MORPHEME-BASED AUTOMATIC SPEECH RECOGNITION FOR A MORPHOLOGICALLY RICH LANGUAGE – AMHARIC

Martha Yifiru Tachbelie, Solomon Teferra Abate, Wolfgang Menzel

Department of Informatics, University of Hamburg
Vogt-Kölln Str. 30, D-22527 Hamburg, Germany
abate,tachbeli,menzel@informatik.uni-hamburg.de

ABSTRACT

Out-of-vocabulary (OOV) words are a major source of error in a speech recognition system and various methods have been proposed to increase the performance of the systems by properly dealing with them. This paper presents an automatic speech recognition experiment conducted to see the effect of OOV words on the performance speech recognition system for Amharic (a morphologically rich language). We tried to solve the OOV problem by using morphemes as dictionary and language model units. It has been found that for a small vocabulary (5k) system morphemes are better lexical and language modeling units than words. An absolute improvement (in word recognition accuracy) of 11.57% has been obtained as a result of using a morph-based vocabulary. However, for large vocabularies morpheme-based systems did not bring much performance improvement as they suffer from acoustic confusability and limited language model scope while word-based recognizers benefit much from OOV rate reduction.

Index Terms— Out-of-Vocabulary problem, Morpheme-based speech recognition, Amharic

1. INTRODUCTION

Most large vocabulary speech recognition systems operate with a finite vocabulary. All the words which are not in the system's vocabulary are considered out-of-vocabulary words. These words are one of the major sources of error in an automatic speech recognition system. When a speech recognition system is confronted with a word which is not in its vocabulary, it may recognize it as a phonetically similar in-vocabulary unit/item. That means the OOV word is mis-recognized. This in turn might cause its neighboring words also to be mis-recognized. [1] indicated the fact that each OOV word in the test data contribute to 1.6 errors on the average. Therefore, different approaches have been investigated to cope with the OOV problem and consequently to reduce the error rate of automatic speech recognition systems. One of these approaches is vocabulary optimization [2], where the vocabulary is selected in a way that it reduces the OOV rate. This involves either increasing the vocabulary size or including frequent words in a vocabulary. This approach may work for morphologically simple languages like English where a 20k vocabulary has 2% OOV rate and a 65k one has only 0.6% [3].

However, for morphologically rich languages, for which OOV is a severe problem, a much larger vocabulary is required to reach the 1% OOV rate. [3] indicated the fact that for Russian and Arabic 800k and 400k vocabularies are required, respectively for a 1% OOV rate. Increasing the vocabulary to alleviate the OOV problem is not the best solution especially for morphologically rich languages as the system complexity increases with the size of the vocabulary. Therefore, modeling sub-word units, particularly morphs, has been used for morphologically rich languages. Many researchers [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14] did morpheme-based or sub-word based speech recognition experiments.

In this paper, we show the effect of OOV rate on the performance of an Amharic speech recognition system. We investigate options to reduce the OOV problem using morphemes as a lexical and language modeling unit and study its effect on the performance of the system. Section 2 gives a brief description of the Amharic word morphology. After reviewing previous works on morpheme-based speech recognition for Amharic in Section 3, we present the results of our experiments in Sections 4, 5 and 6. Finally, conclusions are drawn and recommendations for future works are derived in Section 7.

2. AMHARIC MORPHOLOGY

Amharic is a member of the Ethio-Semitic languages, which belong to the Semitic branch of the Afro-Asiatic super family [15]. It is related to Hebrew, Arabic, and Syrian. Amharic is a major language spoken mainly in Ethiopia. According to the 1998 census, it is spoken by over 17 million people as a first language and by over 5 million as second language throughout different regions of Ethiopia [16].

Like other Semitic languages such as Arabic, Amharic exhibits a root-pattern morphological phenomenon. A root is a set of consonants (called radicals) which has a basic 'lexical' meaning. A pattern consists of a set of vowels which are inserted (intercalated) among the consonants of the root to form a stem. The pattern can be combined with a particular prefix or suffix to make a single grammatical form [17] or another stem [18]. For example, the Amharic root *sbr*[1] means 'break', when we intercalate the pattern ä-ä and attach the suffix -ä we get *säbbärä* 'he broke' which is the first form of a verb (3rd person masculine singular in past tense as in other semitic languages) [17]. In addition to this non-concatenative morphological feature, Amharic uses different affixes to form inflectional and derivational word forms.

