# TURKISH DICTATION SYSTEM FOR BROADCAST NEWS APPLICATIONS

*Ebru Arısoy and Levent M. Arslan*

Electrical and Electronics Engineering Department, Boğaziçi University
34342, Bebek, İstanbul, Turkey
phone: + (90) 212 3596414, fax: + (90) 212 2872465, email: arisoyeb@boun.edu.tr, arslanle@boun.edu.tr

**ABSTRACT**

We have designed a Turkish dictation system for Broadcast news applications. Turkish is an agglutinative language with free word order. These characteristics of the language result in the vocabulary explosion, large number of out-of-vocabulary (OOV) words and the complexity of the N-gram language models in speech recognition when words are used as recognition units. Therefore, we proposed new recognition units. We parsed some of the words to smaller recognition units like stems, endings and morphemes, and introduced these smaller units and the unparsed words to the speech recognizer as lexicon entries. This way, we were able to overcome to the problem of large number of OOV words with a moderate vocabulary size and get better estimates for the N-gram language models. However, best recognition result was obtained using the word-based language model.

## 1. INTRODUCTION

Turkish is a challenging language for Large Vocabulary Continuous Speech Recognition (LVCSR) applications. The main reason is the agglutinative nature of the language, from the same root, very high number of words can be formed by suffixation [1]. Other reason is the free word order characteristic of Turkish, which leads to an increase in the perplexity of the language. Especially for English recognition engines, the mostly used language modelling units are the words. However, if this model is used in modelling agglutinative languages like Turkish, Finnish and Korean the OOV rate will be very high [2, 3], because it is impossible that the lexicon will contain all the words.

In a previous research on the statistical language modelling of Turkish [4], new language modelling units like ”stems and endings”, ”stems and morphemes” and ”syllables” are proposed. If words are selected as base recognition units, the OOV rate becomes very high. To overcome this problem, words are parsed into smaller recognition units. However, in terms of speech recognition performance, syllables and morphemes are smaller units and they lack enough acoustic information compared to larger units. Therefore, these language modelling units result in poorer recognition performance. Stems and endings model is proposed as a solution to the trade off between the small coverage in words and poor recognition performance in smaller recognition units. In this paper, we aim to decrease the perplexity and the OOV rate with an increase in the recognition performance. To achieve this goal, our main concern will be to use the combination of all the models given in [4]. Word-based model and the combined model will be applied on the data collected from the broadcast news.

This paper is organized as follows: In the next section, the Turkish morphology is introduced. Section 3 describes the proposed language models. Section 4 gives the statistics of the text corpus. Section 5 describes the recognition experiments and results of the proposed models. Conclusions and further research ideas based on these results are given in Section 6.

## 2. TURKISH MORPHOLOGY

Turkish is an agglutinative language; many new words can be formed by suffixation from the same root. The suffixes of Turkish are categorized as derivational or inflectional in terms of their function. Derivation is used to produce new lexical items, and it may change the grammatical category.

büz+gü (noun derived from verb stem)

kat+la (verb derived from noun stem)

tuz+luk (noun derived from noun stem)

kan+dır (verb derived from verb stem)

Nominal inflection only marks the grammatical notions like number, person, gender, and verbal inflection marks tense, aspect, modality and person. The morphotactics for verbal inflection is more complex than the nominal inflection.
nominal inflection: ev+im+de+ki+ler+den (one of those that were in my house)
verbal inflection: yap+tır+ma+yabil+iyor+du+k (It was possible that we did not make someone do it)

A popular example of word formation showing the complex morphotactics of Turkish is [5]:
”OSMANLILAŞTIRAMAYABİLECEKLERİMİZDENMİŞSİNİZCESİNE”
which can be decomposed into morphemes as:
OSMAN+LI+LAŞ+TIR+AMA+YABİL+ECEK+LER+İMİZ+DEN+MİŞ+SİNİZ+CESİNE
Its meaning is;
“as if you were of those whom we might consider not converting into an Ottoman”

Although there is not a one-to-one correspondence between Turkish morphemes and English words, we can clearly say that a single Turkish word can correspond to a group of English words. This example is the illustration of the drawback that we have to encounter if we apply the same methods used in English speech recognition engines directly to Turkish.

