Multilingual pronunciations of proper names in a Southern African corpus

Jan W.F. Thirion, Marelie H. Davel and Etienne Barnard
North-West University, Potchefstroom, South Africa
E-mail: {thirionjwf,marelie.davel,etienne.barnard}@gmail.com

***Abstract*—We present our process for the development and analysis of a multilingual names corpus, called Multipron-split. It is derived from Multipron, a corpus collected in previous work [1], where names and speakers were drawn from four South African languages, namely Afrikaans, English, isiZulu and Sesotho. The new corpus is more suited for multilingual pronunciation modelling and research as the "words" consist of either a name or surname, rather than a combination of the two. This enables us to model pronunciations from a single language of origin, which has previously been shown to be important in pronunciation modelling for proper names. An algorithm is presented through which the most common pronunciations of names, also called reference pronunciations, can be automatically extracted from the observed pronunciations. We show that the most common pronunciation variants correlate well with the different speaker languages, and that systematic phone substitutions occur when speakers of one language pronounce names from a different language. Also, reasonably accurate automatic pronunciations can be generated with an automatic grapheme-to-phoneme converter, especially when the speaker language agrees with the name language.**

I. Introduction

Various factors such as a speaker's region of origin, mother tongue, age and socio-economic background result in systematic pronunciation differences between speakers [2]. In a multilingual environment, such as in South Africa, this issue is particularly prominent, since most automated speech-processing systems will be required to operate on speech from speakers with a variety of linguistic backgrounds. In particular, it is generally accepted that poor pronunciation modelling can lead to deteriorated automatic speech-recognition (ASR) performance [3]; this is especially true for multilingual proper names as well as loan words, where native pronunciation rules are often inaccurate [4]. For resource scarce environments, such as in South Africa, dealing with this problem adequately remains a challenge [5], [6], especially since resource-scarce languages are currently less important economically to the providers of commercial speech-recognition systems. Speech recognition of proper names is particularly important in applications such as voice search, directory assistance and automated attendants [7], [8], [9].

It is impractical to create a dictionary by hand with all possible pronunciations of all names in all languages (both because of the time and cost involved, and because of the inevitable inaccuracies that will result from such a process). Hence, pronunciation rules are often employed to predict pronunciations [3]. There is a need for a corpus on which the pronunciation rules for South African languages can be trained, where the linguistic origin of the name is taken into account. Earlier work [1] resulted in a multilingual corpus, called Multipron, for four South African languages, but these combined name/surname pairs typically had mixed languages of origin for the names and surnames, making pronunciation modelling problematic.

In this paper we present our process of transforming the Multipron corpus into a "split" corpus, Multipron-split, of individual names and surnames, tagged by their associated language of origin. We then automatically extract the typical pronunciation as would be produced by a native speaker of each name from the pronunciations in the corpus (observations). The Default&Refine algorithm [10] is then used as G2P converter to predict these reference pronunciations. An interpretation of the results gives insight into the structure of the corpus and the variants contained therein.

II. Background

It is well known that knowledge of the mother tongue of the speaker, as well as the linguistic origin of the word, can be beneficial to producing better pronunciation variants [11]. The consistency of cross-lingual pronunciation of proper names was recently studied for four South African languages, namely Afrikaans, English, Setswana and isiZulu [4]. It was confirmed that knowledge of the linguistic origin of each word was an important factor in predicting how it would be pronounced.

The Autonomata Spoken Names Corpus (ASNC) [12] was recently used in state-of-the art work most related to our current investigation [13], [14], [15], [16]. The database contains 3540 unique names of Dutch, French, English, Turkish and Moroccan origin. The corpus contained only names of people (personal names and surnames), street names and city names in a single language (i.e no mixed language names).

In [17] a tandem G2P-P2P approach was used for the G2P conversion of proper names, where an initial transcription generated by the G2P converter is passed to a P2P converter, along with the orthography of the word. The P2P converter applies learned rules (in the form of decision trees or rule networks, automatically learned from the data) that generate alternative pronunciations. In [18] this method was shown to work well for the G2P conversion of proper names, although the linguistic origin of the word was not taken into account. In [13] it was found that ASR accuracy for proper names

increased when pronunciation variants were added to the lexicon. This was true for native speakers speaking foreign names, but not for foreign speakers. Here "native" refers to the target language of the system (e.g. Dutch) and foreign, or "non-native" include English, French and Moroccan.

