Enhanced Morfessor Algorithm with Phonetic Features: application to Turkish

Ebru Arısoy [1], *Thomas Pellegrini* [2], *Murat Saraçlar* [1], *Lori Lamel* [3]

[1] Boğaziçi University, Istanbul, Turkey
[2] INESC-ID, Lisbon, Portugal
[3] LIMSI-CNRS, Orsay, France

{arisoyeb, murat.saraclar}@boun.edu.tr, thomas.pellegrini@l2f.inesc-id.pt, lamel@limsi.fr

Abstract

This paper describes the application of the enhanced Morfessor algorithm with phonetic features to Turkish. Previous research on Turkish Automatic Speech Recognition (ASR) has shown the superiority of sub-word units over words as lexical items. Among the proposed sub-lexical approaches, the statistical Morfessor algorithm is a popular choice due to its ease of use and ASR performance. Here, baseline Morfessor algorithm is enhanced with a basic phonetic knowledge of Turkish. Phone-based distinctive features specific to Turkish and phonetic confusion constraints are incorporated into the Morfessor algorithm. The ASR performance of the proposed modifications are evaluated using the Turkish Broadcast News transcription system. The best performance, achieved with distinctive features of the consonants, is 1.1% better than the baseline Morfessor algorithm. This decompounding configuration provides the most compact model.

1. Introduction

Turkish, being an agglutinative language with rich morphology, presents a challenge for ASR systems. The productive morphology of Turkish yields many unique word forms. Due to the computational limitations and in order to obtain robust language model estimates, only a limited number of words, usually the most common words in the ASR application domain, are used as the recognition vocabulary in state-of-the-art ASR systems. This limited recognition vocabulary results in high number of out-of-vocabulary (OOV) words, especially for agglutinative and highly inflectional languages, such as Turkish, Finnish, Estonian, Czech and Amharic. Even for vocabulary sizes that would be considered as large for English, the OOV rates for agglutinative and highly inflectional languages are quite high. It was shown that with an optimized 60K lexicon the OOV rate is less than 1% for North American Business news [1]. However, the OOV rates are 9.3% for Turkish with a 50K vocabulary, 15% for Finnish with a 69K vocabulary [2], 10% for Estonian [3] and 8.3% for Czech [4], with 60K vocabularies. As a rule of thumb, an OOV word brings up on average 1.5 recognition errors [5]. Therefore, high OOV rates directly translate into high word error rates (WERs).

Using sub-lexical units in ASR is a common approach proposed for agglutinative languages to handle the OOV problem caused by using word vocabularies. In this approach, the recognition lexicon is composed of sub-lexical units instead of words. The sub-lexical units in the lexicon should be capable of covering most of the words of a language to address the OOV problem and is therefore should lead to an improvement in recognition accuracy. Therefore, logical choices of word segments, which are considered as "*meaningful*" in terms of ASR, can be used as the sub-lexical units. The "*meaningful*" word segments are the ones that carry enough acoustic information for discriminating lexical items and that can be used as histories in predicting the next units.

In agglutinative languages words are formed by concatenation of stems and affixes. Therefore, grammatical units such as, stems, affixes or their groupings can be considered as natural choices of sub-lexical units in ASR systems [6, 7, 8, 9, 10, 11]. They are obtained by using language dependent rule-based morphological analyzers. The splitting of words into sub-words is straightforward with morphological analyzers. However, morphological analyzers may suffer from OOV problem due to many proper names and foreign words that usually occur in news texts, since a limited root vocabulary is compiled in the morphological analyzers together with the morphotactic and morphophonemic rules. For instance, a Turkish morphological analyzer [12] with 54.3K roots can analyze 96.7% of the word tokens and 52.2% of the word types in a text corpus of 212M words with 2.2M unique words. Even though stems and affixes are natural sub-lexical choices, the need for expert knowledge of the language makes them inapplicable to languages lacking of morphological tools.

