Comparing grapheme-based and phoneme-based speech recognition for Afrikaans

Willem D. Basson[1,2]
[1]Human Language Technology Competency Area
CSIR Meraka Institute
[2]Multilingual Speech Technologies
North-West University,
Vanderbijlpark, South Africa
Email: wlmbasson@gmail.com

Marelie H. Davel
Multilingual Speech Technologies
North-West University,
Vanderbijlpark, South Africa
Email: marelie.davel@gmail.com

***Abstract*—This paper compares the recognition accuracy of a phoneme-based automatic speech recognition system with that of a grapheme-based system, using Afrikaans as case study. The first system is developed using a conventional pronunciation dictionary, while the latter system uses the letters of each word directly as the acoustic units to be modelled. We ensure that the pronunciation dictionary we use is highly accurate and then investigate the extent to which ASR performance degrades when the dictionary is removed. We analyse this effect at different data set sizes and classify the causes of performance degradation. With grapheme-based ASR outperforming phoneme-based ASR in certain word categories, we find that relative error rates are highly dependent on word category, which points towards strategies for compensating for grapheme-based inaccuracies.**

I. Introduction

In an automatic speech recognition (ASR) system, words are traditionally represented as a sequence of acoustic sub-word units such as phonemes [1]. The mapping from these sub-word units to words are usually contained in some sort of lexicon, that is, a pronunciation dictionary. The overall performance of ASR systems is strongly dependent on the accuracy of the pronunciation dictionary and best results are usually obtained with hand-crafted dictionaries, which often requires expert knowledge. Development of these dictionaries is a time-consuming, costly and labour-intensive process. If expert knowledge is either unavailable or too costly, manually developed or statistical grapheme-to-phoneme (g2p) rules can be used to generalise from small data sets [1]. However, these methods typically produce less accurate results.

Earlier work in grapheme-based systems has shown that for regular languages – languages that exhibit a close relationship between graphemes and phonemes – phone-based dictionary development may be unnecessary [1], [2], [3]. Using grapheme-based sub-word units eliminates the need for expert knowledge and saves time and cost. Other advantages include simplified lexicon definition and relatively noise-free pronunciation models [4].

The regularity of a language can be measured based on g2p consistency: using the average accuracy that is obtained at a specific dictionary size when extracting g2p rules. According to this measure, languages vary considerably, from highly irregular languages such as English, to highly regular languages such as Flemish, with Afrikaans being somewhere in between [5].

Some of the earliest work done on grapheme-based speech recognition proposes using polygraphs i.e. letter based units constructed from the orthographic word form with arbitrary length left and right contexts as sub-word units [3]. More recent work include context-dependent grapheme-based recognisers [1] as well as using a decision tree based on graphemic acoustic sub-word units together with phonetic questions [2].

For this paper we developed a grapheme-based ASR system alongside a phoneme-based ASR system using the same standardised approach in both, in the one case using tied-state triphones and the other, tied-state trigrams. With the only variable between the systems being their respective pronunciation dictionaries, this allows for a fairly direct comparison of strengths and weaknesses.

The remainder of this paper is structured as follows: Section II describes the approach followed, both to construct the gold standard phonemic dictionary and to compare grapheme-based and phoneme-based performance. The data used is presented in section III. The various experiments are described and results presented in section IV. Finally, the paper is ended by a summary of our main observations in section V.

II. Approach

We develop comparable grapheme-based and phoneme-based ASR systems for different training data sizes ranging from 5 to 40 hours, and compare word error rate (WER) using independent test sets and 4-fold cross validation. For the comparison to be fair, we need to ensure the pronunciation dictionary is as accurate as possible. The most comprehensive Afrikaans dictionary currently available is the *Resources for Closely Related Languages Afrikaans pronunciation dictionary (rcrl_apd)* [6]. This dictionary however does not include all the words in the data set we are modelling. The process to develop and verify a more comprehensive dictionary is of interest and results relating to this process are included in this paper.

A. Pronunciation Dictionaries

We develop 3 different pronunciation dictionaries. Firstly, we develop a manually verified pronunciation dictionary which serves as a gold standard. It should be noted that this dictionary contains pronunciation variants where appropriate. The total effort in verifying all the sub-word units is lessened by utilising methods such as:

- known word extraction: accepting known pronunciations from existing dictionaries;
- decompounding unknown words and matching these to known components in existing dictionaries;
- short word extraction: analysing short words – which are often non-standard words such as abbreviations or acronyms – separately; and
- the classification of word types to be pre-processed by appropriate g2p methods.

