A Southern African corpus for multilingual name pronunciation

Oluwapelumi Giwa, Marelie H. Davel and Etienne Barnard
Multilingual Speech Technologies, North-West University, Vanderbijlpark, South Africa
Email: {oluwapelumi.giwa,marelie.davel,etienne.barnard}@gmail.com

***Abstract*—We describe the challenges that arise in predicting the pronunciations of proper names in a multilingual society. In order to improve our understanding of this issue – which is of significant practical importance for applications of speech technology – we have designed and collected a multilingual corpus of proper names. Both the names and the speakers are drawn from four South African languages, namely isiZulu, Sesotho, English and Afrikaans. We describe how the corpus was designed in order to probe the interaction between the speaker's language and the origin of the name, and discuss the practical steps that were taken in collecting the spoken utterances. A statistical investigation of the prompt material reveals some of the systematic differences between the languages.**

I. Introduction

In multilingual environments, such as those that are common in South Africa, names originate from a variety of languages and are then produced by speakers from language backgrounds that may or may not overlap with the original languages. Consequently, these names are often pronounced quite differently depending on the speaker, creating a significant challenge for automated speech recognition (ASR) systems. How should an ASR system be able to predict the way in which a personal name will be pronounced, if there is so much variation in the way in which people pronounce these names? This causes real difficulty in practical applications such as directory assistance or voice search systems.

Initial work with regard to four South African languages [1] has indicated that there is systematicity in the way speakers from the same language background produce proper names across languages. This is important since such systematic effects can potentially be used to improve the accuracy of proper name predictions. However, no study to date has collected a South African corpus that is of sufficient size to allow for the detailed analysis of this phenomenon.

In this paper we present the design and development of a new corpus containing names produced by speakers of four languages that represent a substantial part of the South African linguistic landscape. Names are selected from the same four languages, and all possible combinations of name language/speaker language are collected. We provide an analysis of the data collected and discuss ways in which this corpus can be utilised further in order to better understand multilingual and cross-lingual name pronunciation.

II. Background

The ability to produce accurate predictions of the pronunciation of a given word is an important component of typical speech processing systems [2]. Dealing with proper names – even within one language – is a particularly challenging task as proper names tend to be diverse and may follow idiosyncratic pronunciation (or spelling) conventions [3].

Internationally, several long-term efforts aimed at the development of pronunciation models for proper names have been undertaken. A prominent example was ONOMASTICA [4], a multi-year, international project which developed pronunciation models for a large number of European names. But even in this well-studied context, proper name pronunciation is not seen as a solved problem [5]. Some results that are particularly relevant for the design and analysis of a multilingual names corpus include the following:

- Consideration of the linguistic origin of a name is crucial during pronunciation prediction [6], [7].
- Phoneme-to-phoneme (p2p) conversion in addition to grapheme-to-phoneme (g2p) conversion is particularly important in the context of proper names [5], [8]. This requires access to information on the speaker's default language, the linguistic origin of the target name, and the speaker's assumption about that origin.

Current pronunciation-prediction techniques typically still rely on a combination of manual and automatic processing [3]. Automated methods, such as rule-based methods, can achieve a reasonable level of accuracy in predicting how proper names will be pronounced, but the range of exceptions is so wide that a manually-generated list of common proper names is generally employed as a first processing step.

For the languages of South Africa, this topic has received relatively little attention. Generic g2p rules exist for all the official languages of South Africa [9], but these are not able to produce accurate pronunciations for any words that do not follow the standard pronunciation rules of the language, and are therefore also poor predictors of personal names. While the prediction of loan words using letter-to-sound rules in the speaker's primary language has been shown to provide usable results for loan words which included proper names [10], it is expected that better results can be obtained using a more sophisticated modelling approach. This was demonstrated by Kgampe and Davel [1], who showed that the linguistic origin of the name and mother tongue of the speaker have a system-

atic effect across the four languages that they studied (English, Afrikaans, isiZulu and Setswana). However, their study only included a small set of respondents (20 speakers of 40 names): a larger corpus would be required for the development of a useable pronunciation system.