In order to handle the drawbacks of grammatical sub-lexical units, their statistical counterparts have been proposed [13, 14, 15]. Statistical sub-lexical units are morpheme-like units. They are obtained with data driven approaches, usually in an unsupervised manner, instead of morphological analyzers. The main advantage of this model compared to grammatical models is that it does not require an expert knowledge of the language. Therefore, it can easily be applied to any language. However, the splitting of words into sub-lexical units are not trivial in statistical segmentations. The statistical morpheme-like units are not supposed to match with the exact grammatical morphemes, however, they should yield the meaningful unit criteria. Therefore, different algorithms are investigated to obtain reasonable morpheme-like units with statistical techniques. These algorithms only require a raw text corpus to learn the word segmentations. However, a basic phonetic knowledge of the language can be used to improve the segmentations [16, 17].

Morfessor [14] is one of the popular unsupervised word decompounding algorithms, applied to agglutinative languages. Morfessor-based sub-lexical units gave promising accuracy improvements over the baseline word model for Finnish, Estonian and Turkish ASR systems [2, 3, 18, 19, 20]. Incorporating phonetic features to the baseline Morfessor algorithm is proposed in [16, 17] and accuracy improvements are obtained for a less

	Not rounded		Rounded	
	Open	Close	Open	Close
Posteriors	a,[a]	ı,[ɯ]	o,[o]	u,[u]
Anteriors	e,[e]	i,[i]	ö,[ø]	ü,[y]

Table 1: Turkish vowels with their [IPA] symbols.

represented language, Amharic.

This paper is an application of the enhanced Morfessor algorithm with phonetic features to Turkish. Our results demonstrate the effectiveness of the phonetic features in decompounding Turkish text for ASR. The paper is organized as follows: The characteristics of Turkish is presented in Section 2. In Section 3 we introduce the acoustic and text data used in word segmentations and in ASR experiments. We present the data-driven word decompounding in Section 4. Section 5 explains the Broadcast News transcription system and gives the results for the decompounding algorithms. Finally, Section 6 concludes the paper.

2. Characteristics of Turkish

Turkish is a member of Altaic family of languages. The main characteristics of Turkish are the agglutinative morphology and the vowel harmony. These features distinguish Turkish as a challenging language for natural language processing and speech recognition applications.

As a result of the agglutinative morphology, many new words can be derived from a single stem by addition of several suffixes. There are no prefixes in Turkish. Figure 1 shows concatenated nominal and verbal inflections. The verbal inflection is more complex than the nominal one. Although there is not a one to one correspondence between Turkish morphemes and English words, we can say that one Turkish word may correspond to a group of English words. This agglutinative nature causes the vocabulary to expand significantly which is problematic for speech recognition.

nominal inflection: `ev-im-de-ki-ler-den`
(among those in my house)
verbal inflection: `yap-tır-ma-yabil-iyor-du-k`
(It was possible that we did not make someone do it)

Figure 1: Norminal and verbal inflection examples for Turkish

Vowel harmony is another characteristic of Turkish. According to one of the vowel harmony rules, a stem ending with a back/front vowel takes a suffix starting with a back/front vowel. Vowel harmony is not a problem when word based models are used for speech recognition. However if sub-words are used as language modeling units, we need to take vowel harmony into account since concatenation of sub-words may result in word-like units with incorrect morphophonemics.

Turkish is almost a phonetic language. This property led us to utilize graphemes instead of phonemes in acoustic modeling. Turkish consists of 29 graphemes, 8 vowels and 21 consonants. Tables 1 and 2 give respectively the vowel and consonant inventories for Turkish. These properties will be used in deciding the phonetic features for Turkish word decompounding.

3. Data description

The Turkish Broadcast News (BN) audio corpus is used in acoustic modeling. This corpus has been collected at Boğaziçi University since 2007. It contains BN recordings from a radio channel (VOA) and four different TV channels (CNN Türk, NTV, TRT1 and TRT2). The annotation of the corpus includes topic, speaker and background information according to Hub4 BN transcription guidelines. In this study, we use approximately 71 hours of speech from the corpus as the acoustic data. This data is partitioned into disjoint training and test sets. Table 3 gives the size of the audio corpus for acoustic model training and test. The reference transcriptions of the acoustic training data include 485K words.