All automated methods used to produce pronunciations were manually verified, which allow us to report on the success rates of each of the automated methods. Since Afrikaans contains many compound words, we focused our effort on identifying known compounds from existing dictionaries, using both a form of longest string matching (LSM) and automated morphological decomposition to achieve this aim.

Secondly, the best possible rule set available to date – rules extracted from the *rcrl_apd* pronunciation dictionary [6] – was used to create an automated (state-of-the-art g2p) pronunciation dictionary. Finally, a minimal effort grapheme-based dictionary was developed by simply splitting the orthographical form of words into space-separated single letters.

Given the gold standard dictionary, the relative accuracy of the g2p dictionary is calculated by measuring the difference between pronunciations. Calculating the accuracy of the grapheme-based dictionary is done by converting every grapheme to its default phoneme based on g2p rules and measuring pronunciation similarity relative to the gold standard dictionary. The relationship between differences in dictionaries and resulting WER is investigated.

B. *ASR accuracy*

ASR systems are analysed and compared in terms of WER. All test sets are recognised using the same flat language model containing all the words in the entire data set. While better recognition accuracy can be obtained using a statistical language model, we specifically want to evaluate the effect of the acoustic models without recognition being guided by a language model. This means that the systems are evaluated and compared in terms of WER with the only difference between systems being their pronunciation dictionaries. (For the later category-based analysis, it is particularly important that categories are not influenced by the language model used.)

C. *Error classification*

ASR recognition errors are classified according to word type and compared across systems. Word types include (1) abbreviations, (2) acronyms, (3) foreign words, (4) generic Afrikaans words, (5) partial words, (6) proper names, (7) concatenated words, (8) spelling errors, (9) spelled out words, (10) single spelled out characters and (11) unknown words. Word type categories were determined during the development of the manually verified pronunciation dictionary. Words that belong to more than one category (due to pronunciation variants or context) are classified as multi-category words. Pronunciation variation caused all but one abbreviation to be classified as multi-category words.

III. Data selection

Afrikaans was selected as the experimental language due to its g2p regularity (fairly regular without being fully regular) and the authors' inherent familiarity with the language. The dataset used is a subset of the NCHLT corpus [7] and has a total length of approximately 64 and a 1/2 hours, consisting of 75 150 utterances from 167 speakers with a male to female ratio of 48.5/51.5. Every utterance in this dataset passed basic quality control checks namely: clipping detection, volume detection and speech cutting detection [8]. Also, to ensure a well balanced dataset every speaker contributes exactly 450 utterances. From this dataset a development set of approximately 2 hours and 45 minutes was held out. The remaining utterances were split into 4 folds with 4 mutually exclusive test sets. Each fold's train set is roughly 46 hours long and contains 54 000 utterances from 120 different gender balanced speakers. All 4 the training sets were then individually subdivided into 46 total random, non-sequential incremental segments. In effect each segment contains approximately one hour more data than the previous one. Finally, to study the effect of phone-based and grapheme-based ASR on varying sizes of training data, segments 5, 10, 20 and 40 were selected for training.

F	# utt trn	# hr trn	# spkr trn	# utt tst	# hr tst	# spkr tst
1	54000	46:18:56	120	18000	15:25:9	40
2	54000	46:51:34	120	18000	14:52:31	40
3	54000	45:51:57	120	18000	15:52:8	40
4	54000	46:9:50	120	18000	15:34:15	40

TABLE I
Data selection: Number of utterances (utt), hours (hr) of audio data and number of speakers (spkr) in train (trn) and test (tst) sets across folds (F)

F	seg 5	seg 10	seg 20	seg 40
1	05:05:24	10:05:53	20:07:59	40:14:12
2	05:06:05	10:11:15	20:24:25	40:45:14
3	05:02:28	10:00:23	19:55:38	39:53:24
4	05:02:50	10:03:34	20:05:03	40:07:01
# utt	5870	11740	23479	46957

TABLE II
Training segments: Hours of audio data and number of utterances per segment (seg) across folds (F)

IV. Experiments and Results

Experiments relating to the development of the gold standard pronunciation dictionary are described in sections IV-A to IV-C, while sections IV-D and IV-E compare the ASR results obtained using the three different dictionaries (the gold standard phoneme-based dictionary, the g2p-predicted dictionary and the grapheme-based dictionary).