Some adverbs can be derived from adjectives but, adverbs are not inflected. Nouns are derived from other basic nouns, adjectives, stems, roots, and the infinitive form of a verb by affixation and intercalation. For example, from the noun *lǧ* 'child' another noun *lǧInät* 'childhood'; from the adjective *däg* 'generous' the noun *dägnät* 'generosity'; from the stem *snIf*, the noun *snIfna* 'laziness'; from root *qld*, the noun *qäld* 'joke'; from infinitive verb *mäsbär* 'to break' the noun *mäsbäriya* 'an instrument used for breaking' can be derived.

Case, number, definiteness, and gender marking affixes inflect nouns. Table 1 presents, as an example, the genitive case markers for nouns.

Person	singular		plural
	Vowel ending	Consonant ending	
1^{st}	-ye	-e	-aččn
2^{nd} masculine	-h	-Ih	-aččhu
2^{nd} feminine	-š	-Iš	
2^{nd} polite	-wo	-wo	
3^{rd} masculine	-w	-u	-aččäw
3^{rd} feminine	-wa	-wa	
3^{rd} polite	-aččäw	-aččäw	

Table 1. Genetive Case Markers (Adapted from Titov (1976))

Adjectives are derived from nouns, stems or verbal roots by adding a prefix or a suffix. For example, it is possible to derive *dnIgayama* 'rocky' from the noun *dnIgay* 'rock, stone'; *znIgu* 'forgetful' from the stem *znIg*; *sänäf* 'lazy' from the root *snf* by suffixation and intercalation. Adjectives can also be formed through compounding. For instance, *hodäsäfi* 'tolerant, patient', is derived by compounding the noun *hod* 'stomach' and the adjective *säfi* 'wide'. Like nouns, adjectives are inflected for gender, number, and case [18].

Unlike the other word categories such as noun and adjectives, the derivation of verbs from other parts of speech is not common. The conversion of a root to a basic verb stem requires both intercalation and affixation. For instance, from the root *gdl* 'kill' we obtain the perfective verb stem *gäddäl-* by intercalating pattern ä-ä. From this perfective stem, it is possible to derive the passive stem *tägäddäl-* and the causative stem *asgäddäl-* using prefixes tä- and as-, respectively. Other verb forms are also derived from roots in a similar fashion.

Verbs are inflected for person, gender, number, aspect, tense and mood [18]. Table 2 shows how a perfective Amharic verb inflects for person, gender and number. Other elements like negative markers also inflect verbs in Amharic.

Person	Singular	Plural
1^{st}	säbbärku/hu	säbbärn
2^{nd} masculine	säbbärh/k	säbbäraččhu
2^{nd} feminine	säbbärš	
2^{nd} polite	säbbäru	
3^{rd} masculine	säbbärä	säbbäru
3^{rd} feminine	säbbäräčč	
3^{rd} polite	säbbäru	

Table 2. Inflection of a Perfective Verb

From the above brief description of Amharic word morphology it can be seen that Amharic is a morphologically rich language. It is this feature that makes the OOV problem more serious in Automatic speech recognition system.

3. PREVIOUS WORK

The application of automatic word decomposition (using Harris algorithm) for automatic speech recognition of less-represented languages, specifically Amharic, has been investigated by [12]. In their study, the units obtained through decomposition have been used in both lexical and language models. They reported recognition results for four different configurations: full word and three decomposed forms (detaching both prefix and suffix, prefix only and suffix only). A word error rate (WER) reduction over the base line word-based system has been reported using 2 hours of training data in speech recognition in all decomposed forms although the level of improvement varies. The highest improvement (5.2% absolute WER reduction) has been obtained with the system in which only the prefixes have been detached. When both the prefixes and suffixes have been considered, the improvement in performance is small, namely 2.2%. This might be, as the authors indicate, due to the limited span of the n-gram language models.

Decomposing lexical units with the same algorithm led to worse performance when more training data (35 hours) was used [13]. This can be explained by a higher acoustic confusability. [13] tried to solve this problem by using other modified decomposition algorithms. Their starting algorithm is Morfessor [19] which has been modified by adding different information. They were able to achieve a word error rate reduction only when a phonetic confusion constraint was used

[1] For transcription purposes, IPA representation is used with some modifications.

to block the decomposition of words which would result in acoustically confusable units.

