Extending an Afrikaans pronunciation dictionary using Dutch resources and P2P/GP2P

Linsen Loots[1], Febe de Wet[1,2], Thomas Niesler[1]
[1]Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa.
[2]HLT Research Group, CSIR Meraka Institute, South Africa.
{linsen, fdw, trn}@sun.ac.za

***Abstract*—In this study we explore the extension of a small Afrikaans pronunciation dictionary by applying phoneme-to-phoneme (P2P) and grapheme-and-phoneme-to-phoneme (GP2P) conversion to an existing and more extensive Dutch pronunciation dictionary. This is compared to the more common approach of extending the Afrikaans dictionary by means of grapheme-to-phoneme (G2P) conversion. The results indicate that the Afrikaans pronunciations obtained by P2P and GP2P from the Dutch dictionary are more accurate than the corresponding pronunciations obtained by the application of G2P. This result indicates that under-resourced languages can take advantage of existing and more extensive pronunciations available in a closely-related and better-resourced language in order to improve the extent and quality of a pronunciation dictionary.**

Index Terms: Dutch, Afrikaans, pronunciation modelling, G2P, P2P, GP2P, decision trees

I. Introduction

Extensive phonetic resources are required to develop automatic speech recognition (ASR) and text-to-speech (TTS) systems. The compilation of these resources is an expensive endeavour because it involves specialised human labour. However, it may be possible to expedite the process for new languages by capitalising on existing resources available in related and better-resourced languages. This study elaborates on previous research investigating Dutch and Afrikaans as an example of two such closely-related languages [1], [2]. While extensive resources have already been developed for Dutch, Afrikaans is still severely under-resourced in terms of human language technology (HLT), as are all the indigenous languages of South Africa.

The pronunciation dictionary is a key component of ASR and TTS systems, which are required for the implementation of voice-enabled HLT applications. This dictionary provides the most probable pronunciation(s) of all words in the system vocabulary in terms of a pre-defined phoneme set. These pronunciations can be generated from scratch by human experts, but this is a very expensive process. Alternatively, they can be generated automatically by rule-based or data-driven techniques. However, automatic methods require training data and their output must be verified because the pronunciations they provide are in general not always correct. The challenge therefore is to find a strategy that can learn efficiently and correctly from existing, verified data.

Grapheme-to-phoneme (G2P) conversion is the process of automatically determining a word's pronunciation from its spelling. Data-driven G2P techniques utilise machine learning algorithms to discover the correspondences between the graphology (graphemes) and pronunciation (phonemes) of words, and exploit these to automatically determine the pronunciations of unknown words.

In contrast, phoneme-to-phoneme (P2P) conversion allows the pronunciation of a a new word in a target accent or language to be derived from a known pronunciation in a source accent or language, rather than from the spelling. It has been shown that, when considering accents of the same language, P2P can give more accurate results than G2P conversion [3]. Furthermore, this accuracy can be further improved when both the graphemes and the source phonemes are used as input, yielding a technique known as grapheme-and-phoneme-to-phoneme (GP2P) conversion.

In this study, we compare two methods that can be used to generate the pronunciations of Afrikaans words automatically. First, we consider the straightforward application of G2P methods to a set of existing Afrikaans words and their pronunciations in order to provide pronunciations for words not in this set. Second, we determine the pronunciations of the same new words by the application of P2P and GP2P methods to the corresponding Dutch pronunciations, which are available in a much more extensive Dutch dictionary.

The following two sections elaborate on the lexical relationship between Afrikaans and Dutch, and discuss G2P, P2P and GP2P conversion. This is followed by a description of the experimental materials and method. Finally, the experimental results are presented, followed by a discussion and conclusions.

II. The lexical relationship between Afrikaans and Dutch

In [1] and [2] various aspects of Afrikaans and Dutch as an example of two closely-related languages are discussed. In this study, we will focus on the lexical relationship between the two languages. As was pointed out in [2], 90-95% of all lexical items in Afrikaans are of Dutch origin. However, many of these lexical items are no longer graphologically identical due to the changes in spelling, phonology and morphology that occurred during the development of Afrikaans. A distinction is therefore made between identical cognates, non-identical cognates, false friends and non-cognates [2].

