Southern African Linguistics and Applied Language Studies 2005, 23(4): 459–474
Printed in South Africa — All rights reserved

SOUTHERN AFRICAN LINGUISTICS AND APPLIED LANGUAGE STUDIES
EISSN 1727–9461

Phonetic analysis of Afrikaans, English, Xhosa and Zulu using South African speech databases

Thomas Niesler[1]*, Philippa Louw[2] and Justus Roux[2]

[1] *Department of Electronic Engineering, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa*

[2] *Centre for Language and Speech Technology, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa*

* *Corresponding author, e-mail: trn@dsp.sun.ac.za*

Abstract: We present a corpus-based analysis of the Afrikaans, English, Xhosa and Zulu languages, comparing these in terms of phonetic content, diversity and mutual overlap. Our aim is to shed light on the fundamental phonetic interrelationships between these languages, with a view to furthering progress in multilingual automatic speech recognition in general, and in the South African region in particular.

Introduction

We present a comparison between four South African languages, based on annotated speech databases gathered in the same manner for each language. The analysis is carried out mostly at phonetic level, using the transcriptions that form part of the databases. Both context-independent (monophone) and context-dependent (triphone) phonetic units are considered. The aim of this study is to advance the understanding of the phonetic character of and relationships between the languages spoken within a multilingual society. Ultimately, we hope that such insights will contribute the development of multilingual speech recognition technology in the southern African region.

Languages

South Africa has 11 official languages. Of these, Afrikaans, English, Xhosa and Zulu have been chosen to be the subject of this study, as they are widely spoken. In particular, Afrikaans, English, Xhosa and Zulu are mother tongues for 13.3%, 8.2%, 17.6% and 23.8% of the South African population, respectively (Statistics South Africa, 2004). Furthermore, Afrikaans and English have European roots, while Xhosa and Zulu are indigenous African languages. Hence, we will be in a position to investigate differences between languages that have common roots, as well as between languages that have distinct roots but are nevertheless linked by virtue of being spoken within the same multilingual society.

Afrikaans

Afrikaans is a Germanic language with its origins in 17th century Dutch as brought to the Cape of Good Hope by settlers from Holland in 1652. It is regarded as a continuation of acrolectal Cape Dutch, which was the language of the European rank and file during the Dutch East India Company (VOC) era (Roberge, 1995). The development of Afrikaans is characterised by the incorporation of lexical and syntactical borrowings from Malay, Bantu and Khoisan languages, as well as from Portuguese and a number of other European languages. It is spoken widely in South Africa, as well as in neighbouring states such as Namibia and Botswana. Approximately 6.2 million people speak Afrikaans as a mother tongue, whilst four million use it as a second or third language (Statistics South Africa, 2004).

South African English

The history of the English language in South Africa may be traced back to the first British occupation in 1795. The language was significantly strengthened by the arrival of 5 000 British settlers in the Eastern Cape in 1820

and the subsequent proclamation of English as an official language of the Cape Colony in 1822. South African English (SAE) is today regarded as a specific variety of so-called World Englishes, and is spoken by approximately 3.7 million people as a mother tongue (Statistics South Africa, 2004). As such, SAE comprises only the fifth largest group of mother tongue speakers, following Zulu, Xhosa, Afrikaans and Northern Sotho. However, it finds wide and general use in business and industry. SAE is also widely spoken in neighbouring countries such as Lesotho, Botswana, Namibia and Zimbabwe.

Xhosa

Both Xhosa and Zulu belong to the Bantu language family, which is prevalent throughout Africa south of the equator (Wentzel *et al*., 1972). The term 'Bantu' was introduced to language studies as far back as 1857 (Poulos & Msimang, 1998). Nguni is one of five language groups which comprise the South-Eastern Bantu languages. In turn, Nguni consists of a group of sister languages, which include Xhosa, Zulu, Ndebele and Swazi (Baily, 1995).

Xhosa is widely spoken in South Africa, with approximately 7.9 million mother tongue speakers. These are concentrated mainly in the Eastern and Western Cape regions, with other significant concentrations in Gauteng, KwaZulu-Natal, the Free State, Mpumalanga, the Northern Cape and the North-West Province (Statistics South Africa, 2004). European linguists studied Xhosa from as early as 1837. The first Xhosa grammar was published in 1850, and the orthography, as it is commonly accepted today, was standardised in 1962 (Wentzel *et al*., 1972).

Xhosa is a tone language with two inherent tones, low and high. Furthermore, it exhibits a wide range of 'exotic' sounds, such as an implosive bilabial stop, as well as an extensive range of clicks with a variety of phonetic accompaniments.

Zulu

Like Xhosa, Zulu is a member of the Nguni group of languages. Approximately 10.7 million South Africans speak Zulu as a mother tongue. These speakers are concentrated mainly in the KwaZulu-Natal province and in the urban areas of Gauteng, with other significant concentrations in Mpumalanga, the Free State, the North-West Province and the Eastern Cape (Statistics South Africa, 2004).

Missionaries and other grammarians such as Döhne (1857), Boyce (1863) and Grout (1893) were responsible for the development of the Zulu written form (Poulos & Msimang, 1998). It is a well-developed language with a growing literature and numerous celebrated writers. It shares many phonetic features with its sister language, Xhosa. Zulu and Xhosa have a conjunctive orthography, which implies that a whole sentence like *Ngiyambona* ('I see him/her') may be presented by a single word. The wide range of morphological and morphophonological possibilities associated with this type of orthography poses great challenges to the technological development of these languages.

The speech databases

We have based our experiments on the African Speech Technology (AST) databases, which consist of recorded and annotated speech collected over both mobile and fixed telephone networks (Louw *et al*., 2001; Roux *et al*., 2004). Speakers were recruited from each of the targetted language groups and given a unique datasheet, with items designed to elicit a phonetically diverse mix of read and spontaneous speech. The datasheets included read items such as isolated digits, as well as digit strings, money amounts, dates, times, spellings and also phonetically-rich words and sentences. Spontaneous items included references to gender, age, mother tongue, place of residence and level of education.