III. Corpus Design

The overall corpus design is based on the criteria and transcription protocols that were developed for the Autonomata Spoken Name Corpus [4], while also accounting for some of (a) the particular characteristics of the South African language landscape and (b) the salient factors that have emerged from research on names in multilingual environments, as summarised in Section II. Our most important design choices, and the motivations behind them, are summarised below.

A. Target languages

South Africa has eleven official languages, which fall into two language families (Bantu and Germanic). Seven of the nine Bantu languages in this set also cluster into two subfamilies (Nguni and Sotho); languages within these subfamilies are to a greater or lesser extent similar [11]. Since both the speaker language and the language of origin of a spoken name influence the expected pronunciation, complete coverage of all pairs would require an excessive number of sub-corpora. We have therefore decided to limit our attention to a subset of four languages which are the most common languages in the Gauteng province where the collection is to be done, namely isiZulu, Sesotho, English and Afrikaans. It so happens that these four languages also represent a reasonable fraction of the variability that occurs across the major South African languages: isiZulu and Sesotho are, respectively, languages from the Nguni and Sotho subfamilies, and English and Afrikaans are by some measures the two languages which are individually the least similar to any other official language in South Africa [11].

We have decided to weigh English more heavily than the other languages in our corpus, to compensate for the well-known dissonance between English orthography and pronunciation. Hence, more samples of English words are likely to be necessary if rules of comparable accuracy are to be derived, and our corpus contains twice as many English names (40%) as in each of the other three languages (20% each). Our speakers, however, are drawn in equal measure from all four language groups (more details below).

B. Selection of names and speakers

In ONOMASTICA, special effort was made to achieve a predetermined balance between frequent and rare words. However, for the languages that we have selected, sufficiently diverse corpora are not available to aim for such a balance. Also, given the very limited knowledge that is available on the cross-lingual pronunciation of South African names, we decided that somewhat more prototypical cases would be most useful for the current corpus. We therefore opted to select names that are in current use: a name list of recent and present students at a large residential university was used as starting point, and first-language speakers of the four target languages were asked to select names that are typically associated with their language. In South Africa, full names are most commonly spoken in the format *Personal name – Family name*. In fact, this format is so common that personal names are generally known as "First names" and family names as "Last names". Thus, separate lists were created for first names and last names in each of the four target languages. From these lists of selected names, a total of 600 first names and 600 last names were drawn in specific language ratios, as described in more detail below.

The rapid language changes that are occurring in South Africa imply a substantial diversity of speakers, even when they nominally share the same first language. It is therefore not feasible to aim for a representative sample of any significant subset of South African society within our small corpus. As a consequence, we have decided to employ students on a university campus for all our recordings. (Fortunately, such students are a highly relevant demographic for the types of practical services that may require multilingual name recognition.)

C. Combining personal and family names

We also had to choose how pairs of first and last names would be drawn from the languages in our trial, since it is not generally true that the personal and family name of a particular person will originate from the same language. As a very simple model of cross-language name combinations, and to match our bias towards English names, we have chosen our name combinations according to the following frequencies:

- 25% of the names are English-English (EE);
- ZZ, SS and AA name pairs contribute 15% each (45% in total); and
- the combinations EZ, ES, EA, ZE, SE and AE each constitute 5% of the corpus.

Typical names selected in each of these categories are shown in Table I.

TABLE I
*The names in our prompt list are combinations of **E**nglish, isi**Z**ulu, **S**esotho and **A**frikaans first and last names.*

Languages	Example
EE	Catherine Wallace
ZZ	Nolwazi Zilwana
SS	Mosebo Moremedi
AA	Reinhard Viljoen
EZ	Gareth Mbuyisa
SE	Khothatso Wingard

D. Creating prompt lists

A characteristic of the Autonomata corpus that has been very useful in practice (J-P Martens, personal communication) is the way in which it separates both speakers and content. That is, the prompted content was separated into a number of different lists, and a distinct subgroup of speakers pronounced each list, thus making it possible to analyse both speaker

differences and content-related differences. To obtain this same benefit our names are split into three separate lists, each containing 200 full names (first name and last name). Each list is recorded by 16 different speakers (4 first-language speakers of each language, of whom 2 are female and 2 male), yielding the corpus speaker design shown in Table II. Speakers do not repeat across lists, and 48 speakers are collected in total.