A study on how mother tongue and the linguistic origin of the word affect ASR performance, was reported in [14]. Language-specific G2P converters were used, both monolingual as well as multilingual acoustic models, and language-specific P2P converters. It was found that native speakers used their own non-native G2P rules when pronouncing unfamiliar words from the non-native language and not knowledge from the G2P rules from their native language. Non-native speakers, however, tended to employ their own non-native G2P rules when pronouncing unfamiliar words from the native language, resulting in substantial error increases. When the speaker's mother tongue was used as basis for selecting variants (from that language) for the recognition of foreign names, performance decreased. Also, names with linguistic origins of languages different from that of the native/target language of the system, were found to be easier to recognise due to the names having less chance of being confused with the pronunciations of the native language. An experiment was also done to investigate whether ASR performance increases if the correct transcription is always added to the lexicon. It is encouraging that improved ASR accuracies were observed for all native/non-native combinations. Better pronunciation prediction algorithms may thus lead to even more improved ASR accuracy as the lexicon will contain even better coverage of the true transcriptions.

The work in [15], [16] can be considered as the current state-of-the-art in the multilingual recognition of proper names using knowledge of the speaker's mother tongue and the linguistic origin of the word. Here it was found that nativised transcriptions [19] are appropriate as target transcriptions for P2P learning. P2P transcriptions improved ASR accuracy of non-native words by a native speaker, but not significantly for native and non-native words by a non-native speaker. Automatically generated P2P transcriptions compete well with typical transcriptions from human experts. For non-native words, speakers will attempt to use the non-native G2P rules of that language; hence, knowledge of the speaker's mother tongue is important for accurate P2P converters. When a P2P converter was trained on foreign names, it outperformed a P2P converter trained exclusively on native words.

From the work above, many unanswered questions remain. For example, it is unclear what benefit task-specific (trained on the same type of data we are trying to predict, taking language of origin into account), rather than language-specific G2P converters would have. It is also important to see how well the results obtained generalise to the South African languages. However, in [1], the names and surnames form a word in which the constituent parts could be of different language origins, making pronunciation analysis difficult – hence the need to create a "split" corpus in order to address these questions.

III. Approach

A. The Multipron "split" corpus

In order to split the first name-surname pairs in Multipron, we started with grapheme-to-phoneme alignment of the dictionary. Dynamic programming was used, with the orthography as the reference string (with a special symbol "=" used to join first names and surnames) and a manual transcription as the observation. (These manual transcriptions were created as an approximate starting point for further development by a first-language Afrikaans speaker, after listening to a few samples of each name.) An automatically trained scoring matrix with no gap extension penalties, based on the Needleman-Wunsch algorithm, was used for alignment [20]. Log-likelihood probabilities were used in the scoring matrix. Next, the aligned sequences were split where "=" was aligned to a gap. This resulted in a separate name and surname. In a few cases, the alignment could not be done (e.g. due to incorrect transcriptions), and these were inspected manually. There were 3 such name-surname combinations, of which 3 individual words could not be used. Hence, from the 10130 entries in the dictionary we generated 20257 individual words.

Word boundary effects were subsequently manually checked and corrected. All double graphemes at word boundaries in the orthography (first name ends in the same grapheme as the first grapheme in the surname) were marked to be checked. All double phonemes in the transcriptions (at the first name/surname boundary) were also marked, but none of these required manual intervention. For all /r/ phonemes that were dropped during the splitting process from the first names, no changes were made. All /l/ phonemes split off from the first name resulted in the phoneme being added to the transcription of the first name (at the end). Finally, double-consonant effects were corrected, as well as nasals. Table I shows a few of these manually corrected examples.

Uncorrected	Corrected
amber__rennie	*{ m b @ r\ E n i*
{ m b @ r\ E n i	{ m b @ r\ E n i
donald__day	*d Q n @ l d @i*
d Q n @ l d @i	d Q n @ l d d @i
peaceful__lottering	*p i s f @ l Q t r\ @ N*
p i s f @ l Q t r\ @ N	p i s f @ l l Q t r\ @ N
hellen__nzwakele	*h E l @ n z v a k E l E*
h E l @ n z v a k E l E	h E l @ n n z v a k E l E
markus__stoop	*m a r k @ s t u@ p*
m a r k @ s t u@ p	m a r k @ s s t u@ p
jeanett__taylor	*d Z @ n E t @i l @ r*
d Z @ n E t @i l @ r	d Z @ n E t t @i l @ r

TABLE I
EXAMPLES OF MANUALLY CORRECTED "SPLIT" WORDS.

