STATISTICAL LANGUAGE MODELS FOR LARGE VOCABULARY TURKISH SPEECH RECOGNITION

by

Helin Dutağacı

B.S. in E.E., Boğaziçi University, 1999

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science
in
Electrical Electronics Engineering

Boğaziçi University
2002

STATISTICAL LANGUAGE MODELS FOR LARGE VOCABULARY TURKISH SPEECH RECOGNITION

APPROVED BY:

Assoc. Prof. M.Levent Arslan ………………………………………..
(Thesis Supervisor)

Prof. Bülent Sankur ………………………………………..

Assoc. Prof. Cem Say ………………………………………..

DATE OF APPROVAL: 20.06.2002

to my mother

ACKNOWLEDGMENTS

I would like to thank to my thesis supervisor Levent M. Arslan for his continuous help and guidance. I have always admired his intelligence and his engineering abilities. Beyond all, his patient and gentle attitude was very valuable.

I would like to thank to Bülent Sankur, who always encouraged me to continue my life as an engineer and as a researcher at the university. I owe every single line of this thesis to his personal support. I am grateful to him for introducing me the pleasure of engineering and research as well as the pleasure of exploring different views of life.

I would like to thank to Cem Say and his students, who developed the morphological parser I have used in this thesis. I would also like to thank to Cem Say for participating in my thesis committee.

I would like to thank to N. Engin Uzun for supplying us the necessary information about Turkish grammar and to Hakan Kalyoncu for supplying the archive of Cumhuriyet newspaper. I would once more like to thank to Levent M. Arslan and his colleagues at the GVZ Speech Technologies Company for developing the software I have used in this thesis.

I would like to thank to Cengiz Sucu and other voluntary workers at Gelenek Yayınevi and Dünya Yayıncılık for supplying me Turkish texts, all of which are much more valuable than their use in this thesis.

I am grateful to Gülay Dincel and Gökçe Erbil, who from the beginning to the end supported me. Without their trust I would have definitely given up.

I would also like to thank to my sisters Sevgi and Berçem and to friends who tolerated my bad temper and who willingly listened to my murmurs during the last months of this thesis. Finally I would like to thank to Tırmık, our pretty kitten, for the enjoyable minutes she shared with me.

ABSTRACT

STATISTICAL LANGUAGE MODELS FOR LARGE VOCABULARY TURKISH SPEECH RECOGNITION

In this thesis we have compared four statistical language models for large vocabulary Turkish speech recognition. Turkish is an agglutinative language and has a productive morphotactics. This property of Turkish results in a vocabulary explosion and misestimation of N-gram probabilities while designing speech recognition systems. The solution is to parse the words, in order to get smaller base units that are capable of covering the language with relatively small vocabulary size. Three different ways of decomposing words into base units are described: Morpheme-based model, stem-ending-based model and syllable-based model. These models with the word-based model are compared with respect to vocabulary size, text coverage, bigram perplexity and speech recognition performance. We have constructed a Turkish text corpus of size 10 million words, containing various texts collected from the Web. These texts have been parsed into their morphemes, stems, endings and syllables and statistics of these base units are estimated. Finally we have performed speech recognition experiments with models constructed with these base units.

ÖZET

GENİŞ DAĞARCIKLI TÜRKÇE KONUŞMA TANIMA İÇİN İSTATİSTİKSEL DİL MODELLERİ

Bu tezde, geniş dağarcıklı Türkçe konuşma tanıma için dört dil modeli karşılaştırılmıştır. Türkçe sondan eklemeli bir dildir ve morfolojik üretkenliği yüksektir. Türkçe'nin bu özelliği konuşma tanıma sistemleri tasarlarken, dağarcık patlamasına ve dilin istatistiklerinin yanlış kestirilmesine neden olmaktadır. Bu sorun, sözcükleri bölerek, dili küçük dağarcıklarla kapsama yetisine sahip daha kısa birimler elde ederek çözülebilir. Bu tezde sözcükleri temel birimlerine bölmek için üç yol anlatılmıştır: Biçimbirim tabanlı model, kök ve köksonrası tabanlı model ve hece tabanlı model. Bu modeller, kelime tabanlı modelle birlikte, dağarcık büyüklüklerine, metin kapsama oranlarına, ikili istatistiklerine ve konuşma tanıma performanslarına göre karşılaştırılmıştır. Web'den toplanmış çeşitli metinler kullanılarak 10 milyon kelime büyüklüğünde bir metin veri tabanı oluşturulmuştur. Bu metinler biçimbirimlerine, kök ve köksonralarına ve hecelerine ayrıştırılmış ve bu temel birimlerin istatistikleri kestirilmiştir. Daha sonra bu temel birimler kullanılarak oluşturulan modellerle konuşma tanıma deneyleri gerçekleştirilmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