Identical cognates are lexical items that are graphologically identical in Dutch and Afrikaans, and which can be ascribed to linguistic inheritance, e.g. *boom, tafel, etc.* Non-identical cognates are words which are etymologically related in Afrikaans and Dutch, but which differ systematically in terms of graphology. Most of the words in this category are inflected word-forms, e.g. the Afrikaans verb *praat* has a number of possible inflections in Dutch *praat, praten, etc.* False friends are words which are exactly the same in Afrikaans and Dutch but, due to semantic processes or referent changes, no longer have the same meaning, e.g. *amper* which means "almost" in Afrikaans but "almost not" in Dutch. Non-cognates refer to word forms that are graphologically unrelated, but which have the same meaning, e.g. the Afrikaans word for "banana" is *piesang* while the Dutch word is *banaan*.

In this study we focus on the relationship between pronunciations in the two languages. Non-cognates are therefore not relevant, because they are graphologically and phonetically unrelated. As a result, the Dutch pronunciations of these words will not give any indication of the pronunciations of their Afrikaans equivalents. If the systematic differences between the non-identical cognates were known or can be learnt from suitable data, the differences in form and pronunciation could be combined during the pronunciation prediction process. However, the lexical scope of this investigation will be restricted to identical cognates and false friends. Even though the false friends are semantically different in the two languages, they are graphologically and phonetically related, and the Dutch pronunciation can therefore be used to predict its Afrikaans counterpart. The application of our techniques to other lexical categories in Dutch and Afrikaans will be explored in future research.

III. Pronunciation conversion

As mentioned in Section I, P2P conversion refers to the use of a word's pronunciation in another accent or a closely related language, to determine its pronunciation in the target accent or language [3]. This is done using the algorithms commonly applied to G2P conversion, and it is therefore useful to give a brief overview of G2P conversion.

A. G2P conversion

G2P conversion employs machine learning to convert the known graphology of a word into its unknown pronunciation. Several data-driven G2P methods have been suggested in the literature, including decision trees [4], [5], [6], [7], HMMs [8], pronunciation by analogy [9], default&refine [10] and memory-based learning [11]. Decision trees have been shown to yield competitive accuracy in G2P conversion [12] and to be effective for P2P conversion [3].

For G2P, P2P and GP2P conversion, we employed deterministic binary decision trees, for which each node is associated with a true/false question regarding the input grapheme and its context. The tree is traversed from the root by recursively using the answer of each node's question to determine which child node to choose. Each leaf node is associated with an output phoneme, which constitutes the classification result [13].

Decision trees are grown recursively. For each new node the available training data are split according to all possible questions. The question which results in the greatest entropy gain is then chosen for that node [5]. A more detailed discussion of decision trees and their use for G2P, P2P and GP2P conversion can be found in [3].

For G2P conversion, the graphemes and their context form the input of the decision tree classifier, and the phonemes the output class. Pronunciations are generated by sequentially presenting the graphemes and their context to the tree, assigning an output phoneme to each, and concatenating these output phonemes.

B. P2P and GP2P conversion

P2P conversion has been successfully used to convert pronunciations between different accents of English [3]. In contrast to G2P conversion, for P2P the phonemes comprising the source pronunciation are used as input to the decision tree, rather than the graphemes. In this study, Dutch is regarded as the source language, and Afrikaans as the target language.

GP2P conversion provides a further extension of the P2P algorithm. Here both the graphemes and the phonemes comprising the pronunciation in the source language are used as input to the decision tree [3]. In order to accomplish this, the source graphemes and phonemes are first aligned by means of dynamic programming.

IV. Data

A. Afrikaans pronunciation dictionary

We use the Afrikaans pronunciation dictionary described in [14] for our experimental evaluations. This dictionary was developed as part of a project on Resources for Closely Related Languages (RCRL) and is therefore known as the RCRL Afrikaans pronunciation dictionary (APD). The RCRL APD contains pronunciations for more than 24 000 words, transcribed using SAMPA[1]. Its development was bootstrapped using approximately 5 000 words from the Afrikaans Lwazi dictionary [15] and extended by adding the most frequent outstanding Afrikaans words, in order of descending frequency of occurrence. Frequency of occurrence was estimated from the so-called *Taalkommissie Korpus*, a 60 million word, stratified corpus which was compiled by the Afrikaans Language Commission as an example of standard, formal Afrikaans in its written form. An automatic pronunciation dictionary verification procedure, based on an analysis of conflicting pronunciation rules, was used to find and correct systematic errors in the RCRL ADP [14].

