Natural Language Engineering **20** (2): 235–259. © Cambridge University Press 2012
doi:10.1017/S1351324912000356

Using morphemes in language modeling and automatic speech recognition of Amharic

MARTHA YIFIRU TACHBELIE[1],
SOLOMON TEFERRA ABATE[1] and
WOLFGANG MENZEL[2]

[1]*School of Information Sciences, Addis Ababa University, Addis Ababa, Ethiopia*
e-mail: {marthayifiru, solomon_teferra_7}@yahoo.com
[2]*Department of Informatics, University of Hamburg, Hamburg, Germany*
e-mail: menzel@informatik.uni-hamburg.de

(*Received 26 September 2011; revised 8 November 2012; accepted 12 November 2012; first published online 12 December 2012*)

Abstract

This paper presents morpheme-based language models developed for Amharic (a morphologically rich Semitic language) and their application to a speech recognition task. A substantial reduction in the out of vocabulary rate has been observed as a result of using subwords or morphemes. Thus a severe problem of morphologically rich languages has been addressed. Moreover, lower perplexity values have been obtained with morpheme-based language models than with word-based models. However, when comparing the quality based on the probability assigned to the test sets, word-based models seem to fare better. We have studied the utility of morpheme-based language models in speech recognition systems and found that the performance of a relatively small vocabulary (5k) speech recognition system improved significantly as a result of using morphemes as language modeling and dictionary units. However, as the size of the vocabulary increases (20k or more) the morpheme-based systems suffer from acoustic confusability and did not achieve a significant improvement over a word-based system with an equivalent vocabulary size even with the use of higher order (quadrogram) n-gram language models.

1 Introduction

Language models are fundamental to many natural language applications such as automatic speech recognition (ASR) and statistical machine translation (SMT). The most widely used kind of language models are statistical ones. They provide an estimate of the probability of a word sequence for a given task. The probability distribution depends on the available training data and how much of the context is considered (Junqua and Haton 1996). Therefore, large amounts of training data are required to ensure statistical significance (Young *et al.* 2006). Even if a large training corpus is available, there may still be many possible word sequences which will not appear at all, or which appear with a statistically insignificant frequency (data sparseness problem) (Young *et al.* 2006). There are even individual words that might

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

not be encountered in the training data irrespective of its size [out-of-vocabulary (OOV) words problem].

These problems are more serious for morphologically rich languages such as Amharic, Arabic, Hebrew, and Turkish than for languages like English. Languages with rich morphology have a high-vocabulary growth rate which results in a high perplexity and a large number of OOV words (Vergyri *et al.* 2004). A straightforward solution for the OOV problem would be an extension of the size of the training vocabulary so as to minimize the set of OOV words. However, this is not effective even for morphologically poor languages, let alone for morphologically rich ones as a vocabulary of any size becomes inadequate. Sproat (1992) states that '… *one can have a very large list of words yet still encounter many words that are not to be found in that list*'. He emphasized the problem of OOV words even for English, which is known to be a morphologically simple language compared to many other languages. One can imagine how severe the data sparseness and OOV words problems become in morphologically richer languages. Thus, a possibility for building high-quality language models is required which is applicable even in case where the training material is limited (Geutner 1995). A promising direction is abandoning the word as a modeling unit and splitting words into their parts (Hirsimäki *et al.* 2005).

Such subword units have been used to improve the coverage of language models and consequently the performance of the applications to which the language models are applied (Geutner 1995; Whittaker and Woodland 2000; Byrne *et al.* 2001; Kirchhoff *et al.* 2002; Hirsimäki *et al.* 2005; Creutz *et al.* 2007; El-Desoky *et al.* 2009; Pellegrini and Lamel 2009; Heintz 2010). For example, morphemes have been used to improve the performance of ASR systems for different morphologically rich languages: for Arabic 0.5% (Kirchhoff *et al.* 2002), for German 1% (Geutner 1995), for Turkish 5.5%, for Finnish 7.6%, for Estonian 14.8% absolute word error rate (WER) reduction (Creutz *et al.* 2007).

Since Amharic is a morphologically rich language, an Amharic language model suffers from the data sparseness and OOV words problems. The negative effect of the Amharic morphology on language modeling has already been reported by Abate (2006), who also recommended the development of subword-based language models. In this paper, we present results of experiments with morpheme-based Amharic language models and their application to Amharic speech recognition.

The next section gives a brief description of the Amharic morphology. Section 3 presents the corpus used for our experiment and the perplexity results for statistical morpheme-based language models. In Section 4, perplexities of word-based language models are presented and a comparison with statistical morpheme-based language models is carried out. Section 5 deals with linguistic morpheme-based language models and Section 6 gives the results of morpheme-based speech recognition experiments and compares them with word-based recognizers. Finally, conclusions and future directions for research are presented in Section 7.

2 Morphology of Amharic

Amharic is a member of the Ethio-Semitic languages, which belong to the Semitic branch of the Afroasiatic super family (Voigt 1987). It is related to Hebrew, Arabic,

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

and Syriac. Amharic, which is spoken mainly in Ethiopia, is the second most widely spoken Semitic language, after Arabic. According to the 1998 census, it is spoken by over 17 million people as a first language and by over 5 million as a second language throughout different regions of Ethiopia. Currently, Amharic is the official working language of the federal democratic republic of Ethiopia and several of the states within the federal system (Teferra and Hudson 2007). The language is also spoken in other countries such as Egypt and Israel.

Like other Semitic languages, Amharic exhibits a root-pattern morphological phenomenon. In addition, it uses different affixes to create inflectional and derivational word forms. In this section we give a short description of the Amharic word morphology.

2.1 Root-pattern morphology

Amharic is characterized by the root-pattern nonconcatenative morphology. A root, also called a radical, is a sequence of commonly three consonants which carry the basic lexical meaning. A pattern, also called a vocalic element, consists of vowels which are inserted (intercalated) between the consonants of the root. Stems are, therefore, formed by intercalating the vowels between the root consonants. Finally, prefixes and suffixes are added either to complete the stem as a word or to form another stem.

For example, the Amharic root *sbr*[1] means 'break'. If we insert the pattern ä-ä among the radicals, we obtain the stem *säbbär*-. Attaching the suffix -ä gives *säbbärä* 'he broke' which is the first form of the verb (third person masculine singular in past tense as in other Semitic languages) (Bender *et al.* 1976). Using the same pattern, but without geminating the second consonant, and attaching the suffix -a, the process nominal *säbära* is produced which means 'breaking'. The pattern ä-a combined with the suffix -i makes agent nouns, as in *säbari* 'one who breaks'. The same stem with the suffix -a, gives us the participle *säbara* 'broken'. Intercalating the vowel ä between the second and third consonant results in a jussive stem *sbär*-. Attaching the prefix mä- to it forms an infinitive verb *mäsbär* 'to break' (Bender *et al.* 1976). The imperfective stem is derived from the root by intercalating the vowel ä between the first and the second consonants.

2.2 Derivational and inflectional morphology

There are five parts of speech in Amharic: adjectives, nouns, verbs, adverbs, and prepositions and conjunctions. Prepositions and conjunctions are totally unproductive. Adverbs are few in number and are less productive. They are not inflected but some adverbs can be derived from adjectives, for instance, *kIfuña* 'severely or seriously' is derived from the adjective *kIfu* 'wicked' by suffixing *-ña*.