$\mathbf{A}$	Acoustic signal
C	Counting function
$\mathbf{E}$	Expected value
H	Entropy
L	Alphabet size of X
LP	Logprob
N	Order of N-gram models
N_e	Number of erroneously recognized words
N_t	Number of words to be recognized
n	Number of words in $\mathbf{W}$
P	Probability
P_{LM}	Estimated probability distribution
PP	Perplexity
$p_i, i = 0,...,L-1$	Probability distribution of X
$\mathbf{W}$	String of words
$\hat{\mathbf{W}}$	String of words selected by the recognizer
w_1^{k-1}	String of words preceding w_k
$w_k, k = 1,...,n$	Words in $\mathbf{W}$
X	Random variable
γ	Vocabulary
HMM	Hidden Markov Model
WER	Word Error Rate
PER	Phoneme Error Rate

1. INTRODUCTION

1.1. Language Modeling

Speech recognition is performed with the analysis of two information sources: Acoustic signal and natural language. Actually there are three levels embedded in the speech signal: The acoustic level, the grammar and the semantic level. Human listeners are trained from the babyhood on to consider the three as a whole. How the three levels are organized in human listeners' acoustic and information processors and how much these levels contribute to the speech recognition action are interesting questions and their answer lie in researches in physiology, psychoacoustics, psychology, neural sciences, linguistics, etc.

Computer based automatic speech recognition systems, on the other hand, are not obliged to "understand" the spoken utterance; they just transcribe acoustic signal to symbols. Language understanding, if demanded by the application, is supposed to be handled by separate algorithms that accept the transcribed text as input. However, as far as speech recognition systems deal with elements of the language, they should have some knowledge of language. The question is what and how much they should know.

The language model is responsible for introducing the necessary "linguistic" information to a speech recognition system. Even the simplest recognizer should be fed with a language model in order to know what it should recognize. The isolated digit recognition system, for example, has a vocabulary of 10 words and behaves according to a very simple model telling that only one digit will be uttered at one time and the uttered digit may be any of the 10 possible digits.

It is straightforward to construct language models for small vocabulary speech recognition applications. When the vocabulary size increases, problems arise. If no constraints are defined for the next word to be recognized, the recognizer will have to find out the spoken word from a large set of candidates. Thus after every word or group of words, the recognizer should have an idea of which words will follow more probably.

This group of words after which a new word will be recognized is defined as “history”. The main function of a language model is to determine and formulate the dependencies of words to their histories. The dominant factors that strengthen the dependency of new words to the history are the syntactic rules of the language and the context in which the words are uttered. The syntactic rules determine short range dependencies; i.e. dependencies within a sentence.

It is ideal to extend the history to the beginning of a sentence, even to the beginning of the speech action. This ideal approach demands a great deal of computational efforts. There are a variety of approaches for utilizing simpler definitions of the history. N-gram models and finite state grammars, which will be considered later in this thesis, are the most popular ones.

1.2. The Vocabulary

The recognizer should have a list of language units, words, digits, sentences, etc., in order to compare their acoustic features with the features of the incoming speech signal and to select the one that resembles more. Vocabulary size; i.e. the number of words that the system is supposed to recognize, depends on the application. For a credit-card verification system, a vocabulary consisting of 10 digits and a few words for control actions is sufficient. Again a keyword spotting system has only the words in interest in its vocabulary.

Continuous speech recognition tasks and dictation systems, on the other hand, demand vocabularies with tens of thousands of words. Words that are not present in the vocabulary and that are introduced to the recognizer result in recognition errors. These words are defined to be “out of vocabulary” words. Selection of the appropriate vocabulary is critical for reducing recognition errors arising from the out of vocabulary words.