In contrast to [12] and [13], [14] used morphemes only for the language modeling component. They applied a lattice rescoring framework to avoid the influence of acoustic confusability on the performance of the speech recognizer. Lattices have been generated in a single pass recognition using a bigram word-based language model and rescored using sub-word language models. Improvement in the performance of the speech recognition has been obtained. However, this method does not solve the out-of-vocabulary problem since a word-based pronunciation dictionary has been used.

4. WORD-BASED SPEECH RECOGNITION

4.1. The Speech Corpus

The speech corpus used to develop the speech recognition system is an Amharic read speech corpus [20]. It contains 20 hours of training speech collected from 100 speakers who read a total of 10850 sentences (28666 tokens). Compared to other speech corpora that contain hundreds of hours of speech data for training, our models obviously suffer from a lack of training data.

Although the corpus includes four different test sets (5k and 20k both for development and evaluation), for the purpose of the current investigation we have used the 5k development test set, which includes 360 sentences (4106 tokens or 2836 distinct words) read by 20 speakers.

4.2. Acoustic, Lexical and Language Models

The acoustic model consists of 6610 cross-word triphone HMMs each with 3 emitting states. The states of these models and all the cross-word triphone models that are potentially needed for recognition are tied using decision-tree based state-clustering that reduced the number of triphone models from 77658 logical models to 10215 physical ones. Their mixture is added incrementally and 12 Gaussian mixtures have been found to be the optimal.

Vocabulary of the three full-word form pronunciation dictionaries (5k, 20k and 65k) have been prepared by taking the most frequent words from a text corpus consisting of 120262 sentences (2348150 tokens or 211120 types). Table 3 shows the out-of-vocabulary rates of the 5k development test set against these vocabularies. Although we tried to optimize the vocabularies by taking the most frequent words, the OOV rate is still high.

Vocabulary	Token OOV (%)	Type OOV (%)
5k	36.43	51.55
20k	20.41	29.23
65k	9.33	13.36

Table 3. OOV rate of the 5k development test set

In order to minimize the development effort, the pronunciation dictionaries have been encoded by means of a simple procedure that takes advantage of the orthographic representation (a consonant vowel syllable) which is fairly close to the pronunciation in many cases. There are, however, notable differences especially in the area of gemination and insertion of the epenthetic vowel.

The text corpus from which the vocabularies have been selected has also been used to train language models. As we have three dictionaries (5k, 20k and 65k), we have developed three trigram language models one for each vocabulary using the SRILM toolkit [21]. The language models are made open by including a special unknown word token. The modified Kneser-Ney smoothing method has been used to smooth all the language models.

4.3. Performance of Word-based Speech Recognizers

Speech recognition experiment has been performed using the 5k, 20k and the 65k vocabularies. In each case the systems have been evaluated with the 5k development test set. Figure 1 presents the word recognition accuracy for each vocabulary. As it can be seen from the figure, the OOV rate decreases when the vocabulary size increases. As the OOV rate decreases the performance of the speech recognition system increases. The best performance (78.3%) has been obtained for the 65k which has OOV rate of 9.33%. The results show that the OOV rate highly affects the performance of speech recognition systems. To deal with this problem, morphemes instead of words have been considered as dictionary entries and units in language models.

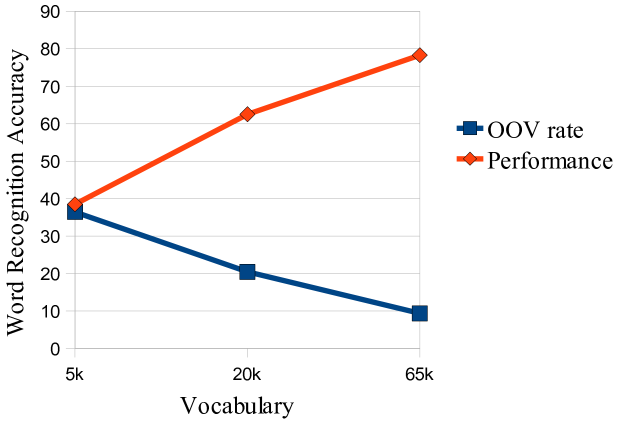

Fig. 1. Word Recognition Accuracy of three Word-based Recognizers.