Labial	b [b], p [p], f [f], m [m], v [v]
Dental	d [d], t [t], s [s], z [z], n [n], l [l], r [r]
Palatal	c [dʒ], ç [tʃ], ş [ʃ], j [ʒ], y[j]
Velar	g [g], k [k], v [w]
Uvular	h [h]

Table 2: Turkish consonants with their [IPA] symbols. The consonant ğ is ignored in the table since it is used to lengthen the previous vowel.

Train	*Test*	*Total*
68h 36min	2h 30min	71h 06min

Table 3: Broadcast News training and test audio data.

For language modeling, the main corpus consists of 96.4M words, coming from news papers, news wires, specialized articles (medicine, technologies, etc.), and also some literature texts. A subset of 11.6M words, called *Boğaziçi* in this article, has been used to train the word decompounding models. Table 4 gives the size of the texts used to train the language models and the word decompounding models.

	Source	*# Words*
Language modeling	*Transcriptions*	485K
	Main corpus	96.4M
Word decompounding	*Boğaziçi*	11.6M

Table 4: The number of word tokens in the text data used to train the languages models and the word decompounding models. *Boğaziçi* is the subset of the *main corpus*.

4. Data-driven word decompounding

To tackle the very high OOV word rates arising in Turkish LVCSR from its very rich morphology, word decompounding is almost a mandatory pre-processing step.

The data-driven word decompounding Morfessor algorithm [14] has extensively been used to decompound lexical units as a prior step for Turkish LVCSR, as in [19, 20, 21]. In [17], an enhanced version of this algorithm has been successfully used for LVCSR for another morphologically rich language, the Amharic Ethiopian official language. New properties incorporating “oral” cues showed 0.7% absolute accuracy improvement on the word baseline, giving 23.6% WER for Amharic BN transcription. These properties, briefly described hereafter, have been adapted and tested for Turkish.

4.1. Modifications to the Morfessor Algorithm

Morfessor has two purposes: first, the training of a word segmentation model given a lexicon with optional frequency counts. Training uses a maximum a posteriori (MAP) criterion based on several text properties, including word frequencies and string probabilities. Second, a previously learnt decomposition model can be used to decompound a new word list. New words, i.e. words that are not in the model, can also be decomposed into morphs that exist in the decompounding model. For further information about Morfessor, please refer to [14].

All the properties used in the Morfessor program are based on written language and do not incorporate any "oral" properties that could be useful for ASR. Two main modifications have been made to enhance Morfessor: a phone-based feature, called 'DF' for distinctive features, and a constraint called 'Cc' that tries to prevent phonetic confusion among units arising from the decompoundings, due to the smaller size of the morphs used as lexical recognition units.

The DF property is language specific since its features depend on the phones of the language. Vowel and consonant distinctive features for Turkish are generated by using the phonetic inventories in Tables 1 and 2. For instance, the properties associated with the vowel a[a] are open, not rounded and posterior. The vowel o[o] differs from the vowel a[a] as being rounded. If only these 3 properties are considered, the feature vector representations of the vowels a[a] and o[o] will be [1 0 1] and [1 1 1] respectively. Being rounded or not rounded is the only distinctive feature between these vowels. The feature vector representation of each consonant is generated in the same way as the vowels using Table 2. The DF property is incorporated in the Morfessor framework as an additional term in the *a priori* probability estimate. For more information about the DF calculation, please see [17].

The 'Cc' option forbids word splits that would result in confusion-prone morphs. Unlike the DF option, Cc is not incorporated in the probability computation, but it corresponds to a yes/no decision to keep a morph candidate. Previous syllabotactic alignments are used to identify syllable confusion pairs. The constraint compares the word split candidates with the other morphs in the lexicon, syllable by syllable. If a candidate differs from another morph by only one syllable that has been found as a confusion pair with this morph, then the split is forbidden.