A. *Identifying known constituents in compounds*

As discussed in section II, we experimented with two different approaches to decompounding. Note that the primary purpose was to lessen the total effort in creating a pronunciation dictionary: not to find linguistic compounds as such, but only to find known constituents from existing dictionaries (i.e. where pronunciations are known.) Since Afrikaans contains many compounds, many words in a word list would be flagged as unknown when measured against existing dictionaries, while the constituents are actually known and pronounced in an identical manner.

In the remainder of this section we describe the two approaches used (variants of Morfessor-based decompounding and Longest String Matching), the post-processing that is required (which is similar for both approaches), and the results achieved.

1) Morfessor: Morphological decomposition was performed using a modified version of Morfessor 1.0 [9], a popular language independent tool for performing unsupervised morphological decomposition. We changed the tool to only use existing words as 'morphemes' and not to create smaller linguistic components, in effect changing it into a decompounding tool. All other settings were left at their default values.

Given as input is a combination of unique words from an existing dictionary and all words with unknown pronunciations, Morfessor then suggests segmentations for all words, based on identified segments that exist as individual words in an existing dictionary. Words that can be segmented are flagged as candidate compounds, new pronunciations are generated based on the pronunciations of the individual words and prepared for review.

2) LSM: An imperfect version of Longest String Matching algorithm similar to that of [10] was used. The difference being that the longest left hand match is performed at the same time as the longest right hand match, possibly causing overlap and missing some compounds. A limited valence morpheme list is used containing only two valance morphemes, namely *s* and *en*. Using a lexicon of known words as a reference, the largest left- and right hand matching strings of each candidate compound is determined. Words are then flagged as possible compounds if: (a) after subtraction of the left and right match, there is no remainder and the length of the compound is equal to the combined length of the largest left and right match, or (b) the remainder of the compound is either a valid word from the lexicon, or (c) the remainder is a valid valence morph from the limited list.

3) Post-processing: After each decompounding method the pronunciations of compound constituents are extracted from existing dictionaries, residual consonant doubling caused by constituent concatenation is removed, and finally, flagged compounds and their accompanying phone strings are manually verified.

4) Results: After verification, we found 1 492 compounds in the data set (containing 3 225 unique words) of which 1 416 had correct pronunciations. A breakdown of our results are shown in Table III. Morfessor decomposition was applied first, then LSM-based decomposition. Note that LSM-based decomposition was only performed on words that Morfessor was not able to decompound, resulting in 179 additional compounds. Since we are not interested in finding linguistically accurate compound boundaries some of the words identified are not actual compounds, yet they still produce correct phone strings. Table IV summarises the effect of decomposition on pronunciation. Most pronunciation errors relate to a few small morphemes (‘ver’, ‘end’, ‘bes’) that were incorrectly predicted as /E/ rather than /@/ (using SAMPA notation).

	Total flagged	Correctly identified	Incorrectly identified
LSM	203	179	24
Morfessor	1419	1313	106

TABLE III
Breakdown of LSM and Morfessor based decomposition showing the number of correctly identified and incorrectly identified compounds

	Pronunciation		
	correct	error	% correct
Correctly decomposed	1 416	76	94.6
Incorrectly decomposed	130	119	8.5

TABLE IV
Effect of decomposition on pronunciations

B. Developing a gold standard dictionary

As described earlier (in section II-A), in order to lessen the total effort of classifying, predicting pronunciations for and verifying 9 375 unique words, we employed various strategies. Initially, all known words from existing dictionaries were extracted: this comprised nearly two thirds of the dictionary. Remaining words were then checked against known word lists and classified as either valid Afrikaans words, valid English words or unknowns words. All valid English words were then removed, their pronunciations predicted with English g2p rules and these were manually verified. The remaining words were then processed concurrently by the two different decompounding methods described in section IV-A1.

Short word extraction was then performed on the remaining words by extracting all words with a length of 1-4 characters. The vast majority of these words fell into the category of spelled out Afrikaans words. High numbers of partials, abbreviations and acronyms were also present. Words were then categorised and pronunciations were generated with appropriate g2p methods after which all words were reviewed manually. A hand made list was crafted for all spelled out single characters. For the remaining 1 351 words pronunciations were predicted and manually verified. All manual verification was performed by two verifiers.