During the suffixation process, the first vowel of the morpheme must be compatible with the last vowel of the stem which is the rule of vowel harmony. Other characteristic of Turkish is the free word order. This is a challenging nature from the perspective of the speech recognizer. Although Turkish characterizes a *subject-object-verb* (SOV) type language, a sentence can be uttered in five different constitute orders [6]. The order of constitutes can be changed without effecting the grammar of the sentence. The effect is only to emphasize the meaning.

In the morphological parser [4, 7] used in this research, the stems and suffixes with their properties are defined; also the transitions between the morphemes are defined by the grammatical suffixation rules. There were 29540 nominal and verbal stems in the parser. Also 5963 new stems and 462 foreign words as nominal stems are added to the parser from the broadcast news corpus. During the parsing process, we do not deal with the morphological ambiguity problem. The parser output with longer stem is selected as longer units are better from the acoustic point of view.

## 3. PROPOSED LANGUAGE MODELS

In this section, we will be concentrating on the selection of base recognition units for speech recognition. Our proposed model will be the combination of word-based model, morpheme-based model and stem-ending-based model, and named as combined model. Word-based model and the combined model will be applied on the broadcast news corpus.

### 3.1 Word-based Model

In word-based model, words are selected as lexicon entries for speech recognition and language modeling probabilities are extracted from the training corpus using the words as units. The system for this model is illustrated in Figure 1. Here, Z is the model used to represent the short pauses between the consecutive words. The transition probabilities between the Z model and the word model, also between word models are calculated using bigram language models.

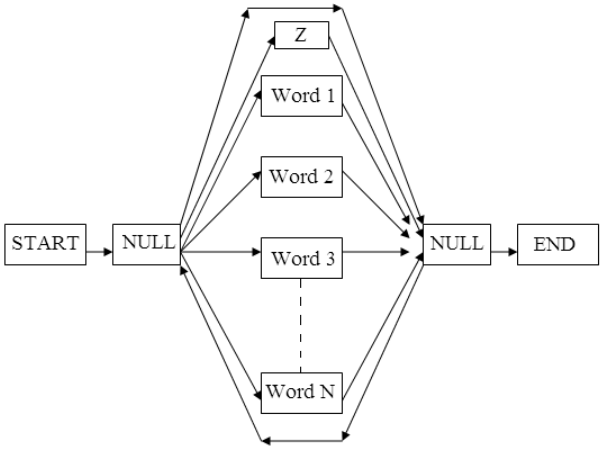


Figure 1: Word-based model

### 3.2 Combined Model

Combined Model is the combination of all the models proposed in [4]. The previously proposed models can be summarized as:

- **Word-based Model:**
  This is the same model explained in Section 3.1.
- **Morpheme-based Model:**
  This is the model where all the words are decomposed into their stems and morphemes and then these parts are taken as lexicon entries. The aim of this model is to reduce the vocabulary size and increase the coverage with the specified vocabulary by using stems and morphemes as vocabulary entries. However, most of the morphemes are small recognition units, which degrade the recognition performance compared to the classical word-based model.
- **Stem-ending-based Model:**
  This is the model where all the words are decomposed into their stems and endings and then these parts are taken as lexicon entries. The concatenation of morphemes which follows a stem is named as the ending. The proposed idea behind this model is to alleviate the problem of small recognition units by concatenating the morphemes. This idea for agglutinative languages is firstly proposed in [8], and this approach is applied to Turkish in [9]. Although the vocabulary size is increased compared to the previous model, recognition performance is better.

The combination of all these models will be our proposed combined model. The idea of the combined model is illustrated in Figure 2. As shown in the figure; the tokens, the base language modeling units, are both the words which are left as stems, stems, endings

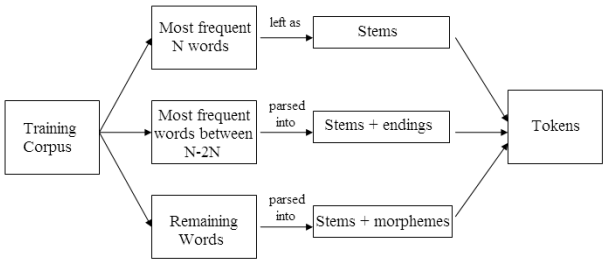


Figure 2: Basic idea behind the combined model

and the morphemes available in the training corpus. Firstly, all the words in the training corpus are sorted according to their frequency of occurrence. Then, most frequent N words are left as stems and most frequent N to 2N words are parsed into stems and endings. The remaining words are parsed into stems and their morphemes. All of the words which are left as stems, all the parsed stems, endings and morphemes are the tokens of the model. The system for this model is shown in Figure 3.