The end-of-word probability has been also modified, to produce more word splits. This last modification over the baseline is called 'H' since it is inspired by the Harris' observation that this number decreases naturally from the word start [22], and the work on German word decomposition [23].

Table 5 summarizes the different options investigated with the decompounding algorithm.

Option	*Comment*
BL	Baseline word based system, no decompounding
M	Baseline Morfessor 1.0
M H	M + modified 'Harris'
M H DFV	M H + distinctive features of vowels only
M H DFC	M H + distinctive features of consonants only
Cc	+ confusion constraint

Table 5: Decomposition options compared in this study.

4.2. Decompounding the training texts

To get rid of misspelled words and artifacts, only the words occurring at least three times were selected from the Boğaziçi corpus. This led to a 197K word lexicon for this corpus.

Only the Boğaziçi lexicon, 197K words, has been used to generate the word decompounding models. The decompounded lexicon was used as decompounding model to decompound all the words of the transcriptions and the 96.4M word text corpus.

The second column of Table 6 gives the number of words or morphs for each system. The baseline, word model without decompounding, gives 1.6M distinct words (also called word types). After decomposition, the lexicons with the Cc constraint yield around 150K words, which corresponds to a 10.7 factor reduction. The 'Cc' constraint limits the number of decompositions, so that the lexicon sizes are bigger with this option. The other option sets, with no Cc, have smaller word lists, around 115K units. Finally, the M H TDC set gave the smallest lexicon, with only 49.5K units.

5. ASR experiments

ASR experiments are performed for each decompounding model using the Turkish BN transcription system [20]. The acoustic and language models used in the ASR system are the same with the ones in [10]. The sub-word-based language models are specific for each option set. The same set of phone-based acoustic model is used for all the experiments.

5.1. Experimental Setup

For the acoustic models, we used Broadcast News data with acoustic signals and their transcriptions. Acoustic models are speaker-independent. They are adapted to each TV/Radio channel using supervised MAP adaptation on the training data, giving us the channel adapted acoustic models. We use decision-tree state clustered cross-word triphone models with approximately 7500 HMM states. Since Turkish is almost a phonetic language, graphemes were utilized instead of phonemes. Each state of the speaker independent HMMs has a GMM with 11 mixture components. The HTK [24] frontend was used to get the MFCC based acoustic features. The training and decoding tasks were performed using the AT&T tools [25]

Language models with interpolated Kneser-Ney smoothing as well as entropy based pruning [26] were built using the SRILM toolkit [27]. In order to reduce the effect of pruning on the recognition accuracy, the first-pass lattice outputs were re-scored with unpruned language models. Language models built with the *main corpus* and the reference *transcriptions* were linearly interpolated to reduce the effect of out-of-domain data. The optimized interpolation coefficient for the transcription language model is 0.4 for all the systems. In order to facilitate converting sub-word sequences into word sequences, the word boundaries are marked with a special symbol (#). The ratio of the sub-word tokens to the word tokens including the word boundary symbol is given in Table 6. The ratios are greater than 2 since word boundary tags (#), which are used to recombine morphs together, are counted in the number of words. The ratios over 2 suggest higher order n-gram language models for sub-word-based modeling. 3-gram language models gave the best performance for the baseline word model. Therefore, 5-gram language models were built for the sub-word based systems to have a comparable span with words.

The lexicon size for all the systems has been limited to the 50K most frequent units due to computational limitations. For the baseline word model, the OOV rate is 9.3%, which is very high compared to the common OOV values, less than 1%, for languages like English or French. With the sub-word unit lexicons, the OOV rate is considered to be "0%", since all the words in the test data can be generated by any combination of the sub-words in the 50K lexicon.