TABLE II
The corpus consists of three separate lists of 200 full names each; each list is recorded by 4 first-language speakers of each language, of whom 2 are female and 2 male.

Language	List A		List B		List C		Total
	M	F	M	F	M	F	
isiZulu	2	2	2	2	2	2	12
Sesotho	2	2	2	2	2	2	12
English	2	2	2	2	2	2	12
Afrikaans	2	2	2	2	2	2	12
Total	8	8	8	8	8	8	48

E. Practicalities

For the recording process, we used mobile telephones running the data-collection application *Woefzela* [12]. The application prompts speakers by displaying the utterance to be spoken (in the format *Firstname Lastname*) on the screen of the mobile telephone, along with controls for starting and stopping the recording of each prompt, as well as repeating unsuccessful prompts. It also monitors the signal-to-noise ratio and duration of each prompt, and requests the speaker to repeat prompts which were apparently not recorded successfully.

Each session starts with the signing of a consent form by the speaker, followed by a training session in which speakers are taught to use the controls. During the main recording session, *Woefzela* keeps track of the recording progress, displaying the number of completed and remaining prompts to the speaker.

Recordings are done in a quiet office environment, but no special precautions are taken to ensure acoustic quality, since our goal with the corpus is to study the broad phonetic details of name pronunciation, rather than fine acoustic details. All recordings are stored in the default format provided by *Woefzela*, namely 16-bit single-channel files, sampled at 16 kHz and saved in Microsoft Wav format.

We also collect relevant meta-information from each speaker, including the following:

- Number of years lived in the current dialect region.
- Primary language(s) at home while growing up.
- Current primary language(s) at home, if different.
- Additional languages that the respondent either understands or speaks.
- Education level (highest qualification).
- Age (in years).
- Gender.

Apart from the pronunciation of each name, we do not capture any additional name-specific information. This is different from the corpus of [1] where the speaker's familiarity and knowledge of the language of origin of each name were also captured. It is interesting to consider whether a speaker will pronounce a name differently if previously heard (and repeating from memory) and if the same name was read from a list. Both these scenarios occur in practice (for example, in a directory assistance system), and will require further analysis for which the current corpus will not provide an exact answer.

The overall corpus design is summarised in Table III.

IV. Analysis

In order to provide an intuition for some of the language differences that occur in our corpus, Tables IV, V and VI list the ten most common letter unigrams, bigrams and trigrams of the prompted names in each of the four languages. It is obvious that the two Germanic languages are somewhat similar, and that the two languages from the Bantu family also share commonalities (though apparently less so than English and Afrikaans).

TABLE IV
The ten most frequent letter unigrams in our corpus, for each language.

isiZulu		Sesotho		English		Afrikaans	
a	0.120	o	0.131	e	0.120	e	0.144
i	0.106	a	0.127	a	0.096	a	0.109
n	0.097	e	0.120	n	0.091	n	0.092
e	0.084	m	0.079	r	0.089	r	0.088
l	0.081	t	0.073	l	0.067	i	0.076
o	0.058	l	0.068	i	0.061	l	0.061
m	0.056	s	0.065	o	0.060	t	0.052
u	0.053	h	0.057	s	0.054	s	0.047
s	0.053	i	0.052	t	0.041	o	0.042
h	0.050	k	0.045	h	0.037	h	0.034

TABLE V
The ten most frequent letter bigrams in our corpus, for each language.

isiZulu		Sesotho		English		Afrikaans	
an	0.037	mo	0.051	er	0.029	an	0.043
le	0.031	ts	0.039	on	0.024	er	0.034
si	0.030	le	0.032	an	0.023	ie	0.032
la	0.027	ma	0.029	ar	0.020	ri	0.028
ng	0.022	an	0.029	en	0.019	en	0.028
el	0.022	se	0.027	ne	0.016	ar	0.023
ma	0.019	lo	0.022	in	0.016	ma	0.022
nd	0.019	th	0.020	ll	0.015	el	0.021
il	0.018	di	0.020	ri	0.015	li	0.018
zi	0.017	ok	0.019	le	0.015	te	0.018