5. MORPHEME-BASED SPEECH RECOGNITION

5.1. Morphological Segmentation

To use morphemes in speech recognition system a word parser, which splits word forms into their constituents, is needed . Different attempts [22, 23, 24] have been made to develop a morphological analyzer for Amharic using different methods. However, none of the systems can be directly used for our purpose. The systems developed by [22] and [23] suffer from lack of data. The morphological analyzer developed by [24] seems to suffer from a too small lexicon. It has been tested on 207 words and analyzed less than 50% (75 words) of them. Moreover, the output of the system is not directly useful for our study which needs the morphemes themselves instead of their morphological features. Since the source code of the analyzer has not yet been made available, it is not possible to customize it.

An alternative approach is offered by unsupervised corpus-based methods that do not need annotated data. These methods are particularly interesting for resource scarce languages like Amharic. Thus, Morfessor [19] which is a freely available, language independent unsupervised morphology learning tool that tries to identify all the morphemes found in a given word has been used to morphologically segment our text corpus. The morphologically segmented text consists of 15,925 distinct morphs. Figure 2 shows the morph length (in terms of number of characters) distribution of the corpus. As can be observed from the figure, the length of most of the morphs is between four and six characters. In order to facilitate the conversion of morpheme sequences to words, a special word boundary marker has been attached to word boundary morphemes which made the morphemes context-sensitive and consequently increased the number of distinct morphemes to 28,492.

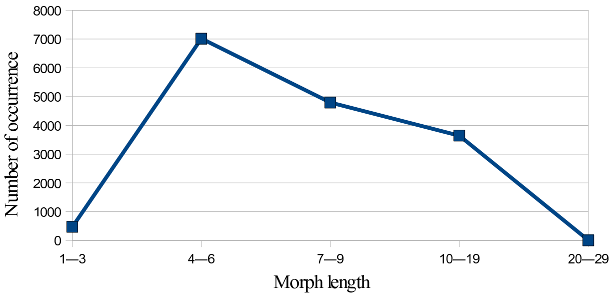

Fig. 2. Morph Length Distribution.

5.2. Acoustic, Lexical and Language Models

The acoustic model has been developed in a similar fashion as for the word-based recognizers. The training data has a set of 6459 cross-morph triphone HMMs each with 3 emitting states. The states of these models and all the possible cross-morph triphone models are tied and, therefore, the number of triphone models is reduced from 57799 logical to 7685 physical models. Similar to the word-based models, 12 Gaussian mixtures have been found to be the optimal.

The entries in the pronunciation dictionary are morphemes. From the morphologically segmented corpus, three dictionaries have been prepared: 5k and 20k by taking the most frequent morphs and 28.4k by considering all the morphemes. The morph-based OOV rates of these vocabularies on the 5k development test set are presented in Table 4 which shows that the OOV rate is highly reduced as a result of using morphs. The token OOV rate for the 5k morph vocabulary, for instance, is only a little bit higher than the token OOV rate of the 65k full-word vocabulary.

Vocabulary	Token OOV (%)	Type OOV (%)
5k	10.75	28.43
20k	0.67	1.83
All (28.4k)	0.03	0.08

Table 4. Morph OOV rate of the 5k development test set

As we have three dictionaries (5k, 20k and 28.4k), we have developed three open vocabulary morph-based trigram language models, one for each vocabulary. Similar to the word-based language models, the morpheme-based ones have also been smoothed using modified Kneser-Ney smoothing technique.

5.3. Performance of Morph-based Speech Recognizers

The morpheme-based speech recognition system has been evaluated on the 5k development test set using the 5k, 20k and 28.4k morph vocabularies. The results are reported in terms of morph recognition accuracy (MRA) and word recognition accuracy (WRA). The word recognition accuracy has been computed after words have been obtained by concatenating the recognized morph sequence. The best performance (see Table 5) has been obtained with the 28.4k morph vocabulary which has an OOV rate of 0.03. Since the OOV rate is very small, an accuracy even higher than the one reported here was expected. The reasons for this disappointing performance (in spite of having a small OOV rate) might be a higher acoustic confusability and the limited language model scope.