B. Reference extraction

For pronunciation variation analysis and evaluation, the typical pronunciation of a word by a native speaker is needed, called references here. These can either be obtained from experts, or be extracted automatically. In the work presented here, a semi-automatic process was employed.

In order to create reference pronunciations, the following was done:

1) **Extract references**: References were extracted by first counting the number of occurrences of every observation/transcription for every word (orthography) from a given language origin, per speaker language. The observation (per speaker language) with the most occurrences was taken as the starting reference. If a name was not pronounced by a certain speaker language, then speaker language was ignored and the observation with the maximum occurrence irrespective of speaker language taken as the starting reference for that word-speaker language. A scoring matrix was then trained [20] between the transcriptions/observations and starting references. The average dynamic programming (DP) score between all observations of a word, per speaker language, was then computed. Two methods of reference selection were compared:
   - **OPTMAX**: We take the reference to be the transcription with the highest average DP score per speaker language. If there are ties (unlikely) then the transcription with the highest number of occurrences is taken. If there are still ties, then the first transcription is taken.
   - **MAXOPT**: We take the reference to be the transcription with the highest number of occurrences per speaker language. Ties are resolved by taking the transcription with the highest average DP score (to all observations) as the reference.

   Names that were not pronounced in certain speaker languages were again treated in the appropriate language-independent fashion for the respective reference selection method (i.e. the observation with the maximum number of occurrences or maximum average DP score, irrespective of speaker language was taken as the reference).
   A total of 5176 unique "split" references were extracted in this way. The reference for a name-surname combination is then the reference per speaker language for each part (name or surname) of the entry independently.
2) **Manual correction**: The references where the speaker language and name language were the same ("in-language") were checked and corrected by a human expert. It is assumed here that speakers with the same home language as that of the word origin would know best how to pronounce it.
3) **Create references**: The "in-language" references were used as the reference for every word. If a reference was not available for a word in a specific language, then one was selected from the other languages. Table II shows the preferences given. No attempt was made in this work to investigate how similar languages are.

	Language			
Choice	A	E	Z	S
1	E	A	S	Z
2	Z	Z	E	E
3	S	S	A	A

TABLE II
SELECTION OF ALTERNATIVE "REFERENCES" FROM "IN-LANGUAGE" REFERENCES FROM OTHER LANGUAGES WHEN AN "IN-LANGUAGE" REFERENCE FOR A WORD DOES NOT EXIST.

   If a reference could still not be found, an automatic reference from step 1 (for the same speaker language and name language as the word in question) could be used as back-off; however, we did not encounter any such cases. Name and surname references were combined to give references for all entries in the original Multipron dictionary. There were 261 Afrikaans, 517 English, 254 isiZulu and 262 Sesotho "in-language" references, for a total of 1294.

C. Reference prediction

Default&Refine [10] is a rule-based algorithm that can be used to perform grapheme-to-phoneme (G2P) conversion. In our task here, it is used to extract rules from the pronunciation dictionary and then predict the pronunciation of the references from the orthography alone. We explore two cases:

- Generic G2P trained on a variety of texts is used to predict the pronunciations for the names. The G2P rules were language-dependent, based on the Lwazi corpus [21].
- A task-specific G2P is developed, with rules trained on the names corpus developed here, using 10-fold cross-validation.

For both these cases, we compare the prediction results against the references extracted from the MAXOPT, OPTMAX and manually verified references. We consider two cases for each of these "target" reference sets:

- References dependent on speaker language only (we temporarily ignore name language).
- References dependent on speaker language and name language, where the speaker language and name language are the same - the so-called "in-language" references.

D. Variant analysis

The variant analysis we present gives insight into the reasons behind the variation between observed pronunciations and reference pronunciations. We take the manually verified reference pronunciations and extract simple P2P rules (no context) using the observed pronunciations. The resultant rules have the form

$p_r \rightarrow p_o$

where p_r is the phoneme from the reference pronunciation and p_o is the phoneme from the observed pronunciation.