The AST databases were collected in five different languages, as well as in a number of non-mother tongue variations. In this work, we have made use of the Afrikaans, English, Xhosa and Zulu mother tongue databases.

Together with the recorded speech waveforms, both orthographic (word-level) and phonetic (phone-level) transcriptions were available for each utterance. The orthographic transcriptions were produced and validated by human transcribers. Initial phonetic transcriptions were obtained from the orthography, using grapheme-to-phoneme rules, except for English where a pronunciation dictionary was used instead. These were subsequently corrected and validated manually by human experts.

Phone set

A global set of 154 phones, based on the International Phonetic Alphabet, has been used during phonetic speech transcription (International Phonetic Association, 2001). These phone units were chosen primarily with their expected utility to speech recognition systems in mind. Hence, they have been chosen to reflect the variety of distinct sounds which are present in each language, and that will be important in modelling pronunciations.

The phone set includes labels for diphthongs and tripthongs which occur across word boundaries in spontaneous speech. Furthermore, in several cases, diacritic symbols, in combination with consonants or vowels, have been considered to be unique phones. The diacritics in question include ejection, aspiration, syllabification, voicing, devoicing and duration. The appendix presents a table of the phones which occur in the Afrikaans, English, Xhosa and Zulu databases.

Stop sounds

In Xhosa and Zulu, ejection and aspiration occur frequently and therefore the ejective and aspirated versions of a stop sound were included in the phone set. While ejection does not occur in Afrikaans or in English, aspiration may be present but is not phonemic, as it is in Xhosa and Zulu. From a speech recognition perspective, one might therefore argue that, because aspiration does not affect lexical meaning, it need not be modelled explicitly in pronunciations. For this reason, it was decided during the annotation of the AST databases that aspiration would not be transcribed for Afrikaans and English. Instead, the non-aspirated versions of the stop sounds were used exclusively in the transcriptions of these databases.

Since we will compare the phonetic content of the four languages by means of their phonetic transcriptions, we must ensure that the method of annotation is uniform over all corpora. In particular, it would not be correct to directly compare two corpora in which aspiration has been transcribed with two in which it has not. Therefore we have re-labelled all occurrences of the aspirated stops [pʰ], [tʰ] and [kʰ] with their non-aspirated, non-ejective counterparts [p], [t] and [k] in the Xhosa and Zulu data, so as to conform to the convention used for Afrikaans and English.

As a result, the phone set used in our analysis includes at least two versions of each stop sound e.g. [p] and [p'], and in some cases also devoiced and implosive versions e.g. [b̥] and [ɓ]. The aspirated stops [pʰ], [tʰ] and [kʰ] are included in the phone table presented in the appendix, for completeness.

Fricatives

The transcriptions of the four databases used in this work distinguish between 13 different fricatives. The voiced and voiceless alveolar lateral fricatives [ɮ] and [ɬ] are unique to Xhosa and Zulu and occur in the Afrikaans data only in place names. The same set of nine fricatives occurs in both Xhosa and Zulu, except for the voiceless velar fricative [x], which occurs only in Xhosa. The voiced and voiceless dental fricatives [θ] and [ð] are unique to English and occur in the other databases only in borrowed words or names.

Affricates

The African languages exhibit a significantly larger number of affricates than Afrikaans or English. The four databases we have used in this study are annotated, using nine distinct affricates, of which seven are unique to Xhosa and Zulu.

Clicks

Both Xhosa and Zulu use dental, lateral and palatal click sounds. These three basic clicks are extended to a total of 15 by including combinations of aspiration, nasalisation and voicing, which can be phonemic (Jessen & Roux, 2002).

Trills and flaps

Two trills and one flap were found in our data. The alveolar trill [r] was transcribed in all databases, although this sound occurs only in phonologised loanwords or borrowed words in English, Xhosa and Zulu. The uvular trill [ʀ] occurs only in Afrikaans and Xhosa. In the Afrikaans data, the alveolar flap [ɾ] is only found where assimilation occurs, for example in the words *vat dit* ('take it'), while in Xhosa and Zulu it is only found in borrowed words such as *quarter*.

Approximants

Four approximants are transcribed, of which two, the palatal [j] and the labio-velar [w], are

glides or semi-vowels. The alveolar approximant [ɹ] occurs only in the English language, and is therefore only found in the borrowed words and names of the other databases.

Nasals

Five different nasals occur in our data, and three of these occur in all four languages. The syllabic bilabial nasal [m̩] is unique to the African languages and is not used in Afrikaans or in English. The palatal nasal [ɲ] occurs in all but the English databases.

Vowels

A total of 29 different vowels is found in the four databases. For all except lax vowels, a longer-duration counterpart is included; for example, for the phone [i] in the Afrikaans word *siek* ('ill'), there is a phone [iː], as found in *vier* ('four'). Both lax and tense vowels are included in the phone set because lax vowels occur frequently in English, but not at all in Afrikaans or the African languages. Lax vowels in English are typically shorter, lower and slightly more centralised than the corresponding tense vowels (Ladefoged, 1975: 74). Examples are found in the words 'hi̱m' and 'pu̱t'.

Diphthongs and tripthongs

A total of 30 unique diphthongs has been identified and transcribed in the four databases. The English and Afrikaans diphthongs are similar, but no diphthongs are shared between the two languages, mainly due to the use of lax vowels in English. In the African languages, diphthongs occur only during code-switching[1] and are not intrinsic to the languages themselves.

A number of diphthongs which occur across word boundaries in spontaneous speech are also included in our phone set, for example in the Afrikaans words *drie-en-twintig* ('twenty-three'). Finally, three tripthongs occurring across word boundaries in spontaneous speech have been identified and transcribed. Examples include the Afrikaans words *hy is* ('he is') or the Afrikaans mother tongue pronunciation of the English word 'about'.

Training and test sets

Each database was divided into a training and a test set. The training sets were used to gather phone statistics, while the test sets were reserved for subsequent independent testing. Each of the four training sets were prepared to contain approximately six hours of audio data, as indicated in Table 1.