Vocabulary	MRA (%)	WRA (%)
5k	55.34	50.04
20k	67.67	62.00
28.4k	68.26	62.78

Table 5. Performance of morph-based speech recognizer

6. COMPARISON OF WORD- AND MORPH-BASED SPEECH RECOGNIZERS

The morph vocabularies have a very low OOV rate compared to the word vocabularies. This has a positive effect on speech recognition accuracy, especially for small vocabularies, namely 5k. The word-based model has a word recognition accuracy of 38.47% when the 5k vocabulary has been used. On the other hand, the morpheme-based system reaches a word recognition accuracy of 50.04% for the 5k morph vocabulary[2], which means an absolute improvement of 11.57%. However, for the 20k the morpheme-based speech recognizer performed slightly worse (62.00%) than the equivalent word-based system which has a word recognition accuracy of 62.51%. The 28.4k vocabulary has morph and word recognition accuracies of 68.26% and 62.78%, respectively. The performance of the recognizer with 28.4k morph vocabulary is only slightly better than the 20k word-based recognizer although it includes all the morphs in the text and has a very low OOV rate. As we have already mentioned, besides the acoustic confusability, the limited scope of the morpheme-based n-gram language model might contribute to the poor performance of the morpheme-based speech recognizer since taking three morphemes might not mean taking three words. This has also been commented by [12] who suggested the use of higher order n-gram models. Thus, higher order morpheme-based language models have been used in our morpheme-based speech recognizers.

We generated lattices using the 20k and 28.4k vocabulary morpheme-based recognizers and rescored the lattices with a quadrogram morpheme-based language model which has been developed in the same manner as the trigram models. The best path transcription decoded from the rescored lattices have morph and word recognition accuracy of 69.70% and 64.46%, respectively for the 28.4k vocabulary and 68.92% and 63.51% for the 20k one (see Table 6). Absolute 1.95% and 1% word recognition accuracy improvement (over the 20k word-based recognizer) have been obtained for the 28.4k and 20k vocabulary morpheme-based recognizers, respectively, as a result of rescoring the lattices with a quadrogram language model.

Vocabulary	MRA (%)	WRA (%)
20k	68.92	63.51
28.4k	69.70	64.46

Table 6. Lattice rescoring with quadrogram morpheme-based language model

As it can be seen from Table 7, rescoring with a pentagram language model did not lead to further improvement. Rather, the morph and word recognition accuracies (for both 20k and 28.4k vocabularies) became worse than the recognizer that used the quadrogram morph-based language model. This might be due to data sparseness. As the language model training corpus is very small many of the pentagrams might not be encountered in the training data and therefore estimated in terms of lower order n-grams. Regarding the language models quality, the pentagram language models did not bring much perplexity improvement (less than 1%) over the quadrogram ones for the 20k and the 28.4k vocabularies. The perplexity gains of the quadrogram language models over the trigram ones are 8.291% and 8.386% for the 20k and 28.4k vocabularies, respectively.

Vocabulary	MRA (%)	WRA (%)
20k	67.69	62.17
28.4k	68.48	63.17

Table 7. Lattice rescoring with pentagram morpheme-based language model

7. CONCLUSIONS AND FURTHER WORK

Speech recognition experiments for Amharic have been conducted to study the effect of OOV words problem for a highly inflectional language and to find out whether the problem can be reduced by using morphemes as lexical and language model units. We did both word-based and morph-based speech recognition experiments. For the word-based systems the OOV rate decreases as the vocabulary size increases and word recognition accuracy increases as the OOV rate decreases. It has also been found that using morphemes as dictionary entries and language model units highly reduces the OOV rate and consequently boosts the word recognition accuracy, especially for small vocabularies (5k). However, as the morph vocabulary grows, the morpheme-based recognizers did not bring notable improvement in word recognition accuracy, which might be due to higher acoustic confusability and a limited language model scope. Rescoring lattices with higher order morpheme-based language model (quadrogram) brought word recognition accuracy improvement.

Although the morpheme-based recognizer benefits from the low OOV rate, it is disadvantaged from the small and acoustically confusable units. Therefore, further improvement can be obtained if care is taken (for instance, using confusion constraints as in [13]) to avoid acoustically confusable units. Moreover, we just concatenated recognized morpheme sequences up to a word boundary marker and no effort has been made to avoid concatenation of illegal morpheme sequences. Attempts in this line may also boost the performance of morpheme-based speech recognizer. For example, rules (such as *ignore the subject marker morph if it comes at the beginning of a morph sequence*) that obstruct the concatenation of illegal morph sequences can be used.

[2] Comparing the morph-based systems directly with the word-based ones may not be fair because they have a higher coverage than word-based systems of the same vocabulary size. On the other hand, the morph-based systems are also dis-favoured by the concatenation of illegal morph-sequences, increasing number of small and acoustically confusable units and a limited language model scope.

8. REFERENCES

[1] P. C. Woodland, C. J. Leggetter, J. J. Odell V. Valtchev, and S. J. Young, "The 1994 HTK large vocabulary speech recognition system," in *Proceedings of the 1995 International Conference on Acoustics, Speech and Signal Processing*, 1995, vol. 1, pp. 73–76.