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

[1] For transcription purpose, International Phonetic Association (IPA) representation is used with some modifications.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

Table 1. *Genitive case markers*

Person	Singular		Plural
	Vowel ending	Consonant ending	
1st	-ye	-e	-aččIn
2nd			
Masculine	-h	-Ih	-aččIhu
Feminine	-š	-Iš	
Polite	-wo	-wo	
3rd			
Masculine	-w	-u	-aččäw
Feminine	-wa	-wa	
Polite	-aččäw	-aččäw	

noun *lIğğ* 'child' another noun *lIğnät* 'childhood'; from the adjective *däg* 'generous' the noun *dägInnät* 'generosity'; from the stem *sInf*, the noun *sInfInna* 'laziness'; from root *k'ld*, the noun *k'äld* 'joke'; from infinitive verb *mäsbär* 'to break' the noun *mäsbäriya* 'an instrument used for breaking' can be derived. Case, number, definiteness, and gender marker affixes inflect nouns. Table 1 presents, as an example, the genitive case markers for nouns.

Adjectives are derived from nouns and stems by adding a prefix or a suffix, and they can also be derived from verbal roots by intercalating patterns. For example, it is possible to derive *dIngayama* 'stony' from the noun *dIngay* 'stone' and *zIngu* 'forgetful' from the stem *zIng*, by suffixation; *sänäf* 'lazy' can be derived from the root *snf* by 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 (Yimam 2007).

Unlike the other word categories such as nouns 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 the pattern ä_ä. From this perfective stem, it is possible to derive a passive (*tägäddäl-*) and a 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 (Yimam 2007). Table 2 shows, as an example, the inflection of perfective and imperfective verbs for person, gender, and number. Other elements like negative markers and prepositions also inflect verbs in Amharic. A single Amharic word may, therefore, represent a complete English sentence. For instance, *amt'Iččewalähu* is equivalent to the sentence 'I have brought it'. A more detailed description of Amharic morphology can be found in (Tachbelie 2010).

From the above brief description, it can be seen that Amharic is a morphologically rich language. It is this feature that makes the development of language models for Amharic challenging.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

Table 2. *Inflection for person, gender, and number*

Person	Perfective	Imperfective
1st	säbbär-ku/hu	?-säbr
1st plural	säbbär-n	?-n-säbr
2nd		
Masculine	säbbär-h/k	t-säbr
Feminine	säbbär-š	t-säbr-i
Polite	säbbär-u	t-säbr-u
Plural	säbbär-aččIhu	t-säbr-u
3rd		
Masculine	säbbär-ä	y-säbr
Feminine	säbbär-äčč	t-säbr
Polite	säbbär-u	y-säbr-u
Plural	säbbär-u	y-säbr-u

3 Statistical morpheme-based language models

3.1 Corpus preparation

To train language models, a text corpus (called ATC_48k hereafter) consisting of 48,090 sentences and 1,542,697 tokens has been prepared. The electronic text has been obtained from the ethiozena[2] archive which contains written newscast. Since the target application domain is speech recognition and since we wanted to merge the text corpus with the one prepared by Abate, Menzel and Tafila (2005) for speech recognition, the text has been normalized accordingly. The normalization task includes:

- Expansion of abbreviations and contractions. The most common contractions are Ethiopian personal names, titles, and names formed through compounding. For example, the personal name gäbrämaryam is commonly written as gä/maryam, the title t'äk'lay ministär usually contracted as t'ä/ministär or sometimes as t'ä/mi.
- Textual transcription of numbers. Telephone numbers, dates, times, etc. written in numbers have been transcribed.
- Removal of punctuation marks.
- Removal of foreign words, particularly English words provided that their removal does not affect the grammatical structure of a sentence.

After normalization, the ATC_48k corpus was merged with the text database prepared by Abate *et al.* (2005) from the same domain. The combined text corpus (hereafter called ATC_120k) was used in the experiments and consists of 120,262 sentences with 2,348,150 tokens and 211,120 word types. Table 3 presents the frequency distribution of words in the ATC_120k corpus.

It closely follows Zipf's law with a few very common words, a middle number of medium frequency words, and very many low frequency words (Manning and

[2] http://www.ethiozena.net/

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

Table 3. *Word frequency distribution*

Frequency	Number of words
1	121,285
2–10	69,526
11–100	17,356
101–1,000	2,655
1,001–10,000	293
10,001–20,000	3
above 20,000	2

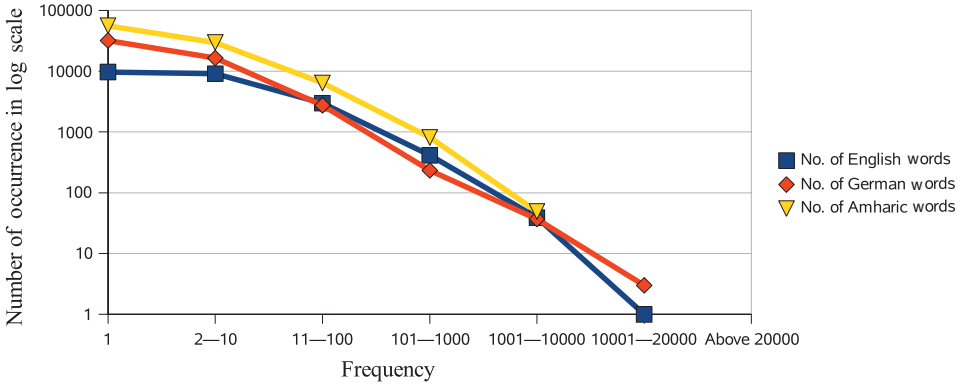


Fig. 1. (Colour online) Word frequency distribution of Amharic, English, and German.

Schütze 1999), which implies that the data is sparse. As can be seen from Table 3, more than 50% (121,285) of the words occur only once (hapax legomena) in the corpus. On the other hand, only five words appear with a frequency of above 10,000. These words are function words such as wIsT 'in'.

We compared the frequency distribution of Amharic words with other languages, namely English and German. To make a fair comparison, we normalized the size of the available corpora to the smallest one, Negra which consists of 20,602 German sentences. The same number of sentences has been selected for English and Amharic from the Wall Street Journal (WSJ) and the ATC_120k corpora, respectively. Figure 1 depicts the frequency distribution of words for the three languages. As can be clearly seen, the number of hapax legomena for Amharic is much bigger than for German and English.

This issue attributes to the morphological richness of Amharic, which is revealed in a very high-vocabulary growth rate as presented in Figure 2.

We have also computed the distribution of word lengths in terms of the pronunciation units.[3] As can be seen from Figure 3, there are Amharic words with up to

[3] We have used our grapheme-based pronunciation dictionary for Amharic, the beep1.0 pronunciation dictionary for English (available from ftp://svr-ftp.eng.cam.ac.uk/comp.speech/dictionaries/beep.tar.gz) and the Verbmobil pronunciation dictionary (Finke *et al.* 1997) for German.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

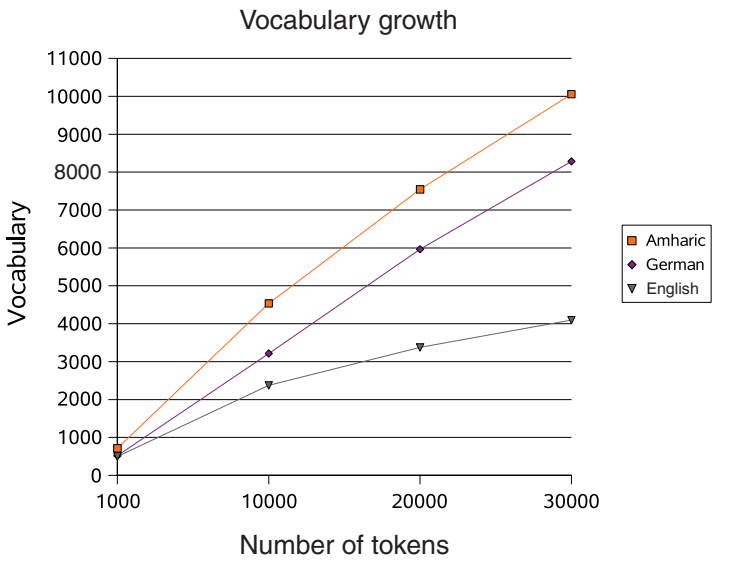


Fig. 2. (Colour online) Vocabular growth rate of Amharic, English, and German.