Results for each step in this process is given in Table V.

Process	Words identified	Valid categories	Valid pron
extr known Afr words	5 925	5 925	5 925
g2p valid Eng	225	189	163
id comps (morfessor)	1 419	1 313	1 265
extract short words	253	196	-
id comps (LSM)	203	179	151
review remaining	1 351	-	-

TABLE V
Per step of the dictionary development process: the number of words correctly identified and the number of valid pronunciations prior to manual correction

C. Dictionary analysis

Using the gold standard dictionary as a reference the phoneme accuracy of the g2p dictionary measured 96.31% with 85.33% of words being identical. This indicates that there is a strong similarity between the two dictionaries. A relative phoneme accuracy of 63.27% was obtained by comparing the grapheme dictionary to the gold standard dictionary. The categorisation of specific differences still requires further investigation. Our findings are presented in Table VI.

Dictionary	Total words	Total phones	Words correct	Phone accuracy
phone	9 374	78 621	-	-
graph	9 374	86 883	6.37%	63.27%
g2p	9 374	78 063	85.33%	96.31%

TABLE VI
Relative phoneme accuracy and percentage of correct words for the g2p dictionary and grapheme dictionary using the gold standard dictionary as reference

D. Effect of dictionary on WER

To evaluate the effect of the dictionaries, we develop three different ASR systems using a relatively standard approach. We use the hidden Markov model toolkit (HTK) [11] and develop context-dependant tied-state acoustic models. Feature extraction on the speech audio data realised 13 Mel Frequency Cepstral Coefficients (MFCCs) with their first and second order derivatives as 39 dimensional feature vectors. MFCC window size was set at 25ms with a frame rate of 10ms. Cepstral mean normalisation was applied at speaker level. With regard to modelling structure, each triphone or trigraph has three emitting states with eight Gaussian mixtures per state and a diagonal covariance matrix. Where parameters are optimised, the development set is used.

Figure 1 shows the effect of different dictionaries on WER at four different training sizes of 5, 10, 20 and 40 hours. At the smallest data set size (5 hours) the gold standard dictionary outperforms the other approaches, with the g2p-based system also outperforming the grapheme-based system. At the largest data set size (40 hours) the grapheme-based system had a WER of 41.13%, the g2p-based system a WER of 39.82% and the phoneme-based system a WER of 38.03%. As is evident in the convergence of WER between the phoneme-based and grapheme-based ASR systems, the more training data that is available the less the degradation in performance is of the grapheme-based ASR system.

Figure 2 shows the difference in relative percentage of WER between (1) grapheme-based and g2p-based ASR, (2) grapheme-based and phoneme-based ASR and (3) g2p-based and phoneme-based ASR. The highest inter-system difference measured 8.25% between grapheme-based and phoneme-based ASR at 5 hours of training. As then expected, the highest total gain in performance of 5.15% is also measured between grapheme-based and phoneme-based ASR. As training hours increase, g2p-based ASR consistently performs approximately 1.93% worse than phoneme-based ASR. This indicates that even with an increase in training size g2p-based ASR is unlikely to outperform phoneme-based ASR. The lowest inter-system difference measured a very promising 1.31% between g2p-based ASR and grapheme-based ASR.

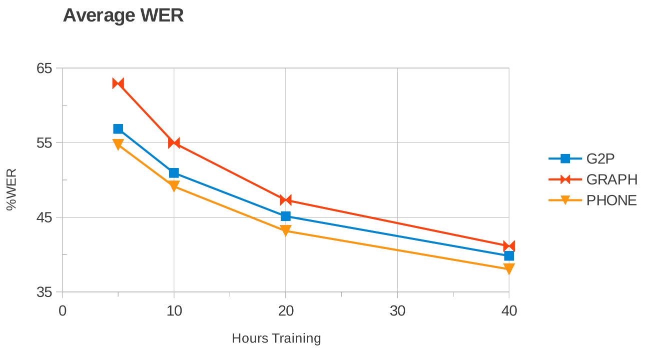

Fig. 1. Average WER of *grapheme-based*, *g2p-based* and *phoneme-based* ASR for training sizes of 5, 10, 20 and 40 hours across 4 folds