An effort to use a dictionary of all words of a specific language is a hopeless attempt. The number of words in the Oxford English Dictionary is about 500,000 excluding technical and scientific terms. With technical and scientific terms this number approaches to one million. Dictionary size of German is 185,000 and French has a vocabulary size of fewer than 100,000 (McCrum *et al.*, 1992). The Turkish dictionary published by Turkish Language Association contains about 70,000 words. These figures do not count for word inflections and derivations, proper nouns and technical and scientific terms. Turkish version of encyclopedia AnaBritannica contains about 85,000 items. The vocabulary increase arising from the morphological productivity is the subject of this thesis, and will be discussed in detail. One can produce thousands of different words, which sound grammatically correct, from a single Turkish stem.

The optimum vocabulary can be selected by data driven methods. The most frequent words are selected from a large text corpus as the items of vocabulary. A few thousands of these words are usually able to cover more than 80 per cent of the text data. After a certain vocabulary size, adding new words to the vocabulary results in very little increase in the percentage of coverage.

1.3. Selection of Base Units

For the "acoustic point of view", it is preferred to work on units that are as long as possible. Longer units are acoustically less confusable; choosing one from a set of sentences is easier compared to trying to decode words of a sentence one by one. On the other hand, it is too difficult to supply groups of words as vocabulary items to the acoustic processor since such a vocabulary would be drastically large. The convention is to consider words as the basic units of the language and as the items of the vocabulary. This choice is the optimum solution for the acoustic, linguistic and computational constraints of speech recognition tasks of English.

Leaving their detailed analysis to following chapters we can define a set of criteria for the choice of base units of a language model; i.e. the items of a vocabulary:

- The units should be acoustically distinguishable.
- The units should be able to cover the language with moderate sizes of vocabulary.
- The units should carry information of the semantic content of the language; i.e. they should possess meaning.
- The units should enable detection of word boundaries.

- The units should not be too sensitive to domain. On the other hand they should carry domain specific information.

1.4. Modeling Turkish

The theoretical background of language modeling for speech recognition has been developed via the research on English. Although many concepts of this background may be taken as universal, different frameworks should be constructed for other languages. Language modeling, especially statistical language modeling primarily relies on text data. Therefore orthographic variations of each language should be taken into account. For example Swedish compound nouns are not separated by a space while similar compound nominals in English are written as two words (Carter *et al.*, 1996). In Japanese there are even no delimiters between words, and segmenting a text into "words" is not a straightforward task (Ando and Lee, 2000). The situation is similar for other Asian languages such as Chinese and Korean.

In general, language is seen as a dynamic concatenation of static building blocks, i.e. static words forming sentences of infinite variability. For agglutinative languages this is not the case; words also have a dynamic character. It is theoretically and practically possible to construct a dictionary of all legal morphological forms of English words, but it is practically impossible for Turkish. Oflazer notes that verbs in Turkish have 40,000 forms not counting derivational suffixes (Jurafsky and Martin, 2000).

Agglutinative character of Turkish will cause a vocabulary explosion if words are selected as base units. This thesis is an attempt to decide on appropriate base units that will fit the criteria we have mentioned.

2. STATISTICAL LANGUAGE MODELING

Noam Chomsky finds the notion of “probability of a sentence” useless (Chomsky, 1969). Probabilistic approaches fail when unexpected events occur. Their power lies on the fact that unexpected events rarely occur. Probabilistic methods would be successful 99 per cent of the time if the statistics were correctly estimated. If they claim to obtain 100 per cent success then we will not be talking about a stochastic system, rather it should be defined as deterministic.

Neither natural languages nor acoustic production of speech are stochastic processes. The underlying mechanisms, rules, laws, etc. of their occurrence are too complex, and exact explanations of these mechanisms are still hidden from human experts. Additionally, a specific application usually does not need to employ all rules of those processes, and may want to extract only the relevant parts of the evidence. The application will then be overloaded if it tries to obtain these relevant parts through a vast amount of rules. Therefore probabilistic approaches are adopted as pragmatic solutions to problems arising from complex phenomena such as natural language.

Today the techniques used in speech recognition are dominated by statistical approaches. The underlying stochastic processes of both speech production and natural language are assumed to be Markov models.