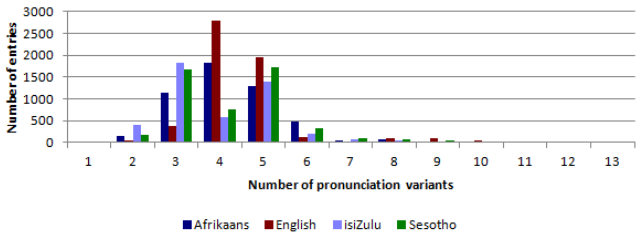

Fig. 1. Relationship between the number of entries (over all speaker and name languages) and the number of variants for an entry.

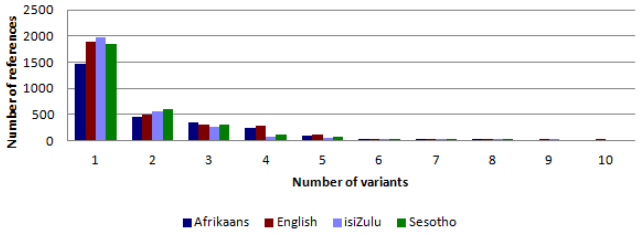

Fig. 2. Relationship between the number of references for each of the "in-language" references and the number of variants for such a reference.

IV. Experiments and Results

A. Corpus analysis

	Speaker language			
Name language	A	E	Z	S
A	1069	1144	912	976
E	2105	2289	1872	2020
Z	961	1063	907	985
S	971	1094	895	994

TABLE III
Number of entries in the Multipron-split corpus based on speaker language (mother tongue) and name language (language origin).

Table III shows the number of entries in the Multipron-split corpus split according to speaker and name language. A total of 20257 entries exist and these are fairly evenly distributed over the speaker and name language pairs, except for English words, which (by design of the Multipron corpus) were more frequent than those from other languages.

Figure 1 shows how many entries exist in the corpus with a given number of variants. The graph gives some insight into the variedness of the pronunciations – we see that most words in the corpus have around 3 to 5 pronunciation variants. This correlates well with Figure 2, from which we deduce that most variants consist of a single pronunciation in each of the speaker language/name language pairs.

Figure 2 shows the relationship between the number of references for each of the "in-language" references and the number of variants for such a reference. Here it can be seen that most of the "in-language" references (dependent on both a speaker and name language) had only a single pronunciation variant which we selected as reference . This is likely to be a typical scenario for multilingual corpora, due to the scarcity of data.

B. Reference extraction

To evaluate how well the reference extraction methods worked, we compared references extracted with those from a manually corrected version of the references. Dynamic programming alignment (Needleman-Wunsch) was performed, where a similarity score of 2 was given if the symbols were identical, -2 for a gap and -1 if they differed. Accuracy (Acc) was calculated as the average accuracy over all of the "in-language" references. The accuracy (percentage) for a single reference was calculated as:

$$Acc = 100 \cdot \frac{Num - Ins - Del - Sub}{Num}$$

(See the Appendix for additional information on the difference between accuracy and correctness, as well as other definitions of terminology.) We also counted the number of references that were perfectly predicted.

Lang	Acc	$Perf/T$	Ins	Del	Sub	Num
A	94.29	200/261	8	16	67	1564
E	94.84	411/517	23	23	93	2711
Z	94.29	178/254	5	22	72	1735
S	90.27	151/262	8	17	145	1816

TABLE IV
Accuracy (Acc), Perfectly predicted ($Perf$) and Total (T) references, Insertions (Ins), Deletions (Del), Substitutions (Sub), and Number (Num) of phonemes for the MAXOPT reference extraction method.

Lang	Acc	$Perf/T$	Ins	Del	Sub	Num
A	91.69	172/261	4	27	98	1560
E	92.18	362/517	19	37	157	2707
Z	92.75	160/254	2	30	89	1732
S	88.69	132/262	6	35	161	1814

TABLE V
Accuracy (Acc), Perfectly predicted ($Perf$) and Total (T) references, Insertions (Ins), Deletions (Del), Substitutions (Sub), and Number (Num) of phonemes for the OPTMAX reference extraction method.