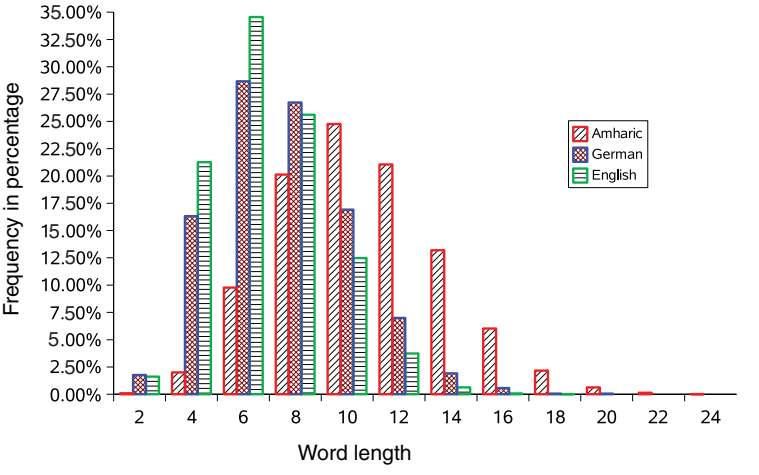


Fig. 3. (Colour online) Word length distribution of Amharic, English, and German.

24 pronunciation units long but the maximum word lengths observed in English and German are 18 and 20, respectively. The highest proportion of Amharic words are 8–12 pronunciation units long while it falls between 4 and 8 for English and German.

We can understand, therefore, that the use of words as lexical and language model units for morphologically rich languages like Amharic imposes a severe data sparseness problem, high-vocabulary growth rate, high-OOV word rate, computational load of the pattern matching component that handles concatenation of lexemes (words) into whole hypothesized utterances. These issues motivate the use of subword lexical and language model units for Amharic.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

3.2 Word segmentation

Developing a subword language model requires use of a word parser which splits word forms into its constituent morphemes. There have been a number of attempts to develop morphological analyzers for Amharic based on different approaches (Bayou 2000; Alemayehu and Willett 2002; Bayu 2002; Fissaha and Haller 2003; Adafre 2005; Amsalu and Gibbon 2005; Argaw and Asker 2007; Gasser 2011). However, most of the systems were either not suited for our purpose or have not been publicly available when needed. Moreover, morphologically annotated data for Amharic does not yet exist. Therefore, we adopted Morfessor (Creutz and Lagus 2005), a language-independent unsupervised morphology learning tool, to produce a morphologically segmented corpus.

The Morfessor baseline model that we have used in our experiment is based on the minimum description length (MDL)[4] criterion. Based on the MDL principle, the Morfessor baseline learns a lexicon of morphemes which is concise and produces a compact representation for the words in the corpus (Creutz and Lagus 2005). Details about Morfessor and how it works can be found in (Creutz and Lagus 2005).

A word type list of the ATC_120k corpus has been used as input to Morfessor. Morfessor learns the morpheme segmentation from the input data and presents as output a word and its constituent morphemes (called statistical morphemes hereafter).

3.2.1 Evaluation of the Word Segmentation

The segmentation performance can be evaluated by comparing morpheme boundaries proposed by the algorithm with a linguistic gold standard. Creutz (2006) indicated that such an evaluation is straightforward and intuitive provided that an adequate gold standard exists. Since, to our knowledge, there is no gold standard segmentation available for Amharic,[5] we manually prepared one for about 1,000 words. The words have been taken from the output segmentation (type-based segmentation) systematically. That means after taking the first word, every fiftieth word has been taken and the correct segmentation for it has been provided. Table 4 shows examples for the output of Morfessor and that of our manual segmentation. We can see the errors in Morfessor's segmentation.

However, similar to Creutz (2006), we considered Amharic stems as a correct segmentation although they should have been analyzed into root and pattern. Morfessor's segmentation of words (type-based segmentation) has been compared with our manual segmentation and precision, recall and f-measure (see Table 5) have been calculated using the evaluation scripts of Hutmegs package (Creutz and Lind 2004). The f-measures for Amharic are higher than the one reported by Creutz

[4] Readers who are interested in MDL can consult the resources available at http://www.mdl-research.org/reading.html.

[5] Our manually segmented corpus that is presented in Section 5.1 cannot be used to evaluate the performance of Morfessor's segmentation because it includes nonconcatenative morphology, which is not the case in Morfessor segmentation.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

Table 4. *Morfessor segmented data*

Words	Morfessor's segmentation	Manual segmentation
č'ama	č'ama	č'ama
č'amačäw	č'a + mačäw	č'ama + čä + w
č'amačäwn	č'ama + čäwn	č'ama + čä + w + n
č'amačäwnIna	č'ama + čäwnI + na	č'ama + čä + w + nI + na
č'amačnIn	č'ama + čnIn	č'ama + č + nI + n
č'amana	č'ama + na	č'ama + na
č'aman	č'ama + n	č'ama + n
č'amaw	č'ama + w	č'ama + w
č'amawn	č'ama + wn	č'ama + w + n
č'amawočačäwn	č'ama + wočačäwn	č'ama + woč + ačä + w + n
č'amawoč	č'ama + woč	č'ama + woč
č'amawočn	č'ama + wočn	č'ama + woč + n
č'amawočnIna	č'ama + wočnI + na	č'ama + woč + nI + na
č'amawoču	č'ama + woču	č'ama + woč + u

Table 5. *Evaluation results of the segmentation*

Measures	Type-based seg.
Precision	82.19%
Recall	54.59%
f-Measure	65.60%

(2006) for Arabic (41.7%) using the same version of Morfessor but on a data set consisting of 17,000 distinct words. This might be due to the amount of data used for evaluation and the way the gold standards have been prepared. The Arabic data set is very big compared to the one used to evaluate the Amharic segmentation (only 1,000 word types). While the Arabic gold standard has been generated automatically based on a lexicon of Egyptian Colloquial Arabic, the Amharic one was prepared manually which would result in a more accurate segmentation compared to the automatic method.

3.3 The language models

The words in our corpus (ATC_120k) have been replaced automatically by their constituent morphemes. The resulting morpheme-segmented corpus (called ATC_120k_Seg, hereafter) consists of 4,035,656 tokens and 15,923 morpheme types. This corpus is divided into three parts: training set, development set, and test set with proportions of 80:10:10. The splitting is only for the language-modeling experiment. Perplexities, log probabilities, and OOV rate are calculated on the development set and test set. The development set is used to find the optimal options of language modeling (parameter tuning), for example, the best smoothing technique. The best performing model is then evaluated on the test set. However, word recognition accuracies and

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

Table 6. *Perplexity of statistical morpheme-based language models*

N-gram	Smoothing technique	Perplexity
Pentagram	Natural Discounting	117.37
Quadrogram	Good-Turing with Katz backoff	113.24
Pentagram	Absolute Discounting	112.79
Pentagram	Witten–Bell	110.88
Quadrogram	Modified Kneser–Ney	107.54
Pentagram	Unmodified Kneser–Ney	103.63

OOV rates reported in Section 6 (speech recognition experiments) are not based on the development and test sets of this corpus. They are calculated based on the 5k development set which is part of the read speech corpus (Abate *et al.* 2005) used in our investigation.