2.1. Formulation of the History

2.1.1. Roles of Acoustic and Language Models

I will summarize the mathematical formulation of the speech recognition design given by Jelinek (Jelinek, 1997) in this section. The question of a speech recognizer is as follows: Given an acoustic signal $\mathbf{A}$, which were the words the speaker uttered?

$\mathbf{W}$ is a string of n words. These words belong to a vocabulary γ.

$$\mathbf{W} = w_1, w_2, \ldots, w_n \qquad w_i \in \gamma \tag{2.1}$$

$P(\mathbf{W}/\mathbf{A})$ denotes the probability that the speaker uttered the word string $\mathbf{W}$, given the speech signal $\mathbf{A}$ is observed. The recognizer chooses a word string $\hat{\mathbf{W}}$ which maximizes this conditional probability, such that

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W}/\mathbf{A}) \tag{2.2}$$

Using Bayes' formula we can rewrite the conditional probability to be maximized:

$$P(\mathbf{W}/\mathbf{A}) = \frac{P(\mathbf{W})P(\mathbf{A}/\mathbf{W})}{P(\mathbf{A})} \tag{2.3}$$

where $P(\mathbf{W})$ is the a priori probability of $\mathbf{W}$, $P(\mathbf{A}/\mathbf{W})$ is the probability of acoustic signal $\mathbf{A}$ will be produced when the speaker intend to utter $\mathbf{W}$. Finally $P(\mathbf{A})$ is the probability that $\mathbf{A}$ will be observed. Since $\mathbf{A}$ is given, and $P(\mathbf{A})$ is the same for all word strings in question, we can discard $P(\mathbf{A})$ and rewrite (2.2) as

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W})P(\mathbf{A}/\mathbf{W}) \tag{2.4}$$

So the word string $\hat{\mathbf{W}}$ that maximizes the product of $P(\mathbf{W})$ and $P(\mathbf{A}/\mathbf{W})$ is recognized by the speech recognizer system.

$P(\mathbf{A}/\mathbf{W})$ is calculated by the acoustic processor. Estimating $P(\mathbf{W})$ is the responsibility of the language model.

2.1.2. Probability of a Sentence

$P(\mathbf{W})$ defined above can be viewed as "the probability of a sentence", the notion Chomsky had found useless. Actually we need this notion in order to select the string of words which maximizes $P(\mathbf{W}/\mathbf{A})$.

Probability of a word string can be formulated as follows:

$$P(\mathbf{W}) = P(w_1, w_2, ..., w_n) \tag{2.5}$$

Using the chain rule:

$$P(\mathbf{W}) = P(w_1)\,P(w_2 / w_1)\,P(w_3 / w_1^2), ..., P(w_n / w_1^{n-1}) \tag{2.6}$$

$$P(\mathbf{W}) = \prod_{k=1}^{n} P(w_k / w_1^{k-1}) \tag{2.7}$$

where w_1^{k-1} is the sequence of words preceding w_k. It is the history of w_k.

If there were no dependencies between the words; i.e. occurrence of a word did not depend on the previous word sequence w_1^{k-1} then we could compute the probability of $\mathbf{W}$ as

$$P(\mathbf{W}) = P(w_1)\,P(w_2)\,P(w_3), ..., P(w_n) \tag{2.8}$$

where $P(w_k)$ is the probability of the occurrence of the word w_k. The assumption that each word is uttered independently employs a “unigram model”. The formulation of the assumption is as follows:

$$P(w_k / w_1^{k-1}) = P(w_k) \tag{2.9}$$

This assumption is not valid for continuous speech. The occurrence of a word depends on its history by means of syntax and semantic context. The range of this dependency can go beyond to the beginning of a sentence or utterance, even to the beginning of a speech or conversation. Unfortunately there is no known way to formulate this dependency for every string of words in both linguistic and probabilistic terms.

The solution is to limit the range of dependency to N words and define the history of a word as the previous $N-1$ words. In other words the following assumption is made:

$$P(w_k / w_1^{k-1}) \approx P(w_k / w_{k-1}, w_{k-2}, \ldots, w_{k-N+1}) \tag{2.10}$$

The language models based on this assumption are called "N-gram models", and the word sequences $w_{k-N+1}, w_{k-N+2}, \ldots, w_k$ are called "N-grams".