TABLE VI
The ten most frequent letter trigrams in our corpus, for each language.

isiZulu		Sesotho		English		Afrikaans	
ile	0.016	ane	0.017	son	0.011	mar	0.015
ele	0.011	ets	0.012	ell	0.007	rie	0.010
and	0.011	tsh	0.011	ers	0.007	ari	0.010
ane	0.010	ots	0.011	ine	0.006	tte	0.009
ela	0.009	mok	0.011	har	0.005	ter	0.008
nga	0.009	ele	0.010	ton	0.005	ett	0.008
lan	0.009	tha	0.010	enn	0.005	lie	0.007
ani	0.009	mot	0.010	nne	0.005	ize	0.007
osi	0.007	tsi	0.010	lin	0.005	lan	0.007
ong	0.007	tse	0.009	ill	0.004	van	0.007

TABLE III
Overall corpus design; to be read in conjunction with Table II

Names	Categories	Personal names: first names and surnames
	Languages	isiZulu, Afrikaans, Sesotho, English
	Distribution	Frequent names, contextually diverse
Respondents	Primary language(s)	isiZulu, Afrikaans, Sesotho and/or English
	Dialect	Gauteng
	Gender	Balanced between male and female
	Age	Adult speakers (>18)
Recordings	Type	Native and non-native pronunciations of personal names
	Style	Read prompts
	Quality	Low noise, natural environment
	Encoding	16 kHz, Microsoft wav files, 16 bit, mono
Prompts	List constitution	200 full names per list (200 first names and 200 last names)
	Language origin	25% EE, 15% AA, 15% SS, 15% ZZ 5% AE, 5% SE, 5% ZS, 5% EA, 5% ES, 5%EZ
Meta-data	Per name	Orthography, name type, name language.
	Per respondent	Primary languages, additional languages, gender, age, dialect grouping, educational level grouping.
	Per recorded name	Name, respondent, pronunciation language, phonemic transcription.
	Phonemic transcription	Phoneme string. Addition of stress, syllable information. Combination of manual and automated transcription and verification.

To investigate this intuition, we also computed cross entropies between the n-grams that occur in the various languages. Table VII shows those values calculated on the trigrams for all pairs of languages. Here it is clear that the languages have significantly different trigram entropies, with Sesotho having the lowest entropy and English the highest; the ordering of entropies agrees with our intuition of the amount of regularity in names in each of the languages. The cross entropies for different languages confirm our observations derived from the frequent n-grams above.

TABLE VII
Cross entropies for all language pairs, as computed from letter trigram statistics

	isiZulu	Sesotho	English	Afrikaans
isiZulu	-6.012	-7.401	-7.754	-7.782
Sesotho	-7.378	-5.944	-7.816	-7.913
English	-7.889	-7.954	-6.660	-7.474
Afrikaans	-7.885	-7.985	-7.373	-6.261

These results are all based on the name orthographies – it would, of course, be interesting to also compute measures of similarity for the spoken forms. As an initial step in this direction, we have simply phonetized each of the words in our corpus, using the pronunciation rules that were derived for generic words in the appropriate language [9]. For the reasons discussed in Section I, these rules are not expected to be accurate for name pronunciations, but they do at least give an indication of how the languages compare with one another. As shown in Table VIII these (cross-)entropies are quite similar to those computed for the letter n-grams, except for an apparently increased shift in the difference between the two Germanic languages in Table VIII. However, that increase is somewhat misleading: the cross-entropy measure is based on a binary distinction (either two phonemes are the same or they are different), whereas phonemes can actually be more or less similar. Thus, the vowel shifts that are quite common in Germanic languages are exaggerated by this measure – for the Bantu languages, which use a smaller set of vowels, this phenomenon is less important.