Then, statistical morpheme-based n-gram models of order 2–5 were trained using the SRI Language Modeling Toolkit (SRILM) (Stolcke 2002). The effect of different smoothing techniques (Good-Turing, Absolute Discounting, Witten–Bell, Natural Discounting, Modified and Unmodified Kneser–Ney[6]) on the quality of a language model were then studied. The best results (in terms of perplexity) obtained for each smoothing technique are presented in Table 6. As can be seen from Table 6, the best performing model is a pentagram model with unmodified Kneser–Ney smoothing. This result is in line with the findings of Chen and Goodman (1998), that Kneser–Ney and its variations outperform other smoothing techniques.

Probability estimates of different n-gram order have been interpolated for Witten–Bell, Absolute Discounting, and Modified Kneser–Ney smoothing techniques. Table 7 shows the best results for each smoothing technique. Interpolating n-gram probability estimates of a specific order n with lower order estimates sometimes yield better models. This is because, especially for low counts, lower order distributions provide valuable information about the correct amount to discount (Chen and Goodman 1998). Our experiment verified this fact. A pentagram model with Modified Kneser–Ney smoothing and interpolation of n-gram probability estimates has a perplexity of 101.38. For the other smoothing techniques an increase in perplexity has been observed.

The best performing model had a perplexity of 102.59 on the test set.

As indicated by Stolcke (Stolcke 2002), discarding unknown words or treating them as a special 'unknown word[7]' (unk) token affects the quality of language models. A language model in which unknown words are treated as unk assigns a nonzero probability to OOV words by virtue of the mapping of OOVs to unk.

[6] Modified Kneser–Ney smoothing method differs from Unmodified Kneser–Ney in that, it uses three different discounts, instead of a single one, that are applied to n-grams with one, two, and three or more counts, respectively.

[7] In morpheme-based language model, unknown words means unknown morphemes since morphemes are considered as tokens.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

Table 7. *Perplexity results with interpolation*

N-gram	Smoothing technique	Perplexity
Quadrogram	Witten–Bell	112.1
Pentagram	Modified Kneser–Ney	101.38
Quadrogram	Absolute Discounting	118.38

Table 8. *Perplexity of word-based models*

N-gram	Smoothing technique	Perplexity
Trigram	Good-Turing with Katz backoff	1,151.29
Pentagram	Absolute Discounting	1,147.04
Pentagram	Witten–Bell	1,236
Pentagram	Natural Discounting	1,204.14
Quadrogram	Modified Kneser–Ney	1,107.32
Pentagram	Unmodified Kneser–Ney	1,078.16

This might improve the predictive power of the language model. Thus, unknown morphemes have been mapped to a special 'unknown morpheme' token for the best model from Table 7 but instead of the expected improvement an increase in perplexity (to 102.26) was observed. This might indicate that the OOV problem is already handled by the use of morphemes as units of the language model.

4 Word-based language models

To relate the results for statistical morpheme-based language models, we have also developed word-based ones. For this purpose, we used again the corpus (ATC_120k) from which the morpheme-segmented corpus has been derived. Table 8 shows the perplexity of word-based models. The pentagram model with Unmodified Kneser–Ney is the best model compared to the other word-based language models.

The best perplexity has been obtained with a pentagram language model using Modified Kneser–Ney smoothing, interpolation of n-gram probability estimates, and a mapping of unknown words to a special 'unknown word' token. This model yields a perplexity of 879.25 and 873.01 on the development and test sets, respectively. Unlike the morpheme-based models, the word-based model benefited from the mapping of unknown words to the unk token.

Abate (2006) reported a maximum perplexity of 167.89 for a bi-gram word-based model. Although it is not possible to directly compare these results, because different test sets and measurement procedures have been used, we did not expect such a big difference. To find a possible explanation, we trained and evaluated word-based language models using the ATC_120k corpus under the same conditions as described in Abate (2006), where the HTK toolkit (Young *et al.* 2006) has been

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

used. To compute a perplexity, Hidden Markov Model Toolkit (HTK) converts a bigram model into a lattice format and then generates sentences from the lattice.

Using this method the perplexity of our bigram word-based language model decreased to 239.45. The remaining difference could be explained by the nature of the data used in the experiments. While ATC_120k corpus contains many long sentences (35,626 and 4,434 sentences in the training and test sets, respectively) with more than 20 words, the sentences in the training data used in Abate (2006) are usually shorter (only 51 long sentences out of a total of 73,895). The test set contains none at all. To test this hypothesis, we conducted another experiment following the same procedure and using ATC_Short corpus obtained by excluding the long sentences from ATC_120k. Under these conditions the perplexity was reduced to 181.63.

Obviously, the problem with the method used by Abate (2006) is that it calculates the perplexity from automatically generated sentences and there is no guarantee for the correctness of these sentences. In addition, when the test is conducted repeatedly, the perplexity values vary as the sentences generated are different. In our experiment with short sentences (containing 20 or less words), the perplexity values varied between 181 and 202. This makes a direct comparison with perplexities as measured on real test sentences even more difficult.

Although we expect that the high perplexity of our word-based language models is caused by the morphological richness of the language, spelling errors might also contribute. To estimate the influence of spelling errors, we have conducted control experiments. For these experiments, two data sets have been prepared: data_set_I and data_set_II. Note that, 10,000 sentences of the ATC_48k have been manually checked for spelling errors. The resulting spelling error-free sentences have been merged with 11,917 sentences that are read for a speech corpus (Abate *et al.* 2005) as this will assure us that there is no spelling error. This forms data_set_I. Data_set_II is prepared in the same way except that the spelling errors in the 10,000 sentences have not been corrected. These data have been divided into training set, development set, and test set with a proportion of 80:10:10. Word-based language models have been developed using the training data and perplexities have been computed on the test sets. The result shows that spelling errors can be excluded as an explanation of the high perplexity and the morphological feature of the language seems to be the sole source for the high perplexity of our word-based language models.

4.1 Comparison of word-based and statistical morpheme-based language models

Since test sets have quite different token counts for word- and morpheme-based language models, the perplexity values of these two kinds of language models are not directly comparable. Even so, a comparison can be made based on the probabilities the models assign to the test sets, where a model with a higher probability is considered a better one. To avoid underflow problems, log probabilities are commonly used. Table 9 depicts the log probabilities (calculated from the language model development set which is a 10% of the available data and is different from the training set and test set) of the best statistical morpheme- or word-based

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

Table 9. *Log probabilities*

Model	Log probabilities
Best performing morpheme-based model	−834,495
Corresponding word-based model	−705,218
Best performing word-based model	−726,095
Corresponding morpheme-based model	−836,215

models as well as the corresponding[8] word-based or statistical morpheme-based models.

As can be seen in Table 9, word-based models have a higher log probability and, therefore, should be considered superior ones although their perplexity is higher than that of the morpheme-based ones. On the other hand, subword-based language models offer the benefit of reducing the OOV words by 99%. This is a great achievement, as the OOV words problem is severe in morphologically rich languages in general, and Amharic in particular.

