance measures for speech recognition, psychological and instrumental'. In: *Pattern recognition and artificial intelligence* 116, pp. 374–388.

Michael, Erica B and Tamar H Gollan (2005). 'Being and becoming bilingual: Individual differences and consequences for language production.' In: New York, NY: Oxford University Press. Chap. 19, pp. 389–407. ISBN: 9780198034612.

Mielke, Jeff (2005). 'Ambivalence and ambiguity in laterals and nasals'. In: *Phonology* 22.2, pp. 169–203.

Mielke, Jeff (2008). *The Emergence of Distinctive Features.*

Moran, Steven, Daniel McCloy and Richard Wright, eds. (2014). *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Mufwene, Salikoko S (2001). *The ecology of language evolution*. Cambridge University Press.

Nagarajan, Hemalatha (2014). 'Constraints through the ages: loanwords in Bangla'. In: *The EFL Journal* 5.1.

Nakhleh, Luay, Don Ringe and Tandy Warnow (2005). 'Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages'. In: *Language*, pp. 382–420.

Nakhleh, Luay, Tandy Warnow et al. (2005). 'A comparison of phylogenetic reconstruction methods on an Indo-European dataset'. In: *Transactions of the Philological Society* 103.2, pp. 171–192.

Nerbonne, John and Wilbert Heeringa (1997). 'Measuring dialect distance phonetically'. In: *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON-97)*.

New, B., C. Pallier et al. (2001). 'Une base de données lexicales du français contemporain sur internet: LEXIQUE'. In: *L'Année Psychologique* 101. `www.lexique.org` Version 3.80, pp. 447–462.

New, Boris, Marc Brysbaert et al. (2007). 'The use of film subtitles to estimate word frequencies'. In: *Applied Psycholinguistics* 28.04, pp. 661–677.

Nichols, Johanna (1992). *Linguistic diversity in space and time*. University of Chicago Press.

Odden, David (1995). 'Rules v. Constraints'. In: *The Handbook of Phonological Theory*. Ed. by John A. Goldsmith. Blackwell. Chap. 1, pp. 06–244.

Odden, David (2005). *Introducing Phonology*. Cambridge University Press.

Parker, Stephen George (2002). 'Quantifying the sonority hierarchy'. PhD thesis. University of Massachusetts at Amherst.

Parker, Steve (2012). 'Sonority distance vs. sonority dispersion–a typological survey'. In: *The Sonority Controversy*. Ed. by Steve Parker. Vol. 18. Walter de Gruyter. Chap. 4, pp. 101–165.

Pattanayak, Debi Prasanna (1966). *A controlled historical reconstruction of Oriya, Assamese, Bengali, and Hindi*. Ed. by Cornelis Hendrik van Schooneveld. Vol. 31. Linguarum Series Practica. Mouton.

Peirce, Jonathan W (2007). 'PsychoPy—psychophysics software in Python'. In: *Journal of neuroscience methods* 162.1, pp. 8–13.

Pinet, Melanie, Paul Iverson and Mark Huckvale (2011). 'Second-language experience and speech-in-noise recognition: Effects of talker–listener accent similarity'. In: *The Journal of the Acoustical Society of America* 130.3, pp. 1653–1662.

Protopapas, Athanassios et al. (2012). 'IPLR: An online resource for Greek word-level and sublexical information'. In: *Language resources and evaluation* 46.3, pp. 449–459.

Pulleyblank, Douglas (1996). 'Neutral vowels in Optimality Theory: A comparison of Yoruba and Wolof'. In: *The Canadian journal of linguistics* 41.4, pp. 295–347.

Rakow, Martin and Conxita Lleó (2011). 'Comparing cues of phrasing in German and Spanish child monolingual and bilingual acquisition'. In: *Intonational Phrasing in Romance and Germanic: Cross-linguistic and Bilingual Studies*. Ed. by C. Gabriel and C. Lleó. Hamburg studies on multilingualism. John Benjamins Publishing Company. Chap. 8, pp. 213–234. ISBN: 9789027219305.

Remijsen, Bert (2007). 'Lexical tone in Magey Matbat'. In: *LOT Occasional Series* 9, pp. 9–34.

Remijsen, Bert (2010). 'Nouns and verbs in Magey Matbat'. In: *Typological and Areal Analyses: Contributions from East Nusantara* (618). Ed. by M.C. Ewing M. Klamer, pp. 281–311.

Remijsen, Bert (2015). *Matbat_MageyDialect_2003_Lexicography*. 1998-2003 [text]. URL: `https://datashare.is.ed.ac.uk/handle/10283/796`.

Remijsen, Bert and Leoma Gilley (2008). 'Why are three-level vowel length systems rare? Insights from Dinka (Luanyjang dialect)'. In: *Journal of Phonetics* 36.2, pp. 318–344.

Rice, Keren (2002). 'Vowel place contrasts'. In: *Language universals and variation*. Ed. by Mengistu Amberber and Peter Collins. Greenwood Publishing Group, pp. 239–69.

Ridouane, Rachid (2008). 'Syllables without vowels: phonetic and phonological evidence from Tashlhiyt Berber'. In: *Phonology* 25.02, pp. 321–359.

Ringe, Don, Tandy Warnow and Ann Taylor (2002). 'Indo-European and Computational Cladistics'. In: *Transactions of the philological society* 100.1, pp. 59–129.

Ringe, Donald A (1992). 'On calculating the factor of chance in language comparison'. In: *Transactions of the American Philosophical Society*, pp. 1–110.