In Table 1, 'phone types' refers to the number of different phones which occur in the data, while 'phone tokens' indicates their total number. Note that in the training sets, a slightly lower speech rate was observed for Xhosa and Zulu compared with English and Afrikaans, resulting in the smaller numbers of phone tokens.

Each test set was designed to contain approximately 25 minutes of speech data, as shown in Table 2. There was no speaker over-

Table 1: Training sets for each database

Database	Speech (hours)	No. of speakers	Phone types	Phone tokens
Afrikaans	5.99	235	82	171 189
English	5.98	245	73	169 512
Xhosa	6.03	191	104	154 705
Zulu	6.19	202	91	157 239

Table 2: Test sets for each database

Database	Speech (mins)	No. of speakers	No. of phone tokens
Afrikaans	24.5	18	10 828
English	24.2	18	10 685
Xhosa	25.4	17	10 396
Zulu	24.9	16	10 067

lap between the test and training sets. Each contained both male and female speakers.

Because the test and training sets are disjoint, the former may be considered to represent 'unseen data', relative to the latter. Hence, for example, a test set may contain phones or combinations of phones which do not occur in the corresponding training set. We have employed the test sets as validation metrics in order to ensure objectivity in our subsequent phonetic comparisons.

Non-speech sounds

In addition to phones representing speech sounds, periods of silence, as well as noises produced by the speaker (like lip smacks and coughs) and background noises, were labelled in all databases. When background noises occur while a word is being spoken, the utterance is in some cases considered to be unintelligible, and marked accordingly.

Phonetic corpus analysis

This section presents a comparative analysis of the phonetic content and character of each of the four language databases under study.

Monophone statistics

We first investigate the extent to which the most frequent monophone types in each language's training set cover the monophone tokens in the same language's test set. Figure 1 illustrates how this coverage changes as a function of the number of most frequent monophone types in the training set. Noise and silence phones, as described in the section entitled 'Non-speech sounds', are excluded from this analysis.

It is evident from the graph that a larger number of phones is required for near full coverage of the test set in Xhosa and Zulu than in Afrikaans and English. For example, 99% coverage of the Afrikaans and English test sets can be achieved by retaining the 34 and 39 most frequent monophones, respectively. For Xhosa and Zulu, on the other hand, 56 and 62 monophones, respectively, must be retained to achieve the same coverage. Hence, not only do the Xhosa and Zulu training sets contain a comparatively large number of phone types (as is evident from Table 1), but the least frequent phones in Xhosa and Zulu also occur relatively more frequently than their counterparts in English

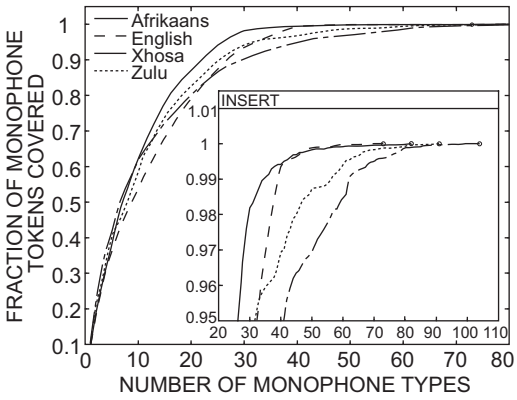

Figure 1: Coverage of training set monophone tokens by the most frequent monophone types

and Afrikaans. This indicates that a greater variety of sounds are in active use by speakers of Xhosa and Zulu than by speakers of English and Afrikaans. It also implies that Xhosa and Zulu will require a substantially larger phone set for accurate phonetic modelling in speech recognition applications.

We now turn our attention to the degree to which each language's set of monophones covers the test sets of the other languages. Figures 2–5 illustrate the degree to which the most frequent Afrikaans, English, Xhosa and Zulu monophone types respectively cover the monophone tokens in each of the four test sets.

From Figure 2, it is apparent that Afrikaans monophones sometimes cover the Xhosa and the Zulu test sets to a larger degree than the English test set. For example, when retaining the 40 most frequent monophones in the Afrikaans phone set, 80% of the phones in both the Xhosa and Zulu test sets but only 68% of the phones in the English test set are covered. However, from Figure 3, it is evident that English monophones practically always cover the Afrikaans test set to a greater extent than they cover the Xhosa and Zulu test sets. It appears therefore that there is often a greater phonetic commonality between Afrikaans and the two African languages than between Afrikaans and English. Closer inspection of the data reveals that this is largely due to the vowels [a],[i],[u] and [ɔ], which are frequent in Afrikaans and very frequent in Xhosa and Zulu, but infrequent in English. These four tense vowels are used often in Afrikaans, Xhosa and Zulu

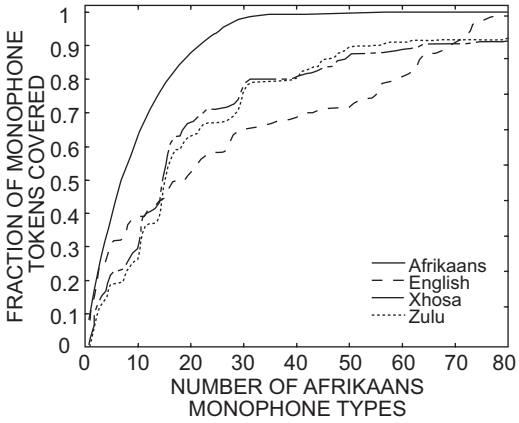

Figure 2: Coverage by the most frequent Afrikaans monophone types of the test set monophone tokens in each language

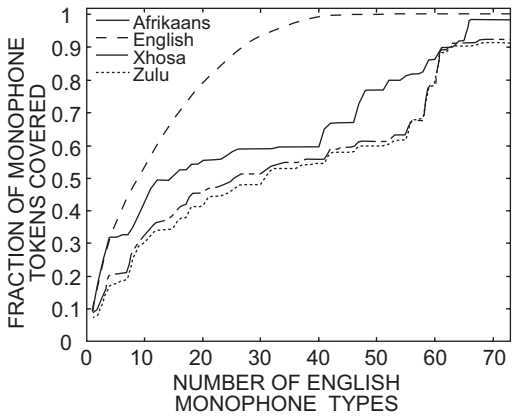

Figure 3: Coverage by the most frequent English monophone types of the test set monophone tokens in each language

but infrequently in English, due to the preference for lax vowels.