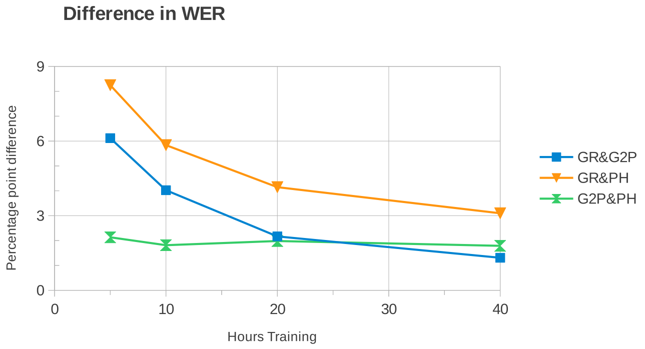

Fig. 2. Average difference in relative percentage of WER between *grapheme-based* and *g2p-based* ASR, *grapheme-based* and *phoneme-based* ASR, and *g2p-based* and *phoneme-based* ASR for training sizes of 5, 10, 20 and 40 hours across 4 folds

E. Error analysis

With the difference in WER being the most pronounced at 5 hours, we analyse the errors made according to word category. As mentioned in section II-C, the abbreviation category contains only one word namely *mej*, and since it doesn't occur in every fold's test set the abbreviation category is ignored during error analysis, leaving a total of 11 categories. Also, it has to be pointed out that words in the spelling error category can only be correctly recognised in their erroneous form. The data set has a fairly low saturation of spelling errors but their effect on recognition accuracy requires further investigation. Ideally (if data containing spelling errors are not to be discarded), spelling errors should either be corrected prior to system development, or the correct and incorrect spellings should be considered the same word during scoring. Both these approaches require that the word actually produced by the speaker should be identified. As this information was not available for the current analysis, spelling errors were handled as if they were standard words.

Table VII gives a detailed view of our findings. Scores are given as a percentage of how many times words from a specific category are miss-recognised as other words out of the total number of words from that category in all 4 test sets. Each cell is coloured green, yellow or red to indicate whether the relevant system performed best, second-best or worst. Not surprisingly grapheme-based ASR performed worse than phoneme-based ASR in 10 of the 11 categories It did however outperform g2p-based ASR in 5 categories namely spelled out words, proper names, spelling errors, partial words and multi-category words. The high WER of spelled out characters can be attributed to the language model used: with a flat language model the insertion penalty (the cost of adding an extra word during decoding) must be very high in order to produce sensible results. This causes short words to be miss-recognised very frequently.

Category	g-based WER	g2p WER	gold-dict WER
Spelled out char	73.73%	68.31%	63.65%
Multi-category	38.53%	40.54%	29.36%
Acronyms	32.03%	28.91%	26.95%
Unknown words	28.65%	25.15%	28.65%
Spelled out word	27.96%	30.53%	15.27%
Foreign	16.04%	14.92%	13.84%
Proper names	10.44%	11.00%	9.48%
Spelling errors	10.40%	11.42%	9.68%
Concatenation	7.48%	5.79%	5.67%
Partial words	6.62%	7.31%	6.13%
Generic Afr words	2.81%	2.49%	2.68%

TABLE VII
Word categories of errors observed at 5 hours of training data

Similarly, with the difference in WER being least at 40 hours, we again split errors based on word categories. Our findings are presented in Table VIII. Comparative to the error analysis of the smallest data set size (5 hours), grapheme-based ASR now outperforms g2p-based ASR in 4 out of the 11 categories, tying for an additional 2 categories. With increased training data, grapheme-based ASR managed to outperform phoneme-based ASR in 5 of the 11 categories. Interestingly, one of the categories includes generic Afrikaans words: the largest category of words in the test set. This might be attributed to noise-free pronunciation models or increased language regularity but this also requires further investigation. The biggest disparity in performance occurs in the spelled out words category between g2p-based and phoneme-based ASR, with g2p-based ASR miss-recognising twice as many words as phoneme-based ASR.