Our findings are somewhat contradictory so far. While perplexity measures favor the morpheme-based approach, the probability-based evaluation indicates an advantage for the word-based account. Therefore, a comparison within the framework of a target application is required to find out which model is better suited to improve the performance on the particular task. We used speech recognition for that purpose.

5 Linguistic morpheme-based language models

5.1 Manual word segmentation

Words in Amharic can be decomposed into root, pattern, and one or more affix morphemes, which are combined by intercalation and concatenation. Since Morfessor is not able to deal with the nonconcatenative feature, we manually segmented 72,428 word types found in a corpus of 21,338 sentences (419,660 tokens). By doing so, we aimed at obtaining an optimistic estimate of what a fully automatic procedure could achieve at best if it were available. In contrast to the mere segmentation provided by Morfessor, we complemented the manual segmentation with an annotation of the root and the pattern whenever appropriate.

We do not claim the manual segmentation to be comprehensive. Since a word type list has been used, there is only one entry even in case of polysemy or homonymy. For example, the word *t'Iru* might be an adjective which means 'good' or it might be an imperative verb which has a meaning 'call' (for second person plural). Consequently, the word has two different segmentations. Nevertheless, we provided only one segmentation based on the most frequent meaning of the word in

[8] Corresponding language models means models developed in the same manner as the best language models except different units used.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

our text. The geminated and nongeminated word forms, which might have distinct meanings and consequently different segmentations, have also been treated in the same manner as the polysemous or homonymous ones. Because the transcription system does not indicate the gemination of consonants, a systematic ambiguity remains. For instance, *?at'ägäbu* can be treated as an adverb which means 'next to him' or as a verb with a meaning 'they made somebody else full or they satisfied (somebody else)' based on the gemination of the consonant g. This word could have, therefore, been segmented in two different ways: [*?at'ägäb + u*] if it is an adverb or [*?a + t'gb + ää + u*] if it is a verb.

Another point is related to the plural forms of words borrowed from Geez[9] which are formed by epenthesis and vocalic changes. For example, the plural form of the noun *kokäb* - 'star' is *käwakIbt*. There are also some Amharic words whose plural forms are formed through partial reduplication. For instance, the plural form of the adjective *tIllIk'* - 'big' is *tIlallIk'*. It would have been easier to analyze such words as, for example, *käwakIbt = kokäb + NPl*[10] or *tIlallIk' = tIllIk' + AdjPl*[11]. However, we need an analysis that gives the morphemes themselves. Thus, such words have not been segmented since it is difficult to identify their morphemes. Nevertheless, we believe that these limitations do not significantly affect the quality of our manually segmented corpus since only the predictive power of the language model for the complete input strings is important instead of the particular segmentation.

Based on the list of segmented words, we automatically substituted each word in the corpus (from which the word type list for the manual segmentation was derived) with its segmentation, thus producing a linguistically morpheme-segmented corpus (manually_seg). It consists of 1,141,434 tokens and 11,154 distinct morphemes.

5.2 Comparison of linguistic morpheme-based language model with word- and statistical morpheme-based language models

To find out whether the manual segmentation contributes to an improvement of model quality, we also segmented the same text corpus using Morfessor, and consequently obtained a statistically morpheme-segmented version of the corpus (morfessor_seg). The same text corpus (without being morphologically segmented) has also been used to train word-based models. That means we have three versions of the corpus: an unsegmented version for training word-based models and a linguistically morpheme-segmented as well as a statistically morpheme-segmented ones. Each of the three corpora has been divided into three parts: training, development, and test sets in the proportions of 80:10:10. Language models with several smoothing techniques and n-gram orders have been trained on them. The best model for each corpus is presented in Table 10.

[9] Another Semitic language mostly used in the liturgy of Ethiopian and Eritrean Orthodox täwahdo churches.
[10] A short form for plural noun.
[11] A short form for plural adjective.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

Table 10. *Perplexity difference according to the type of corpus*

Corpus	Smoothing techniques	Perplexity	Logprob	OOV(%)
Manually_seg	Unmodified Kneser–Ney	21.597	−153,178	0.55
Morfessor_seg	Unmodified Kneser–Ney	99.49	−161,241	0.06
Word_based	Unmodified Kneser–Ney	1,000.86	−116,387	11.62

Table 11. *Effect of interpolation*

Corpus	Smoothing techniques	Perplexity	Logprob	OOV(%)
Manually_seg	Modified Kneser–Ney	20.703	−151,071	0.55
Morfessor_seg	Modified Kneser–Ney	97.297	−160,459	0.06
Word_based	Modified Kneser–Ney	997.387	−116,329	11.62

The models developed with manually segmented data outperform the ones developed with automatically segmented data in terms of perplexity.[12] If, however, the log probability is considered, the word-based model still seems to fare better than the morpheme-based ones. With regard to OOV words, the models developed using automatically and manually segmented corpora led to lower OOV rate. In contrast, the number of OOV words in word-based models is very high. One can also see that Unmodified Kneser–Ney smoothing technique performed well irrespective of the corpus/data.

Interpolation improved (see Table 11) all the models regardless of the kind of corpus used. However for the word-based model, the increase in log probability is not as large as it is for the others. The language models developed using manually segmented data consistently outperformed (in terms of perplexity) the ones developed using automatically segmented data.

6 Speech recognition experiments

The language-modeling experiments presented so far have shown that using morphemes in language modeling reduces the OOV rate considerably. This is a big achievement as the OOV words are a major source of error in systems such as automatic speech recognizers. This section presents the experiments conducted to study the effect of OOV rate reduction (as a result of using morphemes) on the performance of speech recognition.

Morpheme-based language models can be applied to a speech recognition system in two different ways. The first is a lattice rescoring framework in which morphemes

[12] The perplexity and the OOV words reported here for manually segmented data differ from the ones reported in (Tachbelie and Menzel 2009). This is because in Tachbelie and Menzel (2009) words have been left unsegmented as it was difficult to decide on their proper segmentation. Later these words have all been segmented in consultation with a linguist and according to recommendations in the literature.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

are used only in the language-modeling component, while the lexicon still consists of words. Therefore, this method cannot help analyze the contribution of morpheme-based language models to a reduction of the OOV rate. On the other hand, it avoids the effect of acoustic confusability[13] on the performance of speech recognition. The second approach uses morphemes as lexical and language modeling units of the speech recognition system. Although this method suffers from acoustic confusability and limited n-gram language model scope,[14] it enables us to see the contribution of morphemes to the reduction of the OOV rate. Thus, the second approach has been adopted.

We first compare three systems with full-word dictionaries of different size. Then, we use the morpheme-based approach and study its effect on the performance of the system. Since the root-pattern morphological feature of the language will pose a problem to reconstruct words from linguistic morphemes, and therefore requires a lattice rescoring approach, only statistical morphemes have been used in the morpheme-based speech recognition experiment. The next section presents the speech corpus used in both word- and morpheme-based speech recognition experiments.

6.1 The speech corpus

The speech corpus used to develop the speech recognition system is a read speech corpus (Abate *et al.* 2005). It contains 20 hours of training speech collected from 100 speakers who read a total of 10,850 sentences (28,666 tokens). Compared to other speech corpora that contain hundreds of hours of speech data for training, this corpus is obviously small in size and accordingly the models will suffer from a lack of training data. Again, the 5k development set has been used for the purpose of this investigation.