From Figure 4, we see that Xhosa monophones are practically as good at covering the Zulu test set as they are at covering the Xhosa test set. Furthermore, Figure 5 indicates that the Zulu monophones cover the Xhosa test set almost as effectively as the Xhosa monophones covered the Zulu test set. Hence, from a monophone viewpoint, the Xhosa and Zulu training sets show much greater similarity to each other than they do to English and Afrikaans. Finally, looking again at Figure 3 and comparing it with Figures 2–5, we see that out

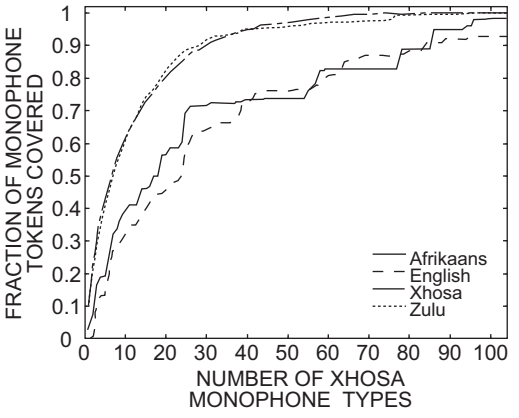

Figure 4: Coverage by the most frequent Xhosa monophone types of the test set monophone tokens in each language

of the four languages, the English monophone set is worst at covering the monophones present in the test sets of the other languages.

Triphone statistics

The variety of sequential combinations in which phones occur within a language directly impacts on the variety of coarticulatory effects that may be expected. From the view of speech recognition, this is important because state-of-the-art recognition technology makes use of context-dependent acoustic units in order to model this coarticulation. Hence, we extended the analysis performed for context-independent monophone units in the previous section to context-dependent triphones.

Proceeding as we had done for monophones, we first investigated how the most frequent triphone types in each language's training set covered those in the same language's test set. These results are shown in Figure 6, and form the context-dependent counterpart of Figure 1.

From Figure 6, we see that English has by far the largest number of triphone types in its training set (approximately 11 300), followed by Xhosa (8 800), Afrikaans (7 800) and Zulu (6 400). However, from Table 1 we recall that English also has the smallest number of monophone types (73) of all four languages. This means that the variety of sequential combinations in which English phones occur is much larger than it is for the other languages. Furthermore, Figure 6 shows that Zulu has the

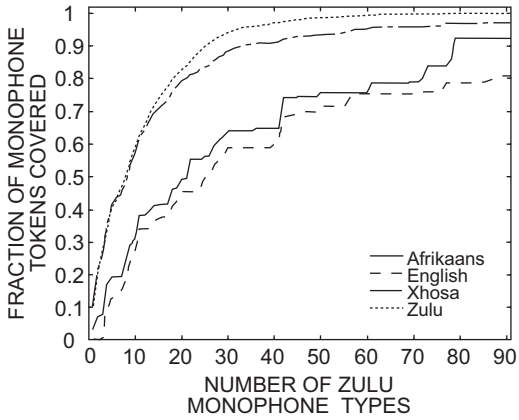

Figure 5: Coverage by the most frequent Zulu monophone types of the test set monophone tokens in each language

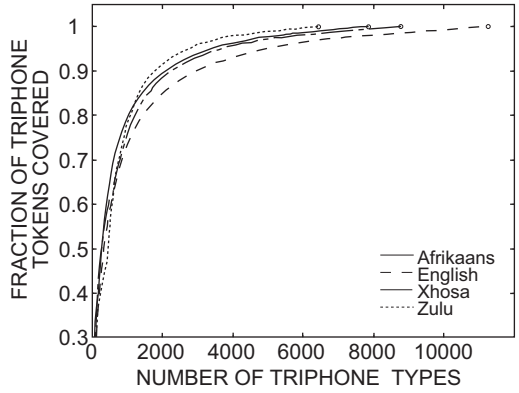

Figure 6: Coverage of training set triphone tokens by most frequent triphone types

smallest number of triphone types, although from Table 1 and Figure 1 we see that it contains the second largest number of monophone types. Xhosa also exhibits a relatively small number of triphone types in relation to its large number of monophone types. This indicates that the variety of sequential combinations in which Xhosa and Zulu phones occur is more highly constrained. This same conclusion will be reached from a different angle in the section entitled 'Language models'.

Now, as for monophones in the previous section, we investigated the degree to which the triphone types of each language covered those in the test sets of the other languages. Figures 7–10 indicate the fraction of triphone tokens in

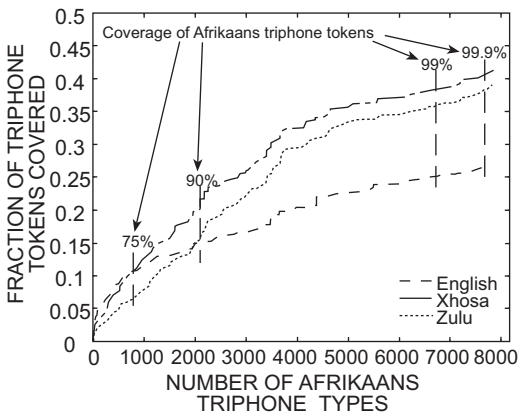

Figure 7: Coverage of test set triphone tokens by the most frequent Afrikaans triphone types in each language. No curve is shown for Afrikaans, which would require a different scale (see Figure 6). Instead, the 75%, 90%, 99% and 99.9% coverage points for Afrikaans are indicated by vertical dashed lines

each language's test set that is covered by the most frequent Afrikaans, English, Xhosa and Zulu triphone types, respectively.