In the sub-word-based models, the word corpora was pre-processed to generate the sub-word units. First, word boundary

Options	*# word or morph types*	*# Morphs / # Words*	AUL (in types)	AUL (in tokens)	*WER(%)*
BL	1.6M	2.0	10.3	6.4	39.6
M	115.6K	2.04	8.0	5.6	36.6
M Cc	152.7K	2.02	8.0	5.7	37.3
M H	115.1K	2.04	8.0	5.6	36.6
M H Cc	151.5K	2.02	8.0	5.7	37.3
M H TDV Cc	153.0K	2.02	8.0	5.7	37.4
M H TDC	***49.5K***	***2.24***	***6.2***	***4.2***	***34.8***
M H TDC Cc	103.7K	2.11	7.8	5.3	35.3
M*	***45.8K***	***2.4***	***6.7***	***4.5***	***35.9***

Table 6: Number of lexical units (# word or morph types), # Morphs / # Words ratio, Average Unit Length (AUL) in types and tokens, and Word Error Rates (WER), for the different option configurations. M* is also the baseline Morfessor model, however, obtained with different parameter settings.

symbols were inserted between words. Second, words in the corpora were decomposed into sub-words with the Viterbi algorithm using the previously learnt models. Then, 5-gram language models were built with the most frequent 50K sub-word units and the sub-word corpora. After decoding the test data and rescoring the lattice output with unpruned language models, the sub-words between consecutive word boundary symbols were concatenated to obtain word-like units. WER performance was evaluated by comparing the word-like units with the reference transcriptions of the test data.

5.2. Experimental Results

The last column of Table 6 presents the performances for all the systems in terms of WER. The best system is the M H TDC which yields a 4.8% absolute WER reduction over the word baseline. This decompounding configuration, modified 'Harris' and distinctive features of consonants, provided the most compact model with the smallest lexicon. All the sub-word unit based systems achieved WER reductions over the word-based system, but the Cc constraint that limits the number of morphs yielded smaller WER reductions. The best results were achieved without the 'Cc' option, i.e. with the systems that presented the smallest lexicons.

The success of the M H TDC model can lead up to the conclusion that models resulting in smaller sub-word lexicons outperform the models with higher lexicon sizes. In our recognition experiments, because of limiting the recognition vocabulary to 50K, lexicon size may have an effect on the accuracy. In order to investigate whether smaller vocabulary size is the only reason of the best model's success, we changed the parameters of the baseline Morfessor algorithm and rerun it to obtain a smaller vocabulary. This model is labeled as M* in Table 6 and it resulted in 45.8K sub-word types. M H TDC model is absolutely 1.1% better than the M* model. This result clearly shows the effect of incorporating distinctive features of consonants into the Morfessor model in decompounding Turkish words.

In [17], the best performance for Amharic was obtained with the 'Cc' option and worse performances were obtained with the models resulting in smaller lexicon sizes. These results are in contradiction with the ones in this paper. This contradiction may arise due to the differences in the experimental setups between Amharic and Turkish. First, the vocabulary sizes for all models were set to 50K for Turkish experiments due to the computational limitations, whereas, all the sub-words were utilized as the lexical items in Ahmaric experiments. Sub-word lexicons with 50K units gave full coverage over the test data, mostly due to the shorter length sub-word units frequently occurring in the corpus, as a result being the lexical items. The fourth and the fifth columns of Table 6 clearly shows how the average unit length (AUL) changes when calculated over types and tokens for the units in the lexicon. Therefore, limiting the vocabulary size for sub-words may favor the models having smaller number of sub-word types after decomposition. Second, handling of the morph recombination into full words are different in the setups. In Turkish, a word boundary tag, (#), is used. This method has the advantage to not distinguish morphs and words, for the morphs that are both affix and single word. In the Amharic experiments, a '+' sign was added to the prefixes, so that some confusions between both forms of a single morph could arise. Another explanation of the contradiction, could be the difference in the baseline OOV rates. The OOV rate is higher in the Turkish experiments than in the Amharic ones, respectively 9.6% and 6.8%. Thus the lexical coverage for Turkish benefits more from numerous word decompositions.