So a bigram model assumes that

$$P(w_k / w_1^{k-1}) \approx P(w_k / w_{k-1}) \tag{2.11}$$

and a trigram model assumes

$$P(w_k / w_1^{k-1}) \approx P(w_k / w_{k-1}, w_{k-2}) \tag{2.12}$$

where w_{k-1} and w_{k-2} are the two words just before the word in question.

2.1.3. Estimation of N-gram Probabilities

Since there exists no authority or expert to tell us the N-gram probabilities, we have to estimate them from data. A text corpus is utilized to extract the statistics and the size and domain of this text corpus have importance. This text corpus is called as "training corpus".

N-gram probabilities are estimated by counting the occurrences of a particular N-gram in the text data and dividing this count to the number of occurrences of all N-grams that start with the same sequence of $N-1$ words; i.e.:

$$P(w_k / w_{k-1}, w_{k-2}, \ldots, w_{k-N+1}) = \frac{C(w_{k-N+1}, \ldots, w_{k-1}, w_k)}{\sum_w C(w_{k-N+1}, \ldots, w_{k-1}, w)} \tag{2.13}$$

So for the particular case of a bigram:

$$P(w_k / w_{k-1}) = \frac{C(w_{k-1}, w_k)}{\sum_w C(w_{k-1}, w)} \tag{2.14}$$

and for trigram:

$$P(w_k / w_{k-1}, w_{k-2}) = \frac{C(w_{k-2}, w_{k-1}, w_k)}{\sum_w C(w_{k-2}, w_{k-1}, w)} \tag{2.15}$$

where $C(w_{k-N+1},..., w_{k-1}, w_k)$ is the count of occurrences of N-gram, $w_{k-N+1},..., w_{k-1}, w_k$ in the training corpus.

Since the vocabulary size and consequently the number of N-grams are too large, estimation of N-gram probabilities by counting methods suffer from data sparseness. Even if it were possible to collect all the material written in a particular language as training corpus, some perfectly constructed N-grams would be missing from this corpus. Therefore this missing data should be interpolated.

On the other hand, the improvement of N-gram probability estimations become too little after some amount of training data. “One informal estimate from IBM shows that bigram models effectively saturate within several hundred million words, and trigram models are likely to saturate within a few billion words” (Rosenfeld, 2000).

2.1.4. Smoothing

However the training text corpus is large, there may appear N-grams that never appear in the training data. In order to include these unseen N-grams to the language model, small probabilities are assigned to them. The probabilities are determined through smoothing methods.

In this thesis, a very simple minded smoothing is applied to the morpheme-based model, which will be described in the following sections. Other models were not smoothed although they should be.

2.2. Entropy and Perplexity

Statistics are estimated in order to extract the information content of particular data. An N-gram language model can be viewed as a statistical model trying to predict the next word given $N-1$ previous words. This approach is borrowed from information theory, which was constructed by Shannon.

The information content of a random variable X is measured by the concept of entropy (Shannon, 1948). Entropy can be viewed as the average uncertainty about occurrence of an event. It is formulated as follows:

$$H(X) = -\mathbf{E}[\log P\{X\}] \tag{2.16}$$

So, when the probability distribution of X with finite alphabet of size L is $\{p_0, p_2, ..., p_{L-1}\}$ then entropy can be written as

$$H(p_0, p_2, ..., p_{L-1}) = -\sum_{i=0}^{L-1} p_i \log p_i \tag{2.17}$$

The main responsibility of a statistical language model is to estimate the "true" probability distribution of the language, $P\{X\}$. If the distribution is misestimated then the entropy, in other words the uncertainty about the next event will be higher. This entropy is called the "cross entropy", which includes the true uncertainty plus the uncertainty added by the wrong estimation of probability distribution:

$$\text{cross}-\text{entropy}(P; P_{LM}) = -\sum_{X} P\{X\} \log P_{LM}\{X\} \tag{2.18}$$

where $P_{LM}\{X\}$ is the estimated probability distribution of a particular language model. Since $P\{X\}$ is unknown, another measure should be utilized to find the cross entropy. That measure is called "logprob (LP)" and is defined as

$$LP = \lim_{n\to\infty} -\frac{1}{n}\sum_{k=1}^{n} \log P_{LM}\left(w_k / w_1, \ldots, w_{k-1}\right) \tag{2.19}$$

where $\sum_{k=1}^{n} \log P_{LM}\left(w_k / w_1, \ldots, w_{k-1}\right)$ is the log probability of the long sequence $w_1, \ldots, w_n$ estimated by the language model. So instead of the ensemble average, time average is taken to estimate the cross entropy with an assumption that language is stationary and ergodic. Since language is neither stationary nor ergodic the logprob is an approximation to the cross entropy.