B. Dutch pronunciation dictionary

Dutch pronunciations were obtained from Elex, a Dutch lexical database that is available in electronic format from the Dutch Centre for Lexical Information [16]. This database

[1]Speech Assessment Methods Phonetic Alphabet.

contains over a million entries, including around 200 000 lemma entries and additional information such as part-of-speech, usage, and syntax. We employed a sub-set of the Elex database, consisting of those words/lemmas for which manually verified, canonical pronunciations are provided. These pronunciations comply with the CELEX (the Dutch Centre for Lexical Information) transcription standard and are coded in the phoneme set defined for the Spoken Dutch Corpus (Corpus Gesproken Nederlands (CGN)) [16], [17]. In order to facilitate comparison and experimentation, the chosen Elex pronunciations were also mapped to the SAMPA phoneme set.

V. Experimental method

Three experiments were carried out to determine whether the Dutch dictionary can be successfully used as a source of Afrikaans pronunciations. Firstly, a direct comparison between the two dictionaries was performed. This indicates how successfully Dutch pronunciations can be used without any modification, and provides a baseline for the later G2P, P2P and GP2P experiments. Secondly, G2P conversion was used to provide Afrikaans pronunciations from the existing RCRL APD dictionary. This is a conventional approach and provides a second performance baseline. Finally, P2P and GP2P conversion are employed to derive Afrikaans pronunciations from Elex. These results will indicate whether the availability of Dutch pronunciations can be used to improve the accuracy with which Afrikaans pronunciations can be derived.

While the same methods can be applied in reverse, i.e. from Afrikaans to Dutch, this has been omitted since the focus of this study is on obtaining pronunciations for the less well-resourced language.

A. Training and test data

The words in Elex for which manually verified CELEX pronunciations are provided were first ordered according to their frequency of occurrence in the CGN. The 50 000 most frequent words were subsequently automatically translated into Afrikaans using a rule-based Dutch-to-Afrikaans converter [2], [18]. This automatic translation identified more than 13 000 of the 50 000 most frequent words as identical cognates. Of these identical cognates, 5 340 also occur in the RCRL APD. These 5 340 words were used for all the experiments presented in this study. Where Elex contained multiple pronunciations for a single word, only the most frequent pronunciation was used.

B. Cross-validation

There are a number of parameters that need to be specified when training decision trees for G2P, P2P and GP2P conversion. These parameters include the context window size, the direction of the conversion process, and the amount of data reserved for pruning. In order to obtain values for these parameters, the phoneme accuracy of the decision tree was optimised on a held-out development set within a 10-fold cross-validation framework. This was achieved by dividing the 5 340 words into 10 non-overlapping and approximately equally-sized partitions. Reserving each partition in turn for later use as a test set, the remaining 9 partitions (corresponding to 90% of the data) were again divided into 10 equally-sized partitions. The first of these sub-partitions served as a development set, and the remaining 9 as a training set for parameter optimisation. Decision trees were trained on this training set for a range of parameter values, and those that lead to the highest phoneme accuracy, measured on the associated development set, were identified as optimal. This process was repeated for each of the 10 partitions of the 5 340-word data set. In general, different optimal parameter values were obtained in each case. Finally, each of the 10 partitions of the 5 340-word data set was employed as a test set, while decision trees were trained using the remaining 9 partitions and the corresponding optimal parameters.

Results are presented in terms of phoneme and word accuracy. The generated pronunciations for each of the 10 test partitions were aligned with the corresponding dictionary pronunciations by means of dynamic programming. From these alignments, the number of substitutions, insertions and deletions were determined. The phoneme accuracy was subsequently calculated using Equation 1, where N_c, N_i and N_t are the numbers of correct, inserted and total phonemes respectively.

$$Acc = \frac{N_c - N_i}{N_t} \tag{1}$$

Word accuracy indicates the percentage of words for which the generated and correct pronunciations are identical. In each case the reported percentages are averages for the 10 cross-validation splits, with 95% confidence intervals calculated using the bootstrap method described in [19].

The same 10 disjoint partitions of the 5 340-word data set were used for training and testing decision trees for all G2P, P2P and GP2P experiments.

VI. Experimental results

The first experiment was to perform a direct comparison of the Dutch and Afrikaans pronunciations for the 5 340 words. This provides an indication of the similarity between the two dictionaries, and thereby provides a baseline with which later results can be compared.