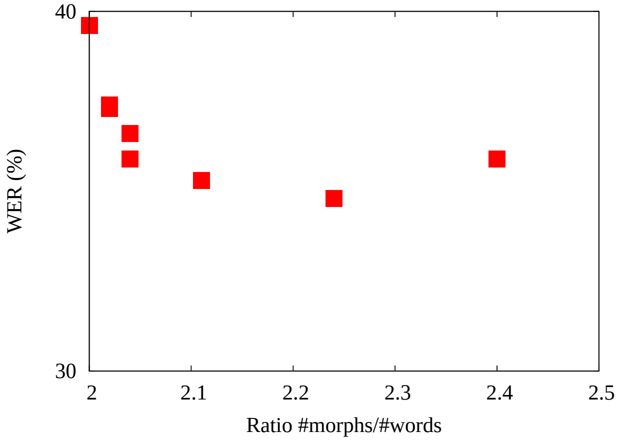

Figure 2: WER as a function of the (#morphs /#words) ratio

There is an interesting apparent relationship between the WER and the ratio of the number of morphs to the number of words. Table 6 and Figure 2 illustrate this relation. The best system M H TDC shows the biggest ratio, after M*, with a 2.24 value. This system is also the one with the lowest AUL. In our experiments, 2.24 can be the optimal morph to word ratio resulting in the most compact model.

6. Conclusion and future work

In this paper, the performance of the data-driven Morfessor algorithm modified with phonetic features is investigated for Turkish. All the sub-word experiments perform better than the reference word baseline and show the superiority of sub-word units in modeling Turkish language. For the enhancement, phonetic features specific to Turkish are incorporated into the Morfessor model. The distinctive features of the consonants achieves the best performance, absolutely 1.1% better than the original algorithm.

Vowel distinctive features are also incorporated into the Morfessor model, however, their performance is even worse than the original one. Our hypothesis for the performance degradation caused by the vowel DF is that the number of vowel features are less than the number of consonant features. Therefore, DF parameter plays a crucial role in consonants and splits more words than the vowel features do.

Future directions of our research will include a key question raised by the observation of the relationship between the WERs and the unit ratios. Is there a way of selecting the best performing model without running ASR experiments? If there are any related parameters with the WER, decompounding algorithms can be modified to produce sub-words that will optimize the WER.

7. Acknowledgements

The authors would like to thank Sabanci and ODTU universities for the Turkish text data and AT&T Labs – Research for the software. This research is supported in part by Turkish State Planning Organization (DPT) under the project number DPT2007K120610, The Scientific and Technological Research Council of Turkey (TUBITAK) under the project numbers 108E113, 105E102 and 107E261 and Boğaziçi University Research Fund (BAP) under the project numbers 05HA202 and 07HA201D. Ebru Arısoy was supported by TUBITAK BDP (Unified Doctorate Program) in 2004-2008. Thomas Pellegrini was supported by the DGA (Doctorate Program) in 2004-2007.

8. References

[1] R. Rosenfeld, "Optimizing lexical and n-gram coverage via judicious use of linguistic data," in *Proc. the European Conference on Speech Communication and Technology*, 1995, pp. 1763–1766.

[2] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen, "Unlimited vocabulary speech recognition with morph language models applied to Finnish," *Comput. Speech. Lang.*, vol. 20, no. 4, pp. 515–541, 2006.

[3] M. Kurimo, A. Puurula, E. Arısoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraçlar, "Unlimited vocabulary speech recognition for agglutinative languages," in *Proc. HLT-NAACL*, New York, USA, 2006, pp. 487–494.

[4] P. Podvesky and P. Machek, "Speech recognition of Czech – inclusion of rare words helps," in *Proc. the ACL Student Research Workshop*, Ann Arbor, Michigan, USA, 2005, pp. 121–126.

[5] I. L. Hetherington, "A characterization of the problem of new, out-of-vocabulary words in continuous-speech recognition and understanding," Ph.D. dissertation, Massachusetts Institute of Technology, 1995.

[6] K. Çarkı, P. Geutner, and T. Schultz, "Turkish LVCSR: Towards better speech recognition for agglutinative languages," in *Proc. ICASSP*, İstanbul, Turkey, 2000, pp. 1563–1566.