[2] I. Bazzi, *Modelling Out-of-Vocabulary Words for Robust Speech Recognition*, Ph.D. thesis, Massachsetts Institute of Technology, 2002.

[3] M. Gales and P. Woodland, "Recent progress in large vocabulary continuous speech recognition: An htk perspective," 2006.

[4] P. Geutner, "Using morphology towards better large-vocabulary speech recognition systems," in *Proceedings of IEEE International on Acoustics, Speech and Signal Processing*, 1995, vol. I, pp. 445–448.

[5] K. Carki, P. Geutner, and T. Schultz, "Turkish lvcsr: towards better speech recognition for agglutinative languages," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1563–1566, 2000.

[6] W. Byrne, J. Hajič, P. Ircing, F. Jelinek, S. Khudanpur, P. Krebc, and J. Psutka, "On large vocabulary continuous speech recognition of highly inflectional language - czech," in *Proceeding of the European Conference on Speech Communication and Technology*, 2001, pp. 487–489.

[7] E. Whittaker and P. Woodland, "Particle-based language modeling," in *Proceeding of International Conference on Spoken Language Processing*, 2000, pp. 170–173.

[8] E. W. D. Whittaker, J. M. Van Thong, and P. J. Moreno, "Vocabulary independent speech recognition using particles," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001, pp. 315–318.

[9] V. Siivola, T. Hirsimki, M. Creutz, and M. Kurimo, "Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner," in *Proceedings of Eurospeech*, 2003, pp. 2293–2296.

[10] T. Hirsimäki, M. Creutz, V. Siivola, and M. Kurimo, "Morphologically motivated language models in speech recognition," in *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, 2005, pp. 121–126.

[11] K. Kirchhoff, J. Bilmes, J. Henderson, R. Schwartz, M. Noamany, P. Schone, G. Ji, S. Das, M. Egan, F. He, D. Vergyri, D. Liu, and N. Duta, "Novel speech recognition models for arabic," Tech. Rep., Johns-Hopkins University Summer Research Workshop, 2002.

[12] T. Pellegrini and L. Lamel, "Investigating automatic decomposition for asr in less represented languages," in *Proceedings of INTERSPEECH 2006*, 2006.

[13] T. Pellegrini and L. Lamel, "Using phonetic features in unsupervised word decompounding for asr with application to a less-represented language," in *Proceedings of INTERSPEECH 2007*, 2007, pp. 1797–1800.

[14] M. Y. Tachbelie, S. T. Abate, and W. Menzel, "Morpheme-based language modeling for amharic speech recognition," in *Proceedings of the 4th Language and Technology Conference - LTC-09*, 2009, pp. 114–118.

[15] R. M. Voigt, "The classification of central semitic," *Journal of Semitic Studies*, , no. 32, pp. 1–21, 1987.

[16] Anbessa Teferra and Grover Hudson, *Essentials of Amharic*, Köppe, Köln, 2007.

[17] M.L. Bender, J.D. Bowen, R.L. Cooper, and C.A. Ferguson, *Languages in Ethiopia*, Oxford Univ. Press, London, 1976.

[18] B. Yimam, *yäamarIña säwasäw*, EMPDE, Addis Ababa, 2nd. ed. edition, 2000EC.

[19] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.1," Tech. Rep. A81, Neural Networks Research Center, Helsinki University of Technology, 2005.

[20] S. T. Abate, W. Menzel, and B. Tafila, "An amharic speech corpus for large vocabulary continuous speech recognition," in *Proceeding of International Conference on Speech Communication and Technology, Interspeech-2005*, 2005.

[21] A. Stolcke, "SRILM — an extensible language modeling toolkit," in *Proceedings of International Conference on Spoken Language Processing*, 2002, vol. II, pp. 901–904.

[22] A. Bayou, "Developing automatic word parser for amharic verbs and their derivation," M.S. thesis, Addis Ababa University, 2000.

[23] T. Bayu, "Automatic morphological analyzer for amharic: An experiment employing unsupervised learning and autosegmental analysis approaches," M.S. thesis, Addis Ababa University, 2002.

[24] S. Amsalu and D. Gibbon, "Finite state morphology of amharic," in *Proceedings of International Conference on Recent Advances in Natural Language Processing*, 2005, pp. 47–51.