In order to perform the comparison, and since the RCRL APD is transcribed using SAMPA, Elex pronunciations were mapped to the SAMPA phoneme set. Corresponding pronunciations from the Elex and RCRL APD were aligned by dynamic programming to determine the number of substitutions, insertions and deletions, and thus the phoneme accuracy. The first line of Table I shows the results of the direct comparison. It is clear that there is a high level of correspondence between the two dictionaries, with 95% of phonemes and 74% of words matching.

In order to obtain a further baseline, G2P conversion was applied to the Afrikaans RCRL APD dictionary. Holding out each of the 10 disjoint partitions of the 5 340-word data set as a test set in turn, a G2P decision tree was trained on the remaining 9 partitions within a 10-fold cross-validation

Source	Target	Method	Phoneme	Word
Dutch (SAMPA)	Afrikaans	Direct	95.01 $\pm$ 0.82%	73.77%
Afrikaans	Afrikaans	G2P	96.80 $\pm$ 0.69%	84.18%
Dutch (CGN)	Afrikaans	P2P	97.26 $\pm$ 0.62%	85.37%
Dutch (SAMPA)	Afrikaans	P2P	97.34 $\pm$ 0.60%	85.69%
Dutch (CGN)	Afrikaans	GP2P	97.67 $\pm$ 0.58%	87.67%
Dutch (SAMPA)	Afrikaans	GP2P	97.65 $\pm$ 0.58%	87.68%

TABLE I
PHONEME AND WORD ACCURACIES OF AFRIKAANS PRONUNCIATIONS DETERMINED BY VARIOUS APPROACHES.

framework, as described in Section V-B. This decision tree was then used to obtain pronunciations for the words in the withheld partition, and these pronunciations were scored against the corresponding pronunciations in the RCRL APD dictionary by dynamic programming, as also described in Section V-B.

The second line of Table I shows the results of G2P conversion applied to the RCRL APD. It indicates that 97% of phonemes and 84% of words can be correctly determined using G2P.

The final experiments, which are the focus of this paper, were the application of the P2P and GP2P strategies to the conversion of Dutch to Afrikaans pronunciations. As only identical cognates and false friends (for which the graphology is the same in both languages) are currently being used, the graphemes of a word are the same in both dictionaries. It does therefore not matter from which dictionary the graphemes are used when training or testing the decision trees for GP2P conversion. The same decision tree training and testing strategy and the same 10-fold data splits used for the G2P experiments above were employed again for both the P2P and GP2P experiments, making the results directly comparable.

Lines three and four of Table I show the results of the P2P experiments. Accuracies are given for the case in which Elex uses its native CGN phoneme set, as well as for the case where its pronunciations are mapped to SAMPA. In both cases, 97% of phonemes are correctly determined, although SAMPA gives a slightly higher word accuracy.

The last two lines of Table I show the results of GP2P experiments. Results are again given for both the CGN and SAMPA phoneme sets. In both cases 98% of phonemes and 88% of words are correctly determined.

VII. DISCUSSION AND CONCLUSIONS

From Table I, we see that for Afrikaans, G2P conversion achieves a phoneme accuracy of 96.80% and a word accuracy of 84.18%. These results are noteworthy when compared to similar G2P experiments performed for English, where phoneme accuracies of approximately 91% are common [4], [5], [3]. This indicates that Afrikaans has a very regular relationship between orthography and pronunciation. It is also worth noting by comparing the first two lines of Table I that G2P conversion provides more accurate Afrikaans pronunciations than using the Dutch pronunciations directly. In particular, G2P provides a significant absolute improvement of 1.79% in phoneme accuracy.

When using P2P conversion to derive Afrikaans pronunciations from Dutch pronunciations, the phoneme accuracy rises to 97%. While this is more accurate than G2P conversion, the improvement is not statistically significant.

Best performance is achieved when using GP2P conversion to derive Afrikaans pronunciations from Dutch pronunciations. In this case, phoneme accuracies of 97.7% and word accuracies of 87.7% are attained. While the improvement of GP2P over P2P is relatively small, the results in Table I show that GP2P conversion represents a statistically significant improvement over G2P conversion.

For P2P conversion, using the SAMPA phoneme set for Elex provides slightly better results than CGN. For GP2P, however, CGN provides slightly better results. In neither case the difference is significant, however, suggesting that the decision trees are able to compensate for most differences between the phoneme sets.