[7] E. Mengüşoğlu and O. Deroo, "Turkish LVCSR: Database preparation and language modeling for an agglutinative language," in *Proc. ICASSP, Student Forum*, Salt-Lake City, UT, USA, 2001.

[8] A. O. Bayer, T. Çiloğlu, and M. T. Yöndem, "Investigation of different language models for Turkish speech recognition," in *Proc. IEEE SIU*, Antalya, Turkey, 2006, pp. 1–4.

[9] T. Çiloğlu, M. Çömez, and S. Şahin, "Language modelling for Turkish as an agglutinative language," in *Proc. IEEE SIU*, Kuşadası, Turkey, 2004, pp. 461–462.

[10] E. Arısoy, H. Sak, and M. Saraçlar, "Language modeling for automatic Turkish broadcast news transcription," in *Proc. Interspeech-Eurospeech*, Antwerp, Belgium, 2007, pp. 2381–2384.

[11] E. Arısoy, H. Dutağacı, and L. M. Arslan, "A Unified Language Model for Large Vocabulary Continuous Speech Recognition of Turkish," *Signal Processing*, vol. 86 (10), pp. 2844–2862, 2006.

[12] H. Sak, T. Güngör, and M. Saraçlar, "Turkish language resources: Morphological parser, morphological disambiguator and web corpus," in *Proc. GoTAL, LNAI 5221*, 2008, pp. 417–427.

[13] J. Goldsmith, "Unsupervised learning of the morphology of a language," *Computational Linguistics*, vol. 27 (2), pp. 153–198, 2000.

[14] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0," Helsinki University of Technology, Publications in Computer and Information Science Report A81, March 2005.

[15] C. Monson, "Paramor: From paradigm structure to natural language morphology induction," Ph.D. dissertation, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2009.

[16] 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*, Antwerp, 2007, pp. 1797–1800.

[17] T. Pellegrini and L. Lamel, "Automatic word decompounding for ASR in a morphologically rich language: application to Amharic," *To appear in IEEE Transactions on Audio, Speech and Language Processing*, vol. Special issue on morphologically rich languages, 2009.

[18] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arısoy, M. Saraçlar, and A. Stolcke, "Analysis of morph-based speech recognition and the modeling of out-of-vocabulary words across languages," in *Proc. HLT-NAACL*, Rochester, NY, USA, 2007, pp. 380–387.

[19] E. Arısoy and M. Saraçlar, "Lattice Extension and Vocabulary Adaptation for Turkish LVCSR," *Speech and Language Processing*, vol. 17 (1), 2009.

[20] E. Arısoy, D. Can, S. Parlak, H. Sak, and M. Saraçlar, "Turkish Broadcast News Transcription and Retrieval,"

IEEE Transactions on Audio, Speech and Language Processing, vol. Special issue on morphologically rich languages, 2009.

[21] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arısoy, M. Saraçlar, and A. Stolcke, "Morph-Based Speech Recognition and Modeling of Out-of-Vocabulary Words Across Languages," *ACM Transactions on Speech and Language Processing*, vol. 5.1 Article 3, 2007.

[22] Z. Harris, "From phoneme to morpheme," *Language*, vol. 31, pp. 190–222, 1955.

[23] M. Adda-Decker, "A corpus-based decompounding algorithm for German lexical modeling in LVCSR," in *Proceedings of EUROSPEECH*, Geneva, 2003, pp. 257–260.

[24] S. Young, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.2), Entropic Cambridge Research Laboratory," 2002.

[25] M. Mohri and M. D. Riley, "Dcd library, speech recognition decoder library, AT&T Labs – Research. http://www.research.att.com/sw/tools/dcd/," 2002.

[26] A. Stolcke, "Entropy-based pruning of backoff language models," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, 1998, pp. 270–274.

[27] A. Stolcke, "SRILM – An extensible language modeling toolkit," in *Proc. ICSLP*, vol. 2, Denver, 2002, pp. 901–904.