Tables IV and V show the accuracy of the reference extraction methods. Here it can be seen that the MAXOPT method outperforms the OPTMAX method due to the most typical variants occurring more frequently than others. When data is particularly scarce, OPTMAX may still be useful to choose the most "average" pronunciation variant as reference. The percentage of references predicted with 100% accuracy using this method ranges between somewhat less than 60% for Sesotho to almost 80% for English. The relatively low accuracy for Sesotho results from inaccurate initial transcriptions. This was confirmed in that many of the errors encountered in the Multipron corpus, most of which were corrected by hand, were of Sesotho origin. The manually corrected Sesotho references were then quite different as a result.

The result of this analysis shows that these automatically extracted references may be very beneficial as a first-round version of pronunciations. A human expert may then check these transcriptions and do the manual corrections. Such semi-automated processes can save a considerable amount of time [6].

C. Reference prediction

In this section we evaluate how well the references can be predicted from trained rules. The accuracy of the conditional pronunciation rules are evaluated directly, in order to gain insight into the predictability of pronunciations under the various combinations of causal factors, using 10-fold cross-validation. In addition, the accuracy of the conditional pronunciation rules are evaluated against the references extracted from the data. The aim is to understand how well the pronunciation rules are able to produce a base reference from which variants can be generated.

1) G2P per speaker language: Here we consider the accuracy with which we can predict the typical pronunciation, of a person with a specific first language, of a name in any of the four languages.

The experiment was performed as follows:

- The effect of name language is ignored temporarily.
- For each name pair, we obtain a reference pronunciation per speaker language (4 references per name).
- Four different dictionaries are created from these references, one per speaker language.
- Finally, we generate pronunciation rules in two different ways:
  - system A: Extract rules directly from the name data; measure accuracy using 10-fold cross-validation. Data is more closely matched, but the training set is very small.
  - system B: Extract rules from generic data, apply to full data set and measure accuracy. Now data is less closely matched, but the training sets are somewhat larger (5,000 to 100,000 words per language).

Results when extracting rules from name data (system A) and generic data (system B) are shown in Table VI. (In this table, both correctness and accuracy are reported.) From the results we see that when name language is not taken into account, task-specific rules outperform generic rules. This may be due to proper names having a less regular spelling system than other more commonly used words in the same language. The generic G2P rules are then insufficient to predict proper name pronunciations accurately.

2) G2P per name language: In this section we consider the accuracy with which we can predict the typical pronunciation of a person with a specific first language of a name in his/her own first language. This can then serve as the basis for adding variants based on the phonemic substitution rules described in Section IV-D.

Experimental setup:

- Only consider pronunciations where the name language is similar to the speaker language.

		Task-specific		Generic	
Lang	Ref	Corr	Acc	Corr	Acc
A	Manual	82.06	79.13	68.95	61.82
	OPTMAX	79.50	75.92	67.48	59.75
	MAXOPT	81.55	78.54	68.52	61.24
E	Manual	80.11	77.57	72.30	67.61
	OPTMAX	77.35	74.34	71.53	66.32
	MAXOPT	80.22	77.47	71.85	67.03
Z	Manual	82.27	78.87	79.94	69.62
	OPTMAX	79.96	76.59	77.78	66.58
	MAXOPT	81.50	78.09	79.39	68.83
S	Manual	80.91	77.76	76.04	67.80
	OPTMAX	78.91	75.30	74.90	65.52
	MAXOPT	81.43	78.29	77.07	68.76

TABLE VI
RESULTS OF G2P PREDICTION (PER SPEAKER LANGUAGE) OF REFERENCES WITH TASK-SPECIFIC AND GENERIC RULES.

- For each name, obtain a reference pronunciation per name language (1 reference per name).
- Create 4 different dictionaries from these references, one per name language.
- Generate pronunciation rules in two different ways:
  - system A: Extract rules directly from the name data; measure accuracy using 10-fold cross-validation. Data is more closely matched, but the training set is very small.
  - system B: Extract rules from generic data, apply to full data set and measure accuracy. Now data is less closely matched, but the training sets are larger (5,000 to 100,000 words per language).