From Figure 7 we see that Afrikaans triphone types exhibit poor coverage of all three other languages. For example, the 6 700 most frequent Afrikaans triphones cover 99% of the Afrikaans test set, but only 38%, 36% and 25% of the Xhosa, Zulu and English test sets, respectively. As was the case for monophones, Afrikaans triphones are often better able to cover the Xhosa and Zulu test sets than the English test set. Figure 8 reveals that English triphones are also poor at covering other languages. For example, the 9 400 most frequent English triphones cover 99% of the English test set but only 26%, 17% and 13% of the Afrikaans, Xhosa and Zulu test sets, respectively. However, the language best covered by the English triphones is Afrikaans, as was the case for monophones (Figure 3).

Finally, Figures 9 and 10 show that Xhosa and Zulu triphones exhibit a much better coverage of each other's test sets than those of Afrikaans and English. In particular, the complete set of Xhosa triphone tokens covers 90% of the triphones in the Zulu test set, and the complete set of Zulu triphone tokens covers 84% of the triphones in the Xhosa test set.

With a view to automatic speech recognition, the significant overlap between Xhosa and

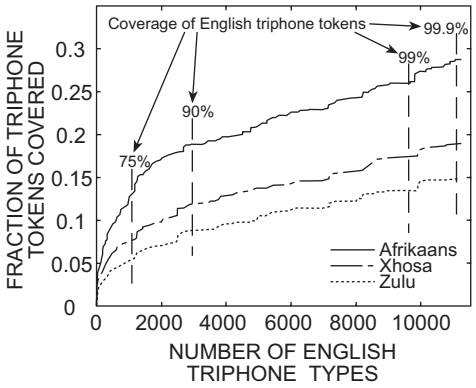

Figure 8: Coverage of test set triphone tokens by the most frequent English triphone types in each language. No curve is shown for English, which would require a different scale (see Figure 6). Instead, the 75%, 90%, 99% and 99.9% coverage points for English are indicated by vertical dashed lines

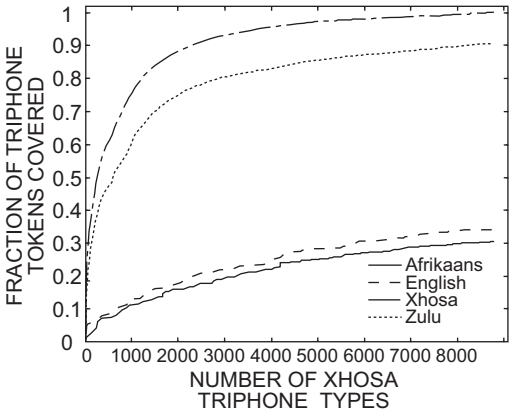

Figure 9: Coverage of test set triphone tokens by the most frequent Xhosa triphone types in each language

Zulu at the triphone level leads us to expect that these two databases can be mutually beneficial in the development of acoustic models. Specifically, we expect that the performance of a Xhosa speech recognition system can be enhanced by the addition of Zulu data, and vice versa. Alternatively, there appears to be scope for the development of a multilingual Nguni recognition system by fusing Xhosa and Zulu phone-based models. On the other hand, the comparatively poor overlap between

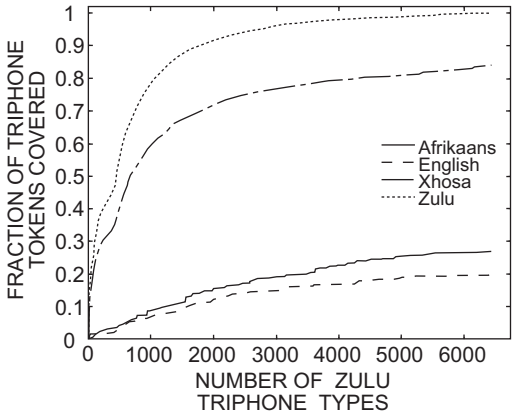

Figure 10: Coverage of test set triphone tokens by the most frequent Zulu triphone types in each language

Afrikaans and English, as well as any other combination of the four languages, is expected to deliver much smaller performance gains, or even none at all. In fact, other researchers in the field of multilingual speech recognition have reported precisely such small or non-existent gains, albeit for different sets of languages than those which are studied here (Köhler, 2001; Schultz & Waibel, 2001).

Borrowed phones and words

South African speakers of one of the four languages studied in this paper will usually also speak at least one of the remaining three. This leads to lexical as well as phonetic sharing between the languages, which we will try to quantify in this section.

We will refer to a word imported from another language during code-switching as a borrowed word. Borrowed words are distinct from loanwords, which are foreign words that have been phonologised and thus integrated into the phonetic structure of a language. For example, the Xhosa word *ifestile* has been derived from the Afrikaans word *venster*, meaning 'window'. This would be referred to as a loanword since the phonology of the original Afrikaans word has been altered to conform to the morphological structure of Xhosa. On the other hand, in the Zulu sentence *Lezi zinto zibuza u-twenty five rand*, meaning literally 'These items cost twenty-five rand', the words 'twenty-five rand' are borrowed words,

because they have not been phonologised. According to our definition, code-switching occurs when borrowed words are uttered but not when loanwords are spoken.

During the orthographic transcription of our speech databases, words considered foreign to the language in question were marked accordingly. For example, the Afrikaans exclamation *ag!* ('oh!') was marked as foreign when it was uttered in the English database. The same applied to English numbers spoken in the Xhosa and Zulu databases. Hence, it was possible to determine a list of the borrowed words present in each of the four databases. Table 3 presents these as percentages of both word types and word tokens in the test set.

Table 3 shows that all languages incorporate a significant proportion of borrowed words, and that this proportion is much higher for Xhosa and Zulu than it is for Afrikaans and English. Closer analysis reveals that borrowed words in English and Afrikaans consist mostly of proper nouns. For Xhosa and Zulu, on the other hand, borrowed words consist most often of English numbers, letters and the names of months.