TABLE VIII
Cross entropies for all language pairs, as computed from triphone statistics

	isiZulu	Sesotho	English	Afrikaans
isiZulu	-6.068	-7.545	-7.876	-7.777
Sesotho	-7.554	-6.170	-7.957	-7.894
English	-7.923	-7.965	-6.667	-7.736
Afrikaans	-7.824	-7.928	-7.776	-6.283

V. Conclusion and status

We have described the development of a multilingual corpus that is specifically aimed at the diverse proper names that occur in South Africa. It is nonetheless likely that the corpus will be useful in other environments where several languages are spoken within the same community, or within closely-interacting communities. In particular, the balance between own-language pronunciations and imitation of foreign-language pronunciations, as well as the different styles of imitation that occur, are of universal importance. By comparing regularities across widely different corpora (such as the one described here, the Autonomata corpus which was collected in Europe, and others), we should be able to gain a better understanding of the parameters that influence these various approaches to proper-name pronunciation.

We are currently completing the collection and verification of the corpus described above; phonemic transcriptions of all

words, based on the phone sets used in [9], will be our next task. As cross-lingual phonemic transcription is a particularly challenging task, we aim to use a combination of automated and manual means for this process. Once transcription has been completed, the corpus will be made available freely under an Open Content license. It is our hope that the release of the corpus will stimulate others to investigate this challenging and practically important subject.

ACKNOWLEDGMENT

This corpus is being developed in collaboration with Jean-Pierre Martens from the University of Ghent (Belgium) and Derik Thirion from Molo Innovations (South Africa). Corpus development is being sponsored by the Department of Arts and Culture of the government of the Republic of South Africa; their support is gratefully acknowledged.

REFERENCES

[1] M. Kgampe and M. H. Davel, "Consistency of cross-lingual pronunciation of South African personal names," in *Proceedings of the 21st Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 2010)*, Stellenbosch, December 2010.

[2] M. Gales and S. Young, "The application of Hidden Markov Models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 2, pp. 195–304, 2007.

[3] M. F. Spiegel, "Proper name pronunciations for speech technology applications," *International Journal of Speech Technology*, vol. 6, p. 419427, 2003.

[4] H. van den Heuvel, J. P. Martens, B. D'hoore, K. D'hanens, and N. Konings, "The Autonomata spoken name corpus. design, recording, transcription and distribution of the corpus," in *Proceedings LREC*, Marrakech, 2008.

[5] H. van den Heuvel, B. Réveil, and J. P. Martens, "Pronunciation-based ASR for names," in *Proceedings Interspeech*, Brighton, 2009, pp. 2991–2994.

[6] A. F. Llitjos and A. W. Black, "Knowledge of language origin improves pronunciation accuracy of proper names," in *Proceedings Interspeech*, Aalborg, Denmark, 2001, pp. 1919–1922.

[7] B. Reveil, J. Martens, and B. Dhoore, "How speaker tongue and name source language affect the automatic recognition of spoken names," in *Proceedings Interspeech*, Brighton, 2009, pp. 2995–2998.

[8] Q. Yang, J. P. Martens, N. Konings, and H. van den Heuvel, "Development of a phoneme-to-phoneme (p2p) converter to improve the grapheme-to-phoneme (g2p) conversion of names," in *Proceedings LREC*, Genua, 2006, pp. 287–292.

[9] M. Davel and O. M. Martirosian, "Pronunciation dictionary development in resource-scarce environments," in *Proceedings Interspeech*, Brighton, 2009, pp. 2851–2854.

[10] T. Modipa and M. H. Davel, "Pronunciation modelling of foreign words for Sepedi ASR," in *Proceedings of the 21st Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 2010)*, Stellenbosch, December 2010, pp. 185–189.

[11] P. N. Zulu, G. Botha, and E. Barnard, "Orthographic measures of language distances between the official south african languages," *Literator: Journal of literary criticism, comparative linguistics and literary studies*, vol. 29, no. 1, pp. 185–204, 2008.

[12] N. J. de Vries, J. Badenhorst, M. H. Davel, E. Barnard, and A. de Waal, "Woefzela - an open-source platform for ASR data collection in the developing world," in *Proceedings Interspeech*, Florence, Italy, August 2011, pp. 3176–3179.