6.2 Word-based recognizers

6.2.1 Acoustic, lexical, and language models

The acoustic model consists of 6,610 crossword triphone Hidden Markov Models (HMMs) each with three emitting states. The states of these models and all the crossword triphone models that are potentially needed for recognition are tied using decision-tree-based state-clustering that reduced the number of triphone models from

[13] A lexical unit u is said to be acoustically confusable when a typical feature representations x_u encoding u will cause acoustic models of other lexical unit o to produce likelihood scores similar to or higher than the score of u's own model (Bouwman *et al.* 2004). In other words, a pair of lexical units are considered as confusable pairs when the first lexical unit is likely to be misrecognized by the second one. This happens most frequently when the lexical units are very short.

[14] The scope of an n-gram language model is the longest distance it covers to capture dependencies. For example, a word-based trigram language model captures the dependencies up to three consecutive words. However, a morpheme-based trigram model may not capture dependencies among three words unless the three consecutive morphemes are words by themselves.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

Table 12. *Word OOV rate on the 5k development set*

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

77,658 to 10,215. Their mixture is added incrementally and 12 Gaussian mixtures have been found to be optimal.

The vocabulary for the three full-word form pronunciation dictionaries has been prepared by taking the most frequent words from the ATC_120k text corpus described in Section 3.1. Table 12 shows the OOV rates on the 5k development set that consists of 360 sentences (4,106 tokens and 2,836 distinct words) for these vocabularies. Although we tried to optimize the vocabularies by taking the most frequent words, the OOV rate is still high. The pronunciation dictionaries have been encoded using a simple procedure that takes advantage of the orthographic representation which is a consonant vowel syllable.

The text corpus from which the vocabularies have been selected has also been used to train the language models. As there are three dictionaries (5k, 20k, and 65k), we have developed three trigram language models one for each vocabulary using the SRILM toolkit (Stolcke 2002). 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.

6.2.2 Performance of word-based speech recognizers

Speech recognition experiments have been performed using the 5k, 20k, and 65k vocabularies. In each case the system has been evaluated with the 5k development set. The results are reported in terms of WER and syllable error rate (SER). SER has been used to provide a fair comparison with the morpheme-based recognition performance as the WER of these systems might be affected by the generation of illegal morpheme sequences. Figure 4 presents both WER and SER for each vocabulary. As it can be noted, both, the OOV rate and the error rates decrease as the vocabulary size grows. The best performance (21.5% WER or 8.1% SER) has been obtained for the 65k vocabulary which has the smallest OOV rate. The results show that the OOV rate highly affects the performance of a speech recognition system.

6.3 Morpheme-based recognizers

To deal with the problems faced with a word-based speech recognition system for Amharic, such as the effect of OOV rate (see Section 6.2.2), morphemes instead of words have been considered as dictionary entries and modeling units for the language models.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

Table 13. *Morpheme OOV rate on the 5k development set*

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

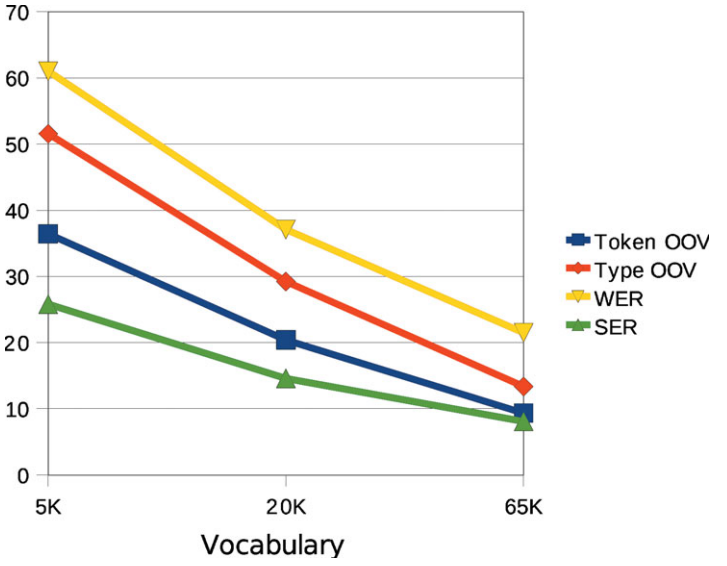


Fig. 4. (Colour online) Recognition error rate of word-based recognizers.

6.3.1 *Acoustic, lexical, and language models*

The acoustic model has been developed in a similar fashion as for the word-based recognizers. The training data is covered by a set of 6,459 cross-morpheme triphone HMMs each with three emitting states. The states of these models and all the possible cross-morpheme triphone models are tied and, therefore, the number of triphone models is reduced from 57,799 to 7,685. Similar to the word-based models, 12 Gaussian mixtures have been found to be optimal.

The Morfessor segmented ATC_120k corpus (see Section 3.3) has been used in the development of acoustic, lexical, and language models of the morpheme-based recognizers. 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. From the morphologically segmented corpus, three dictionaries have been prepared: 5k and 20k by selecting the most frequent morphemes and 28.4k by considering all of them. The morpheme-based OOV rates of these vocabularies on the 5k development set are presented in Table 13 which shows that the OOV rate is considerably reduced as a result of using morphemes.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

Table 14. *Word OOV rate of morpheme-vocabularies on the 5k development set*

Vocabulary	Token OOV (%)	Type OOV (%)
5k	18.39	26.06
20k	1.19	1.73
All (28.4k)	0.05	0.07

Table 15. *Performance of morpheme-based speech recognizers*

Vocabulary	MER(%)	WER(%)	SER(%)
5k	44	49.2	19.7
20k	31.8	37.5	15.1
28.4k	31.2	36.6	14.8

In order to analyze the benefit of morpheme-based system in terms of covering OOV words, we calculated a word-based OOV rate by counting the number of words which can not be rewritten in terms of the morphemes found in the vocabularies. Table 14 shows the word-based OOV rate of the morpheme-based systems. The result is comparable to the content of Table 12. As can be seen, the use of morphemes reduced the OOV rate for words to a similar degree. The token OOV rate for the 5k morpheme vocabulary, for instance, is reduced by almost 50% from the token OOV rate of the 5k full-word vocabulary (*cf.* Table 13).

Using the three dictionaries (5k, 20k, and 28.4k), we have developed three open vocabulary morpheme-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.

6.3.2 Performance of morpheme-based speech secognizers

The morpheme-based speech recognition system has been evaluated on the 5k development set using the 5k, 20k, and 28.4k morpheme vocabularies. The results are reported in terms of morpheme error rate (MER), WER, and SER. The WER has been computed after words have been obtained by concatenating the recognized morpheme sequence. The best performance (see Table 15) has been obtained with the 28.4k morpheme vocabulary which has an OOV rate of 0.03. Since the OOV rate is very small, an even better performance than the one reported here was expected. The reasons for this disappointing result (in spite of having a small OOV rate) might be a higher acoustic confusability and the limited language model scope which are peculiar to morpheme-based speech recognition systems.

6.4 Comparison of word- and morpheme-based speech recognizers

The very low OOV rate of morpheme vocabularies has a positive effect on speech recognition accuracy, especially for small vocabularies, namely the 5k one. As can be

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

Table 16. *Performance of word-based speech recognizers*

Vocabulary	WER(%)	SER(%)
5k	61	25.8
20k	37.1	14.6
65k	21.5	8.1

seen in Tables 15 and 16,[15] while the word-based model produces a WER of 61% and SER of 25.8% for the 5k vocabulary, the morpheme-based system improves WER and SER to 49.2% and 19.7%, respectively, for the 5k morpheme vocabulary.[16] This corresponds to an absolute WER reduction of 11.8% and an SER reduction of 6.1%. The error rate reduction is statistically significant (for both WER and SER) with $p < 0.001$.