Results when extracting rules from name data (system A) and generic data (system B) are shown in Table VII. The results show that when the name language and the speaker language are the same, these "in-language" reference pronunciations can be predicted more accurately than when name language is ignored (Table VI). Furthermore, generic G2P rules outperform the task-specific G2P rules. This is a direct consequence of the limited data available for training. Of interest is the lower accuracy observed for English, which is to be expected, given its less regular spelling system.

		Task-specific		Generic	
Lang	Ref	Corr	Acc	Corr	Acc
A	Manual	89.03	87.08	88.30	84.19
	OPTMAX	83.04	80.01	84.41	78.93
	MAXOPT	86.44	83.62	86.24	81.40
E	Manual	82.44	79.32	90.36	87.13
	OPTMAX	78.07	74.06	87.19	82.92
	MAXOPT	80.72	76.77	89.10	85.49
Z	Manual	96.79	96.27	97.17	96.36
	OPTMAX	91.36	89.93	93.83	91.66
	MAXOPT	93.41	92.31	94.92	93.11
S	Manual	86.37	85.20	87.17	86.39
	OPTMAX	86.38	84.29	88.76	86.62
	MAXOPT	88.22	87.06	91.66	90.66

TABLE VII
RESULTS OF G2P PREDICTION (PER NAME LANGUAGE) OF REFERENCES WITH TASK-SPECIFIC AND GENERIC RULES.

D. Variant analysis

Speaker language	Name language: A	E	Z	S
A	r → r\	r\ → r	a → A:	O → u
	a → A:	@ → a	O → u	u → O
	r →	z → s	E → i	a → A:
	a → @	{ → a	i → @	i → E
	i@ → i	Q → O	s → z	E → @
E	r → r\	r\ → r	a → @	a → A:
	a → @	@ → a	a → A:	u → O
	a → A:	{ → a	i → @	a → @
	a → {	@ → 3:	O → @u	O → @u
	x → g	E → {	E → @	E → @
Z	r →	r\ → r	A: → a	O → u
	@ → i	@ → a	a → A:	a → A:
	@ → E	@ → i	g →	E → i
	A: → a	Q → O	E → i	u → O
	@ → a	{ → a	k → g	i → E
S	r →	r\ → r	A: → a	O → u
	@ → E	@ → a	z → s	u → O
	A: → a	@ → i	E → i	E → i
	r → r\	Q → O	g → k	a → A:
	@ → i	{ → a	a → A:	A: → a

TABLE VIII
SOME OF THE TOP PHONE SUBSTITUTIONS MADE BETWEEN PRONUNCIATIONS OBSERVED AND THE AUTOMATICALLY EXTRACTED REFERENCES BASED ON THE MAXOPT METHOD FOR DIFFERENT SPEAKER LANGUAGE AND NAME LANGUAGE COMBINATIONS. PHONES ARE IN XSAMPA FORMAT, AND ARE SELECTED FROM THE LWAZI PHONE SETS [5].

In Table VIII we see the top 5 substitutions or deletions made by speakers with different mother tongue languages pronouncing words from different language origins. The insertions are not shown here as they require more context to be meaningful. The results reveal a number of interesting patterns: for example, the approximant /r\/ of English and trilled /r/ of Afrikaans are prone to deletion or interchange in all languages, the voiced/voicing feature in /z/ and /s/ is not stable, etc. It is interesting to note that these "rules" are not the same for the different language combinations, even though some commonalities do exist. In the results here, the automatic references were used to compare the pronunciations from different languages. It is also interesting to observe that when a first language speaker pronounces a name in his/her language, the G2P rules of other languages are sometimes employed, e.g. the "$r \rightarrow r\backslash$" mapping for Afrikaans speakers on Afrikaans names. Clearly, determining the correct linguistic origin of a word, is not an easy task and often ambiguous.

When the manual references are used, the results in Table IX are obtained. Here we see that the rules extracted are very similar to those from the automatic references. This is encouraging as it means that the process of variant generation may not be very sensitive to the accuracy of the references extracted. Consequently, it is possible that good variants may still be generated using the automatically extracted references.