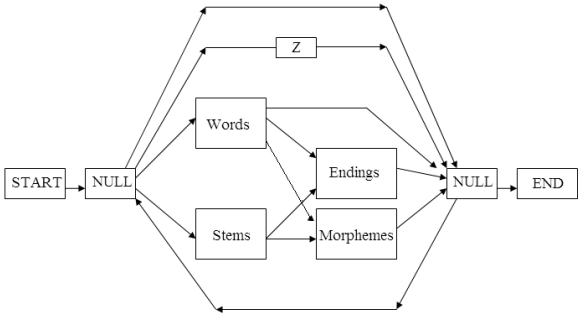


Figure 3: Combined model

In this figure, Z is the model used to represent short pauses between the consecutive words and NULL nodes are used to decrease the number of arcs in the representation. The arcs show the transitions between tokens. Also the transition probabilities between words to endings or morphemes, stems to endings or morphemes, also within the NULL nodes are calculated from the training corpus. In our proposed model, firstly we select N as 2500. This model is named as the combined-2.5K model. Then we increase the number of the most frequent words from 2500 to 5000. The aim is to see the effect of the number of lexicon entries left as words. This model is named as the combined-5K model.

## 4. STATISTICS OF THE CORPUS

### 4.1 Training Corpus

The text materials used in this paper are the articles of Milliyet newspaper belonging to different domains collected in a one month period. There are 355497 words in the training corpus. The training data are grouped as: Train-1, Train-2,.., Train-5, using the domain groups given in Table 1. The reason for grouping is to enlarge the training corpus by adding news from different domains and to see the effect of this on the statistics of the training corpus and test data.

### 4.2 Test Data

Our test data are the articles of the Milliyet newspaper collected in one day from five different domains. Training and test data are disjoint. There are 7016 words in the test data.

Table 1: Different domains in the training data groups

| | |
|---|---|
| Train-1 | World News |
| Train-2 | World News, Economics |
| Train-3 | World News, Economics, Contemporary News |
| Train-4 | World News, Economics, Contemporary News, Politics |
| Train-5 | World News, Economics, Contemporary News, Politics, Daily Life |

### 4.3 Number of Distinct Tokens

Number of distinct tokens is an important concept in the determination of the vocabulary size. It gives the minimum vocabulary size needed to cover 100% of the training data. In word-based model, the tokens will be the words in the training text and in the combined model the tokens will be the words, stems, endings and morphemes which are available in the training corpus. Table 2 gives the statistics of the training corpus in terms of the number of distinct tokens using word-based model. As shown in Table 2, the number of tokens and the number of distinct tokens are very high for this model. Addition of each domain introduces approximately 10 thousand new words to the corpus. Table 3 gives the token statistics of the training corpus for combined-2.5K model. In this model, the number of distinct tokens is comparable smaller than the word-based model. Also, the addition of each new domain to the corpus introduces approximately three thousand new tokens. Table 4 gives the token statistics for the combined model with the most frequent 5000 words. The only difference between this model and the previous model is the number of most frequent words. The addition of each new domain introduces approximately four thousand new tokens.

Table 2: Token statistics for the word-based model

| | Number of Tokens (words) | Number of Distinct Tokens (distinct words) | Number of New Distinct Tokens (distinct words) |
|---|---|---|---|
| Train-1 | 33534 | 10258 | 10258 |
| Train-2 | 119657 | 23275 | 13017 |
| Train-3 | 185037 | 35399 | 12124 |
| Train-4 | 289504 | 46996 | 11597 |
| Train-5 | 355497 | 55931 | 8935 |

Table 3: Token statistics for the combined-2.5K model

| | Number of Tokens | Number of Distinct Tokens | Number of New Distinct Tokens |
|---|---|---|---|
| Train-1 | 47676 | 5584 | 5584 |
| Train-2 | 165225 | 9538 | 3954 |
| Train-3 | 258952 | 13378 | 3840 |
| Train-4 | 406790 | 15762 | 2384 |
| Train-5 | 503863 | 18228 | 2466 |

Table 4: Token statistics for the combined-5K model

| | Number of Tokens | Number of Distinct Tokens | Number of New Distinct Tokens |
|---|---|---|---|
| Train-1 | 43638 | 6738 | 6738 |
| Train-2 | 151936 | 11366 | 4628 |
| Train-3 | 238625 | 15401 | 4035 |
| Train-4 | 374664 | 17874 | 2473 |
| Train-5 | 464214 | 20358 | 2484 |

If we compare these three models in terms of number of distinct tokens, the combined-2.5K model has the minimum number of distinct tokens. This can be explained with the morphological productivity of Turkish. Also the number of distinct tokens increases if we left more words as stems.