The 20k morpheme-based speech recognizer performed slightly worse (with 37.5% word and 15.1% SER) than the equivalent word-based system which has a WER of 37.1% and a SER 14.6%. However, there is no statistically significant difference between the two systems at $p = 0.05$. The performance of the recognizer with the 28.4k morpheme vocabulary is only slightly better than that of the 20k word-based recognizer although it includes all the morphemes 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.

To see the effect of acoustic confusability, we have computed, using National Institute of Standards and Technology (NIST's) scoring package (sclite),[17] the statistics of the confusion pairs for the best performing word- and morpheme-based recognition systems. The result shows that there are a total of 677 confusable word pairs out of which only 2 have a frequency of 3 and only 7 have a frequency of 2. On the other hand, there are a total of 1,376 confusable morpheme pairs and the frequency of confusion goes up to 13. There are 4 and 11 confusable morpheme pairs with frequency of 4 and 3, respectively. We have also seen 32 confusable morpheme pairs with frequency of 2. The summary is given in Table 17.

The problem of acoustic confusability has also been commented on by Pellegrini and Lamel (Pellegrini and Lamel 2006) who suggested the use of higher order n-gram models. Accordingly, we also experimented with higher order morpheme-based language models in our morpheme-based speech recognizers.

[15] This table presents the performance of word-based recognizers presented in Figure 4 for easier comparison.

[16] Comparing the morpheme-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 morpheme-based systems are also disfavored by the concatenation of illegal morpheme-sequences, an increasing number of small and acoustically confusable units, and a limited language model scope.

[17] ftp://jaguar.ncsl.nist.gov/current_docs/sctk/doc/sclite.htm

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

Table 17. *Acoustic confusion of word- and morpheme-based ASR*

Frequency	Word	Morpheme
Above 10	0	1
5–9	0	0
2–4	9	47
1	668	1,328
Total	677	1,376

Table 18. *Lattice rescoring with a quadrogram morpheme-based language model*

Vocabulary	MER(%)	WER(%)	SER(%)
20k	68.92	35.9	14.7
28.4k	69.70	34.9	14.2

Since it is not possible to use n-gram models higher than trigram in the HTK decoding tool, 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. As can be seen from Table 18, a 1.2% absolute WER reduction (compared to the 20k word-based recognizer) has been obtained for the 20k vocabulary morpheme-based recognizers. However, this improvement is statistically insignificant at $p = 0.05$. With respect to SER, the word-based system still performs slightly better than the morpheme-based one (14.6% versus 14.7%). The accuracy of the 28.4k system which has a very low OOV rate (0.03) and covers all the morphemes in the text is still worse than the accuracy of the 65k word-based recognizer (21.5% WER or 8.1% SER) which has a considerably higher OOV rate (9.33). On the other hand, the result is still better than the one achieved with a trigram morpheme-based language model. This might indicate that acoustic confusability is the most influential factor for the performance degradation of a morpheme-based speech recognizer on large vocabularies.

We conclude from these experiments that the use of morphemes as lexical and language modeling units instead of words contributes to a reduction in the number of tied acoustic models, a lower OOV rate of the lexical model and a lower perplexity of the language model. The cumulative effect results in a better performance of the morpheme-based ASR and a higher decoding speed.

As it has been given in Sections 6.2.1 and 6.3.1 the number of cross-word triphones is 10,215, while that of cross-morpheme triphones is 7,685. To see the effect of the difference in the acoustic models on the ASR performance, we have compared the performance of 5k word- and morpheme-language models on both acoustic models. The result shows that with 10,215 cross-word triphone models, the WER of the morpheme-based language model is 42.6% while that of the word-based language model is 61.0% making their absolute difference 18.4%. With 7,685 cross-morpheme

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

triphone models, the WER of the morpheme-based language model is 49.2% while that of the word-based language model is 66.4% that makes their absolute difference 17.2%. This shows that in any of the acoustic models the morpheme-based language model scores a better performance than the word-based language model.

7 Conclusions and future direction

In this paper we described an attempt to develop morpheme-based language models for Amharic and to use them in ASR. Since Amharic is one of the less resourced languages, we have used an unsupervised and language-independent morphological parsing tool (Morfessor) in the course of our experiment.

A substantial reduction of the out of vocabulary rate has been observed as a result of using subwords or morphemes. This addresses a severe problem in morphologically rich languages like Amharic. Moreover, low perplexity values have been obtained with morpheme-based language models compared to word-based ones. However, when comparing the models based on the probability they assign to a test set, word-based models seem to be advantageous. Therefore, recognition experiments have been conducted to study the utility of morpheme-based language models in the context of a real application.

We conducted a series of speech recognition experiments with words and morphemes as lexical units and measured the performance of the systems in terms of WER and SER. We found that using morphemes as dictionary entries and language model units reduces the OOV rate considerably and as a consequence improves the performance of a relatively small vocabulary (5k) recognizers significantly. However, as the morpheme vocabulary grows, the morpheme-based recognizers did not yield a notable reduction in the error rate, which might be due to a higher acoustic confusability and a limited language model scope. Rescoring lattices with a higher order morpheme-based language model (quadrogram) reduced the WER further. However, this reduction is not statistically significant, indicating that acoustic confusability is an influential factor for the poor performance of large vocabulary (20k or more) morpheme-based speech recognition systems.

Although the morpheme-based recognizer benefits from the low OOV rate, it suffers from the smaller and therefore acoustically more confusable units. Further improvement should, therefore, be obtained if care is taken (for instance, using confusion constraints as in Pellegrini and Lamel (2007)) 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 the concatenation of illegal morpheme sequences. Attempts in this direction may also boost the performance of a morpheme-based speech recognizer. For example, constraints (such as 'the subject marker morpheme never appears at the beginning of a morpheme sequence') could be used to prevent the system from considering illegal morpheme sequences.

In order to see the effect of the difference in the number of triphones on the performance of recognition systems, we used the same number of triphones for morpheme- and word-based systems and evaluated their performance on 5k test set.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

As indicated in Section 6.4, the current result shows that morpheme-based system outperformed the word-based one regardless of the number of triphones used in the system. However, further experiments with 20k and above vocabulary size should be conducted in order to make the result generalizable. This experiments will be conducted in the near future.

We applied the morpheme-based language models to an ASR task. Of course, the morpheme-based language models can also be applied to other natural language applications. Since the problem of acoustic confusability is not an issue in SMT, an improvement in performance might be obtained as a result of using morphemes in Amharic SMT. Thus, the use of morpheme-based language models for Amharic SMT is an interesting endeavour. Nowadays, morphological decomposition, and consequently morpheme-based language models are used in SMT into morphologically complex languages (Labaka *et al.* 2007; Badr, Zbib and Glass 2008; Oflazer 2008). Further work on the use of morphemes in Amharic machine translation would be a useful continuation.

In our work, we have used a current state of the art ASR architecture that assumes words/morphemes are concatenated to form whole sentences by means of the language model. It is clear that such an approach is not appropriate to model nonconcatenative phenomena of morphology. Since Amharic exhibits nonconcatenative morphology, it would be good to develop other ASR architectures that could deal with the nonconcatenative morphological feature of Amharic in particular and Semitic languages in general. In the future, we will continue our research in this direction.

Acknowledgment

We would like to thank the anonymous reviewers for their invaluable comments and suggestions on the manuscript of this paper.

References

Abate, S. T. 2006. *Automatic Speech Recognition for Amharic*. PhD. thesis. University of Hamburg.