Speaker language	Name language: A	E	Z	S
A	r → r\	r\ → r	a → A:	u → O
	a → @	@ → a	O → u	i → E
	{ → E	z → s	g → k	a → A:
	h →	d →	E → i	O → u
	r →	{ → a	i → @	E →
E	r → r\	d →	a → A:	u → O
	r →	@ → a	a → @	a → A:
	a → @	r\ → r	i → @	i → E
	a → A:	E → {	g → k	a → @
	a → {	→ 3:	E → @	O → @u
Z	r →	r\ → r	a → A:	u → O
	@ → E	@ → a	g →	i → E
	@ → i	@ → i	g → k	a → A:
	j → Z	Q → O	E → i	E → i
	A: → a	{ → a	K → tl_>	O → u
S	r →	r\ → r	a → A:	u → O
	@ → E	@ → a	g → k	i → E
	r → r\	@ → i	z → s	a → A:
	A: → a	Q → O	g →	A: → a
	@ → i	{ → a	E → i	h →

TABLE IX
SOME OF THE TOP PHONE SUBSTITUTIONS MADE BETWEEN PRONUNCIATIONS OBSERVED AND THE MANUALLY CORRECTED REFERENCES FOR DIFFERENT SPEAKER LANGUAGE AND NAME LANGUAGE COMBINATIONS.

V. CONCLUSION

It was shown that reference pronunciations can be extracted in a semi-automatic process. Although a human expert was required to verify and correct some of the entries, the process was relatively fast and efficient, and the benefit of this process will be even more pronounced when larger dictionaries are being developed.

One of our aims with this research was to determine which is most predictable: cross-lingual reference pronunciations directly, or "in-language" reference pronunciations combined with a number of P2P rules to generate additional variants. (For ASR systems it is not necessary to generate the single-best pronunciation, as long as the most commonly occurring variants can be predicted.) We found that there are numerous P2P effects that occur systematically and that these can be used to generate variants using the "in-language" reference pronunciations, which can be predicted with high accuracy.

When the name language is not taken into account, we found that the task-specific G2P rules outperformed the generic rules, suggesting that proper names pronunciations have a less regular spelling system than generic words. However, for "in-language" prediction the generic rules perform very well, suggesting that "in-language" name pronunciations are quite similar to the pronunciation of generic words. It may be that with more data the task-specific G2P rules will still outperform the generic rules. G2P systems for all languages (name and speaker languages) achieve close to 80% phoneme correctness. The only system that does not achieve this level of accuracy is English, which is not surprising given the general complexity of English G2P. If speaker language and name language overlap, reference pronunciations can be predicted with good accuracy.

Much interesting work remains to be done in order to achieve our goal of accurate pronunciation modelling of South African proper names. Most importantly, the accurate "in-

language" results achieved with generic pronunciation rules (see Table VII), along with the regularities in cross-language pronunciations (Table IX) suggest that significantly improved predictions can be obtained by combining these different knowledge sources – perhaps by using a P2P-based approach similar to that in [17]. Comparing the ASR accuracies that can be achieved with these various approaches on the Multipron corpus will also be of great practical interest.

From a linguistic perspective, it will be interesting to see whether the process of cross-language transfer of pronunciations can be characterized more generically. For example, our four languages are from two different language families; it is reasonable to expect that those family relationships will reveal themselves in the cross-lingual pronunciations. A detailed understanding of this process will be helpful in the development of algorithms that can also be applied to all those language pairs for which cross-lingual data is not available.

ACKNOWLEDGMENT

This corpus is being developed in collaboration with Jean-Pierre Martens from the University of Ghent (Belgium) and Oluwapelumi Giwa from North-West University (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.

TERMINOLOGY

Speaker language - This is the first language of a speaker, also called the native language or mother tongue.

Name language - The language of the word's origin is referred to here as the name language or word language.

Correctness and accuracy - These are two closely related measures that can be used to evaluate the performance of a pronunciation prediction system. As the predicted pronunciation and the reference pronunciation may be of different lengths, these two pronunciations are first aligned on a phoneme-to-phoneme basis. When the two pronunciations are aligned, some of the phonemes will match (predicted correctly), others will not (prediction errors). The number of phonemes that match as a percentage of the total number of aligned phonemes is referred to as "phoneme correctness". "Phoneme accuracy" is a stricter measure whereby the total number of incorrectly inserted phonemes are subtracted from the total number of correct phonemes before the percentage is calculated. We use both measures to quantify our ability to predict different reference pronunciations.

In-language reference pronunciation - This is defined as the single pronunciation per name that is produced most often by first language speakers from the language community where the name originated. (For example, the way an isiZulu speaker would produce an isiZulu name, or an Afrikaans speaker an Afrikaans name.)