V. Conclusion

In this paper, the recognition accuracy of phoneme-based ASR and grapheme-based ASR was compared, using Afrikaans ASR as a case

Category	g-based WER	g2p WER	gold-dict WER
Spelled out char	62.65%	66.90%	63.89%
Multi-category	37.57%	35.87%	27.52%
Acronyms	31.50%	20.47%	25.98%
Unknown words	25.07%	25.07%	25.66%
Spelled out word	23.24%	28.47%	10.89%
Foreign	13.61%	12.81%	10.00%
Proper names	10.26%	11.83%	9.65%
Spelling errors	10.37%	11.38%	9.22%
Concatenation	5.24%	5.12%	6.33%
Partial words	6.20%	6.20%	8.27%
Generic Afr words	1.85%	1.76%	2.15%

TABLE VIII

Word categories of errors observed at 40 hours of training data

study. It was shown that at a context-level of three (using triphones or trigrams), a minimal effort grapheme-based ASR performs nearly on par with g2p-based ASR and converges quickly to the performance of manually verified phoneme-based ASR as the training set size increases.

Grapheme-based systems do not reach the same level of performance as that of a system developed using a manually verified dictionary, but this degradation in word accuracy is primarily caused by very specific word types, namely: spelled out words, acronyms, proper names and foreign words. All these categories (except for acronyms) tend to have highly irregular relationships between graphemes and phonemes confusing both the g2p-based and grapheme-based systems.

Spelled out words, acronyms and foreign words are typically easy to identify: spelled out words and acronyms tend to be short (and generic short words – which are not acronyms or spelled out words – tend to be known), and foreign words can mostly be identified using known word lists in relevant languages. Proper names tend to be more difficult to identify from text (unless capital letters are accurately retained during pre-processing). Luckily, once identified, these categories tend to be small in comparison with the total number of words to be modelled.

In future work, we will investigate an approach whereby the problematic categories are identified automatically and ‘ideal pronunciations’ are created for these. We propose that these ideal pronunciations then be converted to grapheme strings (by training phoneme-to-grapheme rules) in order for the pronunciations to be incorporated in a grapheme-based system. Given sufficient data, it may even be possible to train grapheme-to-grapheme rules: transliterating the original orthography of idiosyncratic words to an ’idealised’ orthography, more amenable to incorporation in a grapheme-based system. This could possibly combine the best of both worlds: the ability of a dictionary to capture idiosyncratic pronunciations, the minimal effort associated with the development of a grapheme-based system, and the ability of a grapheme-based system to remain ‘noise-free’, modelling almost all pronunciation variation at the acoustic level. However, in such a process, care should be taken that the additional variability improves the system, and does not introduce the same dictionary inconsistencies found in phoneme-based systems.

References

[1] M. Killer, S. Stuker, and T. Schultz, “Grapheme based speech recognition,” in *Proc. Eurospeech*, 2003, pp. 3141–3144.

[2] S. Kanthak and H. Ney, “Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition,” in *Proc. ICASSP*, 2002, pp. 845–848.

[3] E. G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck, “Automatic speech recognition without phonemes,” in *Proc. Eurospeech*, 1993, pp. 129–132.

[4] J. Dines and M. M. Doss, “A study of phoneme and grapheme based context-dependent asr systems,” in *Proceedings of the 4th International Conference on Machine Learning for Multimodal Interaction (MLMI)*, 2008, pp. 215–226.

[5] M. Davel and E. Barnard, “Pronunciation prediction with Default&Refine,” *Computer Speech and Language*, vol. 22, pp. 374–393, 2008.

[6] M. Davel and F. de Wet, “Verifying pronunciation dictionaries using conflict analysis,” in *Proc. Interspeech*, Tokyo, Japan, 2010, pp. 1898–1901.

[7] 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 *Proc. Interspeech*, August 2011, pp. 3176–3179.

[8] J. Badenhorst, A. de Waal, and F. de Wet, “Quality measurements for mobile data collection in the developing world,” in *Proc. Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU)*, Cape Town, South Africa, 2012, pp. 139–145.

[9] M. Creutz and K. Lagus, “Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor,” *Publications in Computer and Information Science, Report A*, vol. 81, 2005.

[10] G. B. van Huyssteen and M. M. van Zaanen, “Learning compound boundaries for Afrikaans spelling checking,” in *Pre-Proc. Workshop on International Proofing Tools and Language Technologies*, July 2004, pp. 101–108.

[11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, *The HTK Book*. http://htk.eng.cam.ac.uk/: Cambridge University Engineering Department, 2005.