### 4.4 Coverage

Coverage is an important metric for the recognition performance of a recognizer. The words that are in the test set but not available in the lexicon are called OOV words. If a word is an OOV word, then the recognizer has no chance to recognize it correctly. Therefore OOV words are the main source of recognition errors and the coverage gives us a rough idea about the maximum theoretical performance of the recognizer. Figure 4 gives the coverage of the proposed models over the training and the test text with the specified vocabulary size. As shown from Figure 4, the smallest number of OOV rate is achieved with the combined-2.5K model and the highest OOV rate is achieved with the word-based model if we use the same vocabulary size. This results show that, using smaller tokens as recognition units, higher coverage is achieved with the same vocabulary size.

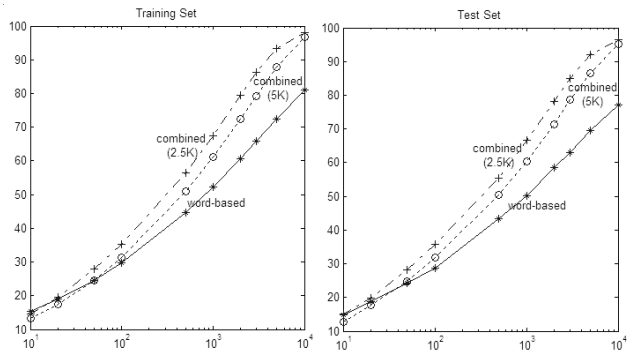


Figure 4: Coverage analysis over the training and the test set

### 4.5 Bigram Models

In this paper, bigram models are constructed using the Language Modeling tools of HTK [10]. Our comparison metric between models will be the perplexity, average branching factor, with respect to the self and the test set.

Table 5: Bigram analysis with respect to the self

| | Word-based model | Combined-2.5K model | Combined-5K model |
|---|---|---|---|
| Train-1 | 753.95 | 208.11 | 322.27 |
| Train-2 | 711.30 | 171.88 | 264.64 |
| Train-3 | 936.18 | 201.14 | 305.72 |
| Train-4 | 957.81 | 192.23 | 285.84 |
| Train-5 | 1063.53 | 197.58 | 291.67 |

Table 6: Bigram analysis with respect to the test set

| | Word-based model | Combined-2.5K model | Combined-5K model |
|---|---|---|---|
| Train-1 | 659.26 | 537.30 | 784.14 |
| Train-2 | 959.06 | 473.73 | 775.97 |
| Train-3 | 1105.38 | 384.82 | 636.86 |
| Train-4 | 1217.26 | 359.75 | 579.04 |
| Train-5 | 1278.17 | 334.45 | 528.91 |

Analyzing the results listed in the tables, we can say that combined-2.5K model is better in language modelling with the available corpus. When number of units are different for each

model, we know that bigram perplexity can not be a reliable comparison metric between models. However, it shows the certainty of each model with the increasing text data size. In word-based model, the addition of new data always increases the perplexity, which means that it increases the uncertainty of the model. However, in combined models the newly added data improves the certainty of the language model by leading to the better estimates over the previous bigram probabilities. In combined-2.5K model, most of the tokens are smaller units compared to combined-5K model, therefore first model is better in modelling the unseen data.

### 4.6 Statistics with Respect to Test Set

As mentioned before our test set is the daily news of the Milliyet newspaper collected from five different domains. The most frequent 10K tokens are selected as the optimum vocabulary size. The statistics of the test set in terms of coverage and perplexity analysis are given in Table 7.

Table 7: Coverage and perplexity analysis of the test set

| | Word-based model | Combined model (2.5K) | Combined model (5K) |
|---|---|---|---|
| Coverage(%) | 77.16 | 96.51 | 95.07 |
| Perplexity | 476.68 | 294.36 | 433.78 |

As shown from the table, the maximum coverage and the smallest perplexity values are achieved using the combined-2.5K model. In the next section, these three models will be compared in terms of speech recognition performance.

## 5. RECOGNITION EXPERIMENTS

We perform recognition experiments on the recordings (16 Khz, 16-bit with head-mounted Plantronics microphone) of the test data with only one female speaker. Each recording contains approximately 10 sentences, and each sentence is uttered in a manner like the continuous speech.