Rose, Yvan et al. (2006). 'Introducing Phon: A software solution for the study of phonological acquisition'. In: *Proceedings of the... Annual Boston University Conference on Language Development. Boston University Conference on Language Development*. Vol. 2006. NIH Public Access, p. 489.

Rothman, Jason (2011). 'L3 syntactic transfer selectivity and typological determinacy: The typological primacy model'. In: *Second Language Research* 27.1, pp. 107–127.

Round, Erich R. (2017). *The AusPhon-Lexicon project: Two million normalized segments across 300 Australian languages*. Paper presented at Poznań Linguistic Meeting 2017.

Rubach, Jerzy (1994). 'Affricates as strident stops in Polish'. In: *Linguistic Inquiry*, pp. 119–143.

Śa', Rāmeśvara (2001). *Synchronic Comparative Phonology of Bengali and German*. Pustak Bipani.

Selinker, Larry and Usha Lakshmanan (1992). 'Language transfer and fossilization: The multiple effects principle'. In: ed. by Susan M. Gass and Larry Selinker, pp. 197–216.

Shannon, Claude E (1948). 'A note on the concept of entropy'. In: *Bell System Tech. J* 27, pp. 379–423.

SIL (2008). *Phonology Assistant v3*. `http://www-01.sil.org/computing/pa/`.

Silverman, Daniel et al. (1995). 'Phonetic Structures in Jalapa Mazatec'. In: *Anthropological Linguistics* 37.1, pp. 70–88.

Singha, Kh. Dhiren and Md. Isthaque Ahmed (2016). 'Phonological Adaptations of Some English Loanwords in Sylheti'. In: *Language in India* 16.7.

SOAS Sylheti Project (2015). *SOAS Sylheti Project FLEx*. URL: `https://public.languagedepot.org/repositories/show/soas-syl-flex` (visited on 11/2015).

Society), Bible Society (British & Foreign Bible and Vereniging Het Nederlands Bijbelgenootschap (1996). *Greek Bible: Today's Greek Version*. Bible Society (British & Foreign Bible Society). ISBN: 9789607847010.

Steriade, Donca (1982). 'Greek prosodies and the nature of syllabification'. PhD thesis. Massachusetts Institute of Technology.

Steriade, Donca (2000). 'Phonetics in Phonology: The Case of Laryngeal Neutralization'. In: 3.

Stevens, Kenneth N and Samuel Jay Keyser (1989). 'Primary features and their enhancement in consonants'. In: *Language*, pp. 81–106.

Stibbard, Richard M and Jeong-In Lee (2006). 'Evidence against the mismatched interlanguage speech intelligibility benefit hypothesis'. In: *The Journal of the Acoustical Society of America* 120.1, pp. 433–442.

Tang, Kevin (2012). *A 61 million word corpus of Brazilian Portuguese film subtitles as a resource for linguistic research*. Tech. rep. University College London Working Papers in Linguistics.

Tang, Kevin and John Harris (In prep[a]). 'A phonemic lexicon of Lithuanian with word frequency estimates'.

Tang, Kevin and John Harris (In prep[b]). 'A phonemic lexicon of Polish with word frequency estimates'.

Tang, Kevin and John Harris (In prep[c]). 'A phonemic lexicon of Romanian with word frequency estimates'.

Teahan, W. J. (1999). 'An improved interface for probabilistic models of text'. In: *Internal report*. Lund, Sweden: Department of Information Technology, Lund University.

Teahan, W. J. (2000). 'Text Classification and Segmentation Using Minimum Cross-entropy'. In: *Content-Based Multimedia Information Access - Volume 2*. RIAO '00. Paris, France: Le Centre de hautes études internationales d'informatique documentaire, pp. 943–961.

Tucker, Benjamin V and Natasha Warner (2010). 'What it means to be phonetic or phonological: the case of Romanian devoiced nasals'. In: *Phonology* 27.2, pp. 289–324.

Vance, Timothy J (2008). *The sounds of Japanese*. Cambridge University Press.

Wang, Hongyan and Vincent J. van Heuven (2005). 'Mutual intelligibility of American, Chinese and Dutch-accented speakers of English.' In: *INTERSPEECH*, pp. 2225–2228.

Weinberger, Steven (2015). *Speech Accent Archive*. URL: `http://accent.gmu.edu`.

White, Geoffrey M, Francis Kokhonigita and Hugo Pulomana (1988). *Cheke Holo dictionary. PL*. Vol. 97. Pacific linguistics C. Dept. of Linguistics, Research School of Pacific Studies, Australian National University, Canberra.

Wiese, Richard (2000). *The phonology of German*. Oxford University Press on Demand.

Wiese, Richard (2001). 'The phonology of /r/'. In: *Distinctive feature theory*. Vol. 2. Phonology and Phonetics. Mouton de Gruyter. Chap. 9, pp. 335–368. ISBN: 9783110886672.

Wijngaarden, Sander J. van (2001). 'Intelligibility of native and non-native Dutch speech'. In: *Speech Communication* 35.1, pp. 103–113.

Williams, Briony, Rhys James Jones and Ivan Uemlianin (2006). 'Tools and resources for speech synthesis arising from a Welsh TTS project'. In: *Fifth Language Resources and Evaluation Conference (LREC), Genoa, Italy*.

Zec, Draga (1995). 'Sonority constraints on syllable structure'. In: *Phonology* 12.01, p. 85.

Zissman, Marc A (1996). 'Comparison of four approaches to automatic language identification of telephone speech'. In: *IEEE Transactions on speech and audio processing* 4.1, p. 31.