REFERENCES

[1] O. Giwa, M. H. Davel, and E. Barnard, "A Southern African corpus for multilingual name pronunciation," in *22nd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 2011)*, Nov. 2011, pp. 49–53.

[2] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for ASR: A survey of the literature," *Speech Communication*, vol. 29, no. 2-4, pp. 225–246, 1999.

[3] M. Adda-Decker and L. Lamel, "Multilingual Dictionaries," in *Multilingual Speech Processing*, T. Schultz and K. Kirchoff, Eds. Berlington, MA, USA: Academic Press, 2006, ch. 5, pp. 123–166.

[4] M. Kgampe and M. H. Davel, "Consistency of cross-lingual pronunciation of South African personal names," in *21st Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 2010)*, Nov. 2010, pp. 123–127.

[5] E. Barnard, M. H. Davel, and G. B. van Huyssteen, "Speech technology for information access: a South African case study," in *Proceedings of the AAAI Spring Symposium on Artificial Intelligence for Development (AI-D)*, Mar. 2010, pp. 8–13.

[6] M. H. Davel and O. Martirosian, "Pronunciation dictionary development in resource-scarce environments," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech)*, Sep. 2009, pp. 2851–2854.

[7] B. Erol, J. Cohen, M. Etoh, H.-W. Hon, J. Luo, and J. Schalkwyk, "Mobile media search," in *ICASSP '09: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 4897–4900.

[8] F. Bechet, R. De Mori, and G. Subsol, "Very large vocabulary proper name recognition for directory assistance," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 2001, pp. 222–225.

[9] F. Bechet, R. De Mori, and G. Subsol, "Dynamic generation of proper name pronunciations for directory assistance," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, May 2002, pp. I–745–I–748.

[10] M. H. Davel and E. Barnard, "Pronunciation prediction with Default&Refine," *Computer Speech and Language*, vol. 22, no. 4, pp. 374–393, 2008.

[11] A. F. Llitjos and A. W. Black, "Knowledge of Language Origin Improves Pronunciation Accuracy of Proper Names," in *Eurospeech*, 2001, pp. 1919–1922.

[12] H. van den Heuvel, J.-P. Martens, K. D'hanens, and N. Konings, "The Autonomata Spoken Names Corpus," in *Proceedings LREC*, 2008, pp. 140–143.

[13] H. van den Heuvel, B. Réveil, and J.-P. Martens, "Pronunciation-based ASR for names," in *Interspeech 2009: 10th Annual Conference of the International Speech Communication Association, Vols. 1-5*, 2009, pp. 2959–2962.

[14] B. Réveil, J.-P. Martens, and B. D'Hoore, "How speaker tongue and name source language affect the automatic recognition of spoken names," in *Interspeech 2009: 10th Annual Conference of the International Speech Communication Association, Vols. 1-5*, 2009, pp. 2971–2974.

[15] B. Réveil, J.-P. Martens, and H. van den Heuvel, "Improving proper name recognition by adding automatically learned pronunciation variants to the lexicon," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds., 2010, pp. 2149–2154.

[16] B. Réveil, J.-P. Martens, and H. van den Heuvel, "Improving proper name recognition by means of automatically learned pronunciation variants," *Speech Communication*, vol. 54, no. 3, pp. 321–340, 2012.

[17] 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*, 2006, pp. 287–292.

[18] H. van den Heuvel, J.-P. Martens, and N. Konings, "G2P conversion of names : what can we do (better)?" in *Interspeech 2007: 8th Annual Conference of the International Speech Communication Association*, vol. 1-4, 2007, pp. 1181–1184.

[19] F. Stouten and J. Martens, "Dealing with cross-lingual aspects in spoken name recognition," in *IEEE Workshop on Automatic Speech Recognition and Understading, Vols. 1-2*, 2007, pp. 419–424.

[20] M. H. Davel, C. J. van Heerden, and E. Barnard, "Validating smartphone-collected speech corpora (accepted for publication)," in *Proc. Spoken Language Technologies for Under-resourced Languages (SLTU)*, May 2012.

[21] Meraka-Institute, "Lwazi ASR corpus," http://www.meraka.org.za/lwazi, 2009.