Code-switching occurs often in both modern Xhosa and Zulu, and leads to a high proportion of borrowed words. In particular, it is accepted practice in several African languages to cite numbers, dates and amounts in either the mother tongue or in English. In some cases, these items are even cited in Afrikaans. This occurs because the English alternative is often much shorter than the equivalent Xhosa or Zulu. In Xhosa, for instance, the number 2 353 is often read simply as 'Two thousand three hundred and fifty-three'. However, it could also be read as *Amawaku amabini namakhulu amathathu namashumi amahlanu nantathu*, meaning literally 'Thousands-that-are-two and hundreds-that-are-three and tens-that-are-five and three'. Code-switching is also likely to appear in the spontaneous citing of dates and times. For example, a Zulu-speaking person might tell the time as *Isikhathi manje u-five past ten*, meaning literally 'The time now is five past ten' (Louw *et al*., 2001).

Once the set of borrowed words for each language had been determined, the set of phones used to pronounce words intrinsic to each of the four languages (i.e. not borrowed) could be determined. From these, we could determine which phones in the phone set of each language are used exclusively to pronounce borrowed words. Table 4 expresses these borrowed phones as percentages of both the number of phone types as well as the number of phone tokens in the training set. The values in Table 4 indicate that each language devotes an approximately equal proportion of its phone types to borrowed words. However, for Zulu and especially for Xhosa, these phones represent a much larger proportion of the training set than they do for Afrikaans and English. We conclude that code-switching has had a greater phonetic impact on the Xhosa database than it has had on the Zulu database. Furthermore, the impact of code-switching on Xhosa and Zulu by far outstrips its impact on Afrikaans and English. In fact, only 0.3% of the phones in the English data have been introduced by code-switching, and are not naturally a part of the English language.

Hence, we could also say that the variety of sounds regularly produced by Afrikaans mother tongue speakers has been expanded more by the presence of the other languages than it has for English mother tongue speakers. To an increasingly greater extent this is true for Zulu and Xhosa mother tongue speakers, respectively.

Language models

In order to obtain some insight into the diversity of the four phone sets, unigram language models were obtained from the

Table 3: Prevalence of borrowed words in each language

Training set	Borrowed words as % of:		Examples of borrowed words
	word types	word tokens	
Afrikaans	5.8%	0.3%	phone, sorry, Brixton
English	2.8%	0.3%	kloof, ja, Dalsig
Xhosa	12.2%	52.0%	ten, o'clock, Durban
Zulu	10.9%	45.3%	eight, and, university

Table 4: Phones borrowed from other languages

Training set	Borrowed phones as % of: phone types	phone tokens
Afrikaans	29%	1.6%
English	31%	0.3%
Xhosa	32%	10.3%
Zulu	34%	6.4%

Table 5: Phone unigram language models

Training set	Number of unigrams	Unigram perplexity
Afrikaans	82	29.2
English	73	34.3
Xhosa	104	35.6
Zulu	91	32.7

phone transcriptions of each training set. A unigram is a statistical language model that assigns a probability to the occurrence of every phone in a test sequence based only on the overall occurrence statistics of all phones in the training set. Perplexities were calculated on the test sets, and are shown in Table 5. Perplexity is a measure of the predictability of a phone sequence (Jelinek *et al*., 1983). A higher perplexity indicates that, on average, the next phone in a sequence is harder to predict.

Afrikaans has a lower unigram phone perplexity than English, even though it has a larger number of phones. The implication is that, relative to English, Afrikaans has a large proportion of phones which are used infrequently. This is confirmed by Figure 1, which shows that 99% of the Afrikaans test set can be covered by retaining only the 34 most frequent monophones.

A backoff bigram language model (Katz, 1987) was obtained from the phone transcriptions of each language's training set. Like the unigram, a bigram language model assigns a probability to each phone in a test sequence. However, in addition to the overall occurrence statistics, this model takes the identity of the predecessor phone in the test sequence into account. Table 6 lists the test set perplexity of each bigram language model. Absolute discounting was used for the estimation of language model probabilities (Ney *et al*., 1994).

It is interesting to note that Xhosa shows a slightly lower bigram perplexity than English, even though it has a higher unigram perplexity and a significantly larger phone set. For Zulu, the bigram perplexity is lower still. This indicates that there is a stronger sequential relationship between consecutive phones for Xhosa and Zulu than there is for English. We believe this to be the case because both Xhosa and Zulu conform to the phonological /V/, /CwV/ or /CV/ syllable structure. This means that a consonant is always followed by a vowel or the glide [w]. Furthermore, since nouns typically begin with a vowel — e.g. *abafana* ('boys'), *abantu* ('people') and *isikolo* ('school') — and a concordial system comprising /V/ or /CV/ elements is in place, co-articulation across word boundaries will largely be confined to the five basic vowels.

Table 6: Phone bigram language models

Training set	Number of bigrams	Bigram perplexity
Afrikaans	1 503	12.4
English	2 012	14.5
Xhosa	2 466	14.3
Zulu	1 651	13.6

Summary and conclusions

This paper has presented a comparative corpus-based analysis at phonetic level for Afrikaans, English, Xhosa and Zulu. The study was based on four annotated databases of recorded speech gathered under similar conditions. An analysis of the phone sets and phone transcriptions of each language shows both Xhosa and Zulu to be substantially more phonetically complex and diverse than the remaining two languages. From this point of view, speech recognition may be expected to be intrinsically more difficult for these languages. In contrast, we find Afrikaans to have the most compact phone set and predictable sequential phonetic structure of the four languages. We also find that the phonetic content of Xhosa and Zulu overlaps to a very large degree, in terms of both monophones and triphones. The same cannot be said for any other combination of the four languages. In particular, there is no strong overlap between Afrikaans and English, despite their common European roots. In fact, it is found that Afrikaans is often more phonetically similar

to Xhosa and to Zulu than it is to English. From these findings, we conclude that there appears to be scope for the sharing of Xhosa and Zulu speech data in the development of phone models for multilingual speech recognition. We do not, however, anticipate that significant gains can be made by pooling data from any of the other languages studied.

Notes

[1] Code-switching occurs when a speaker changes from one language to another during discourse. A related phenomenon is code-mixing, which refers to the use of words from one language in an utterance predominantly in another language. For the purposes of our analysis, we will not distinguish between code-switching and code-mixing.