Abate, S. T., Menzel, W., and Tafila, B. 2005. An Amharic speech corpus for large vocabulary continuous speech recognition. In *Proceedings of 9th European Conference on Speech Communication and Technology, Interspeech-2005*, Lisbon, Portugal, pp. 1601–4.

Adafre, S. F. 2005. Part of speech tagging for Amharic using conditional random fields. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Ann Arbor, Michigan, USA, pp. 47–54.

Alemayehu, N., and Willett, P. 2002. Stemming of amharic words for information retrieval. *Literary and Linguistic Computing* **17**(1): 1–17.

Amsalu, S., and Gibbon, D. 2005. Finite state morphology of Amharic. In *Proceedings of International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria, pp. 47–51.

Argaw, A. A., and Asker, L. 2007 An amharic stemmer: reducing words to their citation forms. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources - ACL 2007*, Prague, Czech Republic, pp. 104–10.

Badr, I., Zbib R., and Glass J. 2008. Segmentation for English-to-Arabic statistical machine translation. In *Proceedings of ACL-08, HLT Short Paper*, Columbus, Ohio State, pp. 153–6.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

Bayou, A. 2000. *Developing Automatic Word Parser for Amharic Verbs and their Derivation.* Master's thesis, Addis Ababa University.

Bayu, T. 2002. *Automatic Morphological Analyzer for Amharic: An Experiment Employing Unsupervised Learning and Autosegmental Analysis Approaches.* Master's thesis, Addis Ababa University.

Bender, M., Bowen, J., Cooper, R., and Ferguson, C. 1976. *Languages in Ethiopia.* London: Oxford University Press.

Bouwman, G., Cranen, B., and Boves, L. 2004. Predicting word correct rate from acoustic and linguistic confusability. In *Proceeding the 8th International Conference on Spoken Language Processing*, Jeju Island, Korea, pp. 1481–4.

Chen, S. F., and Goodman, J. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

Creutz, M. 2006. *Induction of the Morphology of Natural language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition.* PhD thesis, Helsinki University of Technology.

Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar, M., and Stolcke, A. 2007. Analysis of morph-based speech recognition and the modeling of out-of-vocabulary words across languages. In *Proceedings of NAACL HLT 2007*, Rochester, NY, USA, pp. 380–7.

Creutz, M., and Lagus, K. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.1. Technical Report A81, Neural Networks Research Center, Helsinki University of Technology.

Creutz, M., and Lind, K. 2004. Morpheme Segmentation Gold Standards for Finnish and English. Technical Report A77, Helsinki University of Technology.

El-Desoky, A., Gollan, C., Rybach, D., Schlüter, R., and Ney, H. 2009. Investigating the use of morphological decomposition and diacritization for improving arabic LVCSR. In *Proceedings of INTERSPEECH 2009*, Brighton, UK, pp. 2679–82.

Finke, M., Geutner, P., Hild, H., Kemp, T., Ries, K., and Westphal, M. 1997. The Karlsruhe - Verbmobil speech recognition engine. *Proceedings of International Conference on Acoustic, Speech, and Signal Processing, ICASSP 1997* **1**: 83–6.

Fissaha, S., and Haller, J. 2003. Amharic verb lexicon in the context of machine translation. *Proceedings of the 10th Conference on Traitement Automatique des Langues Naturelles*, Butz-sur-Mer, France, **2**: 183–92.

Gasser, M. 2011. Hornmorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya. In *Proceedings of the Conference on Human Language Technology for Development - HLTD 2011*, Alexandria, Egypt, pp. 94–9.

Geutner, P. 1995. Using morphology towards better large-vocabulary speech recognition systems. *Proceedings of IEEE International on Acoustics, Speech and Signal Processing* **1**: 445–8.

Heintz, I. 2010. *Arabic Language Modeling with Stem-derived Morphemes for Automatic Speech Recognition.* PhD thesis, Graduate Program in Linguistics, The Ohio State University.

Hirsimäki, T., Creutz, M., Siivola, V., and Kurimo, M. 2005. Morphologically motivated language models in speech recognition. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, Espoo, Finland, pp. 121–6.

Ircing, P., Krebc, P., Hajič, J., Psutka, J., Khudanpur, S., Jelinek, F., and Byrne, W., 2001. On large vocabulary continuous speech recognition of highly inflectional language - Czech. In *Proceeding of the European Conference on Speech Communication and Technology*, Aalborg, Denmark, pp. 487–9.

Junqua J. C., and Haton J. P. 1996. *Robustness in Automatic Speech Recognition: Fundamentals and Applications.* London: Kluwer Academic.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press

Kirchhoff, K., Bilmes, J., Henderson, J., Schwartz, R., Noamany, M., Schone, P., Ji, G., Das, S., Egan, M., He, F., Vergyri, D., Liu, D., and Duta, N. 2002. Novel speech recognition models for Arabic. Technical Report, Johns-Hopkins University Summer Research Workshop.

Labaka, G., Stroppa, N., Way, A., and Sarasola, K. 2007. Comparing rule-based and data-driven approaches to spanish-to-banque machine translation. In *Proceedings of Machine Translation Summit XI, 2007*, Copenhagen, Denmark.

Manning, C. D., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge and London: The MIT Press.

Oflazer, K. 2008. Statistical machine translation into a morphologically complex language. In *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing CICLing'08*, Haifa, Israel. Published as LNCS vol. 4919, pp. 376–87.

Pellegrini, T., and Lamel, L. 2006. Investigating automatic decomposition for ASR in less represented languages. In *Proceedings of INTERSPEECH 2006*, Pittsburgh, Pennsylvania, USA, pp. 285–8.

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

Pellegrini, T., and Lamel, L. 2009. Automatic word decompounding for ASR in a morphologically rich language: application to amharic. *IEEE Transactions on Audio, Speech, and Language Processing* **17**(5): 863–73.

Sproat, R. 1992. *Morphology and Computation*. London: The MIT Press.

Stolcke, A. 2002. SRILM – an extensible language modeling toolkit. *Proceedings of International Conference on Spoken Language Processing* **2**: 901–4.

Tachbelie, M. Y. 2010. *Morphology-Based Language Modeling for Amharic*. PhD thesis, University of Hamburg. http://www2.sub.uni-hamburg.de/opus/volltexte/2010/4848/pdf/TachbelieDissertation.pdf. Last accessed on the 3rd of December, 2012.

Tachbelie, M. Y., and Menzel, W. 2009. Morpheme-based language modeling for inflectional language – Amharic. In N. Nicolov, G. Angelova, and R. Mitkov (eds.), *Recent Advances in Natural Language Processing V, Current Issues in Linguistic Theory*, pp. 301–10. Amsterdam and Philadelphia: John Benjamin's Publishing.

Teferra, A., and Hudson, G. 2007. *Essentials of Amharic*. Köln: Köppe.

Voigt, R. M. 1987. The classification of central semitic. *Journal of Semitic Studies* **32**: 1–21.

Vergyri, D., Kirchhoff, K., Duh, K., and Stolcke, A. 2004. Morphology-based language modeling for Arabic speech recognition. In *Proceedings of International Conference on Spoken Language Processing*, Jeju Island, Korea, pp. 2245–8.

Whittaker, E., and Woodland, P. 2000. Particle-based language modeling. In *Proceeding of International Conference on Spoken Language Processing*, Beijing, China, pp. 170–3.

Yimam, B. 2007. *YäamarIña Säwasäw*, 2nd ed. Addis Ababa: EMPDE.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. 2006. *The HTK Book*. Cambridge University, Engineering Department. Cambridge, UK.

https://doi.org/10.1017/S1351324912000356 Published online by Cambridge University Press