### 5.1 The Recognizer

The recognizer used in these experiments is HTK [10]. For the training of the HMM's, we have used totally 10143 recordings taken from 344 different speakers. Also context dependent triphones are trained according to the language model. Recognition results are evaluated using per cent of correct and accuracy.

$$\text{Per cent Correct} = \frac{N-D-S}{N} \times 100\% \quad (1)$$

$$\text{Per cent Accuracy} = \frac{N-D-S-I}{N} \times 100\% \quad (2)$$

where S, D and I are the substitution, deletion and insertion errors. Also, N is the total number of the labels. Optimum recognition parameters for each model are decided after recognition experiments on one of the recordings.

### 5.2 Speech Recognition Results

In each proposed model, most frequent 10K tokens are used as recognition lexicon. Also experiments are performed with both speaker independent and speaker dependent HMM's. Some of the problematic recordings are excluded from the experiment. The recognition results in terms of percent of correct and accuracy are given in Table 8.

From the recognition results, we can say that the results for each model are similar to each other. Although, maximum OOV rate is with word-based model, its recognition performance is slightly better than the other models. Also, speaker adaptation increases the recognition performance approximately 18% for all the models.

Table 8: Recognition Results

| Models | Correct(%) | Accuracy(%) |
|---|---|---|
| Word-based | 46.29 | 36.37 |
| Word-based (adapted) | 55.80 | 45.49 |
| Combined-2.5K | 45.38 | 35.12 |
| Combined-2.5K (adapted) | 55.93 | 44.38 |
| Combined-5K | 44.76 | 34.33 |
| Combined-5K (adapted) | 54.37 | 46.84 |

## 6. CONCLUSIONS

In this paper, we have searched for the appropriate base recognition units for LVCSR. In Turkish broadcast news dictation system, we investigated the combination of recognition units like words, stems, endings and morphemes. Using this combined model, we manage to increase coverage and decrease perplexity; however, recognition performance was lower than the classical word-based model. This can be explained with the number of smaller recognition units in the vocabulary. In the combined model, the lengths of the units are very different from each other and this unbalanced vocabulary entry situation generates a handicap from the point of the recognizer. As a consequence, we can say that although coverage is small and the perplexity is high compared to the other models, the best result is with the word-based model. The main drawback of our proposed models is the unbalanced length recognition units. A further research can be to work on balanced length recognition units. Also applying the word based model to more specific news domains like economics, politics and sports can be a solution to unlimited vocabulary size and this may result in better Turkish dictation systems.

## REFERENCES


[1] Hakkani-Tür, D., K. Oflazer and G. Tür, *Statistical Morphological Disambiguation for Agglutinative Languages*, Technical Report, Bilkent University, 2000.

[2] Siivola, V., M. Kurimo and K. Lagus, "Large Vocabulary Statistical Language Modeling for Continuous Speech Recognition in Finnish", *Proceedings of the 7th European Conference on Speech Technology and Communication, EUROSPEECH 2001*, Aalborg, Denmark, 2001.

[3] Kwon, O. W. and J. Park, "Korean Large Vocabulary Continuous Speech Recognition with Morpheme-based Recognition Units", *Speech Communication*, Vol. 39, pp. 287-300, January 2002.

[4] Dutağacı, H., *Statistical Language Models for Large Vocabulary Continuous Speech Recognition*, M.S. Thesis, Boğaziçi University, 2002.

[5] Oflazer, K., "Two-level Description of Turkish Morphology", *Literary and Linguistic Computing*, Vol. 9, No.2, 1994.

[6] Erguvanlı E. E., *The Function of Word Order in Turkish Grammar*, Ph.D. Thesis, University of California, Los Angeles, 1979.

[7] Çetinoğlu, Ö., *A Prolog Based Natural Language Infrastructure for Turkish*, M.S. Thesis, Boğaziçi University, 2001.

[8] Kanevsky *et al*., "Statistical Language Model for Inflected Languages", US patent No:5,835,888,1998, 1998.

[9] Mengusoglu, E. and O. Deroo, "Turkish LVCSR: Database Preparation and Language Modeling for an Agglunitative Language", *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-2001 Student Forum*, Salt Lake City, May 2001.

[10] Young S., D. Ollason, V. Valtchev and P. Woodland, *The HTK Book (for HTK Version 3.2)*, Entropic Cambridge Research Laboratory, March 2002.