The logprob is estimated over an unseen text sequence, which is called the test text. If this text appeared in the training text, its probability would be high; consequently the cross entropy would be estimated lower than it should be.

The cross entropy is always higher then the entropy itself, and the quality of a language model is determined from how much the cross entropy gets closer to the entropy of the language. Entropy is also defined as a lower bound on the number of bits or the number of "yes/no" questions in order to encode or predict the next piece of information. Therefore the average number of choices can be formulated as

$$PP = 2^{LP} \tag{2.20}$$

where the quantity PP is called "perplexity".

From the recognizer's point of view, perplexity gives the average number of choices of the next word to be predicted. When perplexity is high, the recognizer's task becomes more difficult because of the large number of choices. This difficulty arises from the language itself (the entropy of the language) and the language model (additional uncertainty when the statistics are misestimated).

2.3. Selection of Base Units

The difficulty of the recognizer’s task does not depend solely on perplexity since perplexity is not a measure of uncertainty about the acoustic speech signal (Jelinek, 1997). The acoustic realizations of the words in question should be as different as possible for the recognizer to decide on one easily. This desired differentiability becomes higher when the words are long; shorter words cause more acoustic confusion.

The main building blocks of the language are considered as words. The continuous speech recognizer’s concern is to predict the next spoken word through maximization of likelihood among the words in the vocabulary. The recognizer design could be based on language units as “sequences of words” instead of single words, and this choice would reduce acoustic confusability. On the other hand the vocabulary size and the number of statistical parameters would go beyond the limits of current computational power. Additionally and consequently the training corpus needed to satisfactorily estimate the statistics of these language units should be huge.

If the recognizer chose letters (phonemes as their acoustic realizations) as base units, the language model would be too simple and the perplexity would be much smaller than of the model with words as base units. However recognition of utterances phoneme by phoneme would cause high errors since the units are acoustically very confusable; inter-phoneme interactions, which are very powerful, would be ignored.

When there are no obvious words in the language (as in Asian languages) or when the vocabulary size; i.e. number of distinct words is already large (as in agglutinative or compounding languages) then smaller base units other than words should be adopted.

Short units tend to be more frequent taking the advantage of high likelihood arising from the language model. Conversely rare words are usually long, and can be easily recognized. However for this scheme to work well, the base units chosen should be coherent to the semantic, syntactic and morphological properties of the language.

As a result, given the language and the domain the language model should cover, there is a trade-off between perplexity and acoustic confusability while selecting the base units. This trade-off seems to be solved empirically; i.e. relying on the word error rates of the recognizer.

3. TURKISH MORPHOLOGY AND MORPHOLOGICAL PARSER

3.1. Turkish Morphology

There are significant research efforts about computational linguistics of Turkish. Most of the work was dedicated to natural language processing applications; there are very few efforts for utilizing them for speech recognition task (Çarkı *et al.*, 2000), (Mengüşoğlu and Dereo, 2001).

Turkish has a complex morphotactics (Oflazer, 1994). Turkish is a member of Altaic family languages and has an agglutinative structure. Thousands of new words can be formed from a single stem by addition of suffixes to it one after another. The detailed morphotactics of Turkish will not be given here; rather issues related to speech recognition task will be discussed. The discussion aims to give a motivation for the choice of morphemes instead of words as base units of a Turkish speech recognizer.

Oflazer gives a two-level description of Turkish morphology. A two-level description of a language employs two representations of a word: Lexical representation and surface representation. The lexical representation gives the sequence of morphemes in a word while the surface representation is the actual spelling of the word. The mapping of lexical level to surface level is determined by a set of phonetic rules of suffixation. Oflazer gives 22 such rules.

The syntax of a language determines the ordering of words in a sentence, whereas morphotactics defines the rules of concatenation of morphemes. The syntax of English strongly determines the n-gram statistics. In addition, functional words such as prepositions, possessives, auxilary verbs, articles, etc. are the most frequent words of English. Most of the time, these functional words correspond to morphemes of Turkish.