Acknowledgements — The AST project was supported by the Innovation Fund (Project No. 21213) of the Department of Arts, Culture, Science and Technology (DACST) of the South African national government. We thank the reviewers for their constructive comments.

References

Baily R. 1995. The Bantu languages of South Africa. In: Mesthrie R (ed) *Language and Social History*. Cape Town: David Philip Publishers.

International Phonetic Association. 2001. *Handbook of the International Phonetic Association.* Cambridge: Cambridge University Press.

Jelinek F, Mercer RL & Bahl LR. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **5**:179–190.

Jessen M & Roux JC. 2002. Voice quality differences associated with stops and clicks in Xhosa. *Journal of Phonetics* **30**:1–52.

Katz S. 1987. Estimation of probabilities from sparse data for the language model component of a speech recogniser. *IEEE Transactions on Acoustics Speech and Signal Processing* **35**(3): 400–401.

Köhler J. 2001. Multi-lingual phone models for vocabulary independent speech recognition tasks. *Speech Communication* **35**: 21–30.

Ladefoged P. 1975. *A Course in Phonetics.* New York: Harcourt Brace Jovanovich Inc.

Louw PH, Roux JC & Botha EC. *2001. African Speech Technology (AST) telephone speech databases: corpus design and contents*. Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH), Part 3, pp 2055–2058.

Ney H, Essen U & Kneser R. 1994. On structuring probabilistic dependencies in stochastic language modelling. *Computer, Speech and Language* **8**: 1–38.

Poulos G & Msimang CT. 1998. *A Linguistic Analysis of Zulu*. Cape Town: Via Africa.

Roberge T. 1995. The formation of Afrikaans. In: Mesthrie R (ed) *Language and Social History*. Cape Town: John Philip Publishers.

Roux JC, Louw PH & Niesler TR. 2004. *The African Speech Technology Project: an assessment*. Proceedings of the 4th International Conference on Language Resources and Evaluation, Part 1, pp 93–96.

Schultz T & Waibel A. 2001. Language independent and language adaptive acoustic modelling for speech recognition. *Speech Communication* **35**: 31–51.

Statistics South Africa. 2004. *Census 2001 — primary tables South Africa: Census 1996 and 2001 compared.*

Wentzel PJ, Botha JJ & Mzileni PM. 1972. *Xhosa Taalboek*. Johannesburg: Perskor.

Appendix: The following table presents a list of the phones used to transcribe the Afrikaans, English, Xhosa and Zulu databases. An example of a word in which the phone occurs in each language is given. When this example word appears in italics, this indicates that it has been borrowed from one of the other languages during code-switching. Noise and silence phones, as detailed in the section entitled 'Non-speech sounds', are not included in the table

DESCRIPTION	IPA	Afrikaans	English	Xhosa	Zulu
Stops					
Voiceless Bilabial Plosive	p	pop	spit	*Cape*	*Cape*
Aspirated Bilabial Plosive	p^h			phila	phila
Ejective Bilabial Plosive	p'			pasa	impi
Voiced Bilabial Plosive	b	baba	baby	imbuzi	imbuzi
Devoiced Bilabial Plosive	b_0			bhala	bholoho
Voiced Bilabial Implosive	ɓ			ubawo	ubaba
Voiceless Alveolar Plosive	t	tot	total	*Atlanta*	*Atlanta*
Aspirated Alveolar Plosive	t^h			thatha	isikhathi
Ejective Alveolar Plosive	t'			itakane	ikati
Voiced Alveolar Plosive	d	daar	death	indoda	isidala
Devoiced Alveolar Plosive	d_0			amadoda	
Ejective Palatal Plosive	c'			ukutya	
Aspirated Palatal Plosive	c^h			ityhefu	
Voiced Palatal Plosive	ɟ			indyevo	
Voiceless Velar Plosive	k	koek	kick	*Brakpan*	*Brakpan*
Aspirated Velar Plosive	k^h			yikha	-khipha
Ejective Velar Plosive	k'			kakubi	kanti
Voiced Velar Plosive	g	berge	gun	ingubo	ugogo
Glottal Stop	ʔ	ver_as	co_operative	i_oyile	u_opharetha
Fricatives					
Voiceless Labiodental Fricative	f	vier	four	ukufa	ukufa
Voiced Labiodental Fricative	v	water	vat	ukuvula	ukuvula
Voiceless Dental Fricative	θ	*Martha*	thing	*thousand*	*think*
Voiced Dental Fricative	ð	*Northam*	this	*the*	*the*
Voiceless Alveolar Fricative	s	sies	some	sala	ukusala

DESCRIPTION	IPA	Afrikaans	English	Xhosa	Zulu
Voiced Alveolar Fricative	z	zoem	zero	ukuzama	ukuzama
Voiceless Post-Alveolar Fricative	ʃ	sjoe	shine	kushushu	shisa
Voiced Post-Alveolar Fricative	ʒ	genre	genre	*genre*	*genre*
Voiceless Velar Fricative	x	gaan	*Gauteng*	irhafu	*Gauteng*
Voiceless Glottal Fricative	h	*Bethlehem*	hand	huhuza	huhuluza
Voiced Glottal Fricative	ɦ	hand	*Johannes*	ihashe	ihhashi
Voiceless Alveolar Lateral Fricative	ɬ	*Umhlanga*		hlala	hlala
Voiced Alveolar Lateral Fricative	ɮ			dlala	dlala
Affricates					
Ejective Alveolar Lateral Affricate	t͡ɬ’			intloko	inhloko
Post-Alveolar Affricate	t͡ʃ	tjokker	chocolate	tshixa	intsha
Aspirated Post-Alveolar Affricate	t͡ʃʰ			ukutshisa	
Ejective Alveolar Affricate	t͡s’			ukutsiba	tsotsi
Aspirated Alveolar Affricate	t͡sʰ			isithsaba	
Voiced Post-Alveolar Affricate	d͡ʒ	*John*	jug	inja	juba
Voiced Alveolar Affricate	d͡z			amanzi	amanzi
Voiced Lateral Alveolar Affricate	d͡ɮ			indlovu	indlala
Ejective Velar Affricate	k͡x’			ikrele	-kleleba
Clicks					
Dental Click	ǀ			cinga	cima
Nasalised Dental Click	ǀ̃			nceda	ncane
Voiced Nasalised Dental Click	ǀ̬̃			iingcango	ingcazi
Voiced Dental Click	ǀ̬			gcina	gcina
Aspirated Dental Click	ǀʰ			chaza	chitha
Alveolar Lateral Click	ǁ			xela	uxolo
Nasalised Alveolar Lateral Click	ǁ̃			nxila	nxusa
Voiced Nasalised Alveolar Lateral Click	ǁ̬̃			ingxelo	ingxenye
Voiced Alveolar Lateral Click	ǁ̬			gxeka	gxuma
Aspirated Alveolar Lateral Click	ǁʰ			xhasa	xhuma
Palatal Click	ǃ			qiqa	qoma