In conclusion, it has been demonstrated that, even though Afrikaans has a very regular relationship between its orthography and pronunciation, the application of the P2P and especially GP2P approaches to convert Dutch to Afrikaans pronunciations is more effective than than the application of G2P to the Afrikaans pronunciations alone. This suggests that there is scope for utilising the more extensive Dutch resources for the development of Afrikaans pronunciation dictionaries, thereby allowing HLT applications to be developed for Afrikaans more quickly and cheaply.

Future research will focus on testing whether this conclusion holds true for non-identical cognates. Further investigations can also test whether the results presented are valid for other languages.

VIII. ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation (NRF) as part of grant FA207041600015 related to HLT Resources for Closely-Related Languages, and was executed using the High-Performance Computing (HPC) facility at Stellenbosch University. We would also like to thank Liesbeth Augustinus for her work on the automatic translation of the words in Elex.

REFERENCES

[1] W. Heeringa and F. de Wet, "The origin of Afrikaans pronunciation: a comparison to west Germanic languages and Dutch dialects," in *Proc. Conference of the Pattern Recognition Association of South Africa*, F. Nicolls, Ed., Cape Town, South Africa, 2008.

[2] G. B. van Huyssteen and S. Pilon, "Rule-based Conversion of Closely-related Languages: A Dutch-to-Afrikaans Converter," in *Proc. Conference of the Pattern Recognition Association of South Africa*, F. Nicolls, Ed., Stellenbosch, South Africa, 2009, pp. 23–28.

[3] L. Loots and T. Niesler, "Automatic conversion between pronunciations of different English accents," *Speech Communication*, p. DOI: http://dx.doi.org/10.1016/j.specom.2010.07.006, 2010.

[4] A. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," in *Proc. ESCA Speech Synthesis Workshop*, Jenolan Caves, Australia, 1998.

[5] J. Suontausta and J. Häkkinen, "Decision tree based text-to-phoneme mapping for speech recognition," in *Proc. ICSLP*, Beijing, China, 2000.

[6] A. K. Kienappel and R. Kneser, "Designing very compact decision trees for grapheme-to-phoneme transcription," in *Proc. Eurospeech*, Aalborg, Denmark, 2001.
[7] G. Webster and N. Braunschweiler, "An evaluation of non-standard features for grapheme-to-phoneme conversion," in *Proc. Interspeech*, Brisbane, Australia, 2008.
[8] P. Taylor, "Hidden markov models for grapheme to phoneme conversion," in *Proc. Interspeech*, Lisbon, Portugal, 2005.
[9] Y. Marchand and R. Damper, "A multistrategy approach to improving pronunciation by analogy," *Computational Linguistics*, vol. 26, no. 2, pp. 195–219, 2000.
[10] M. Davel and E. Barnard, "Pronunciation prediction with default&refine," *Computer Speech and Language*, vol. 22, no. 2, pp. 374–393, 2008.
[11] W. Daelemans, A. van den Bosch, and J. Zavrel, "Forgetting exceptions is harmful in language learning," *Machine Learning*, vol. 34, pp. 11–41, 1999.
[12] K.-S. Han and G.-L. Chen, "Letter-to-sound for small-footprint multilingual TTS engine," in *Proc. ICSLP*, Jeju, Korea, 2004.
[13] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Pacific Grove: Wadsworth & Brooks, 1984.
[14] M. Davel and F. de Wet, "Verifying pronunciation dictionaries using conflict analysis," in *Proc. Interspeech*, Makuhari, Chiba, Japan, 2010, pp. 1898–1901.
[15] M. Davel and O. Martirosian, "Pronunciation dictionary development in resource-scarce environments," in *Proc. Interspeech*, Brighton, UK, 2009, pp. 2851–2854.
[16] R. G. L. Centre for Lexical Information, U. o. N. Speech, and N. Max Planck Institute for Psycholinguistics, "CELEX Dutch database, version 3.2, http://tst.inl.nl," 1998.
[17] T. H. Dutch Language Union, "Spoken Dutch corpus (Corpus Gesproken Nederlands (CGN)), version 1.0, http://www.inl.nl/en/lexica/cgn-lexicon," 2004.
[18] "Dutch to Afrikaans Converter (D2AC), http://d2ac-a2dc.sourceforge.net/," 2004.
[19] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. ICASSP*, Montreal, Canada, 2004.