DESCRIPTION	IPA	Afrikaans	English	Xhosa	Zulu
Nasalised Palatal Click	ǃ̃			nqaba	inqola
Voiced Nasalised Palatal Click	ǃ̬̃			ngqukuva	ngqatha
Voiced Palatal Click	ǃ̬			gquba	gqabula
Aspirated Palatal Click	ǃʰ			qhuba	qhaqha
Trills and Flaps					
Alveolar Trill	r	roer	*Hartenbos*	isitrato	egaraji
Uvular Trill	ʀ	*roer*		ndigraya	
Alveolar Flap	ɾ	vat dit	forty	*quarter*	*quarter*
Approximants					
Alveolar Approximant	ɹ	*Brixton*	red		*red*
Alveolar Lateral Approximant	l	lag	legs	lala	lala
Palatal Approximant	j	jas	yes	yima	yima
Voiced labio-velar Approximant	w	*William*	west	wela	wathuza
Nasals					
Bilabial Nasal	m	ma	man	mama	umama
Syllabic Bilabial Nasal	m̩			umfana	ngiyambona
Alveolar Nasal	n	non	not	iyana	-na
Palatal Nasal	ɲ	mandjie		inyama	inyama
Velar Nasal	ŋ	lang	thing	ingubo	ingane
Vowels					
High Front Vowel	i	siek	*Piet*	impilo	impilo
High Front Vowel with duration	iː	ski	keep	impilo	impilo
Lax Front Vowel	ɪ	*Brixton*	him	*Cecil*	*Cecil*
Rounded High Front Vowel	y	u		*u*	*u*
Rounded High Front Vowel with duration	yː	ekskuus			
High Back Vowel	u	boek	*Kapkaroord*	vulani	vulani
High Back Vowel with duration	uː	boer	blue	vula	vula
Lax Back Vowel	ʊ	*Woodstock*	push	*Newtown*	*Newtown*
Mid-high Front Vowel	e			ndithengile	ngithengile
Mid-high Front Vowel with duration	eː	been	*Vrede*		

DESCRIPTION	IPA	Afrikaans	English	Xhosa	Zulu
Rounded Mid-high Front Vowel	ø	neus			
Rounded Mid-high Front Vowel with duration	øː	keuse			
Rounded Mid-high Back Vowel	o	*Sibongile*	*Sibongile*	ukubonisa	ukubonisa
Rounded Mid-high Back Vowel with duration	oː	hoop			
Mid-low Front Vowel	ɛ	mes	nest	Themba	thenga
Mid-low Front Vowel with duration	ɛː	lê	fairy	aneesenti	nembeza
Rounded Mid-low Front Vowel	œ	brug	nurse		
Rounded Mid-low Front Vowel with duration	œː	brue	burst		
Central Vowel with duration	ɜː	wîe	turn	*third*	*third*
Rounded Mid-low Back Vowel	ɔ	mos	*Hartenbos*	molo	ubhalo
Rounded Mid-low Back Vowel with duration	ɔː	môre	bore	anethoba	ncokola
Low Back Vowel	ɒ	*McDonald's*	hot	*box*	*box*
Lax Mid-low Vowel	ʌ	*public*	hut	*hut*	*hut*
Low Central Vowel	a	agter	*Garsfontein*	sala	ukusala
Low Central Vowel with duration	aː	saak	*Klerksdorp*	apha	ukusala
Low Back Vowel with duration	ɑː	*master's*	harp		
Central Vowel (Schwa)	ə	niks	the	*degree*	*degree*
Mid-low Front Vowel	æ	eg, help	average	*camp*	*camp*
Mid-low Front Vowel with duration	æː	ver	dad	*camp*	*camp*
Diphthongs					
	ɔi	rotjie			*Joyce*
	əi	allerlei	*Kleinmond*	*Spelled "H"*	*Spelled "H"*
	iu	geoloog		*Beaufort*	*Beaufort*
	ia	inisiatief		*financial*	*financial*
	iɔ	Centurion	*Georgalli*		
	iœ	Pretorius			
	ɪə		here	*years*	
	iə	vegetariër		*nearly*	
	œy	luiaard		*Nelspruit*	
	ui	moeiliker			

DESCRIPTION	IPA	Afrikaans	English	Xhosa	Zulu
	əu	nou		road	road
	əʊ	*phone*	hope	road	
	ai	poegaai	*Laaiplek*	*drive*	*drive*
	aːi	baaie		*Laaiplek*	
	aɪ		*Laaiplek*	*drive*	
	ʌɪ	*Alpine*	fine		
	aʊ	*Camperdown*	power	*south*	
	au	*Landsdown*	power	*south*	*south*
	oi	weggooi		*point*	
	ɔɪ		boy	*point*	
	ɔu			road	road
	eu	eeu			
	ɛi			*Cape*	*Cape*
	ɛɪ		waste	*Cape*	
	eɪ	*gymnasium*	play	*Cape*	*Cape*
	ɛə		there	*there*	
	ʊə		poor	*poor*	
	ua	*Francois*			*February*
	iɛ	twee en			
	ʌə		Bryan		
Tripthongs					
	ɪɔɛ		Spelled "TON"		
	əaʊ		thousand		
	əiə	hy is			