For nominal and verbal Turkish words there are two different morphotactics, and there are transitions from one to another. A nominal structure can be converted to a verbal structure and vice versa. The morphotactics for nominal words is simpler than verbal words. Some of the inflectional suffixes added to a nominal word are given in Table 3.1. They are the plural, possessive and case suffixes.

Table 3.1. Some of the nominal inflectional suffixes

Lexical definition	Surface definition	English correspondence	Example	English translation
Plural	+lAr	+s	ev+ler	houses
Possessive, 1st person singular	+Hm	my	ev+im	my house
Possessive, 2nd person singular	+Hn	your	ev+in	your house
Possessive, 3rd person singular	+sH	his/her/its	ev+i	his/her/its house
Possessive, 1st person plural	+HmHz	our	ev+imiz	our house
Possessive, 2nd person plural	+HnHz	your	ev+iniz	your house
Possessive, 3rd person plural	+lArH	their	ev+leri	their house
Dative	+A	to	ev+e	to house
Locative	+DA	At/in	ev+de	at house
Ablative	+DAn	From/of	ev+den	from house
Genitive	+nHn	+’s/of	ev+in kapısı	door of the house

The verb “to be” is implemented by the suffixes added to the nouns. The subject of the nominal verb is also indicated by suffixes. As examples to nominal verbs following words can be given:

•	genc+im	I am young.
•	genç+ti+m	I was young.
•	genç+ti+k	We were young.
•	genç+miş+sin	(I heard that) you are young.
•	genç+siniz+dir	You are (definitely) young.
•	genç+miş+cesine	as if she is young.
•	genç+se+k	if we are young

Finally, all these suffixes concatenate to a nominal stem in accordance with the nominal morphotactics to form more complex constructions such as:

•	ev+de+ki+ler+den+di+k	We were of those at the house.
•	ev+imiz+de+yse+ler	If they are at our house.

For verbal structures the morphotactics is very complex since causative, reflexive, passive cases, tense information, person, negation, adverbial cases, etc. of a verb can be expressed within a single word. Following examples can be given:

•	gör+ül+me+meli+ydi+k	We should not have been seen.
	gör	see
	gör+ül	be seen
	görül+me	not be seen
	görülme+meli	shall not be seen
	görülmemeli+ydi	should not be seen
	görülmemeliydi+k	we should not be seen
•	gör+ebil+se+ydi+niz	If you were able to see.
	gör	see
	gör+ebil	be able to see
	görebil+se	if were able to see
	görebilse+ydi	if were able to see
	görebilseydi+niz	if you were able to see

Verbal structures can be converted into nominal or adverbial structures such as:

gör+eme+dik+ler+imiz	: those we were not able to see
gör+düğ+ümüz+de	: when we see

There is not always a one-to-one correspondence between morphemes in Turkish and functional words in English. But it is occasionally the case that a Turkish word with suffixes corresponds to a group of English words. When we neglect the acoustic appropriateness, we can select the Turkish morphemes as base units of a speech recognizer.

3.2. Morphological Parser

3.2.1. The Original Morphological Parser

A Prolog-based morphological parser was developed at the Computer Engineering Department of Boğaziçi University (Çetinoğlu, 2001). The parser is based on Oflazer's finite state machines, with a number of changes. Figure 3.1 and Figure 3.2 show the morphotactics the parser is based on. The stems of the parser fall into two categories: Verbal stems and nominal stems. Figure 3.1 gives the finite state machine for nouns while Figure 3.2 illustrates the finite state machine for verbs.

The parsing algorithm is based on left-to-right root matching approach. The input word is analysed from left to right and its successive parts are matched with morphemes in the lexicon. If each substring of the word is matched with a morpheme, and if these substrings are in the "right" order; i.e.the ordering is legal with respect to the finite state machine, the word is parsed. Otherwise it is rejected. The operation of the algorithm can be found in (Çetinoğlu, 2001).

The parser needs two types of information: The lexicon and the transitions. The lexicon defines the morphemes (stems and suffixes) and their properties. The transitions define the grammatical rules of suffixing. The lexicon of the original parser was a small set of Turkish stems. The total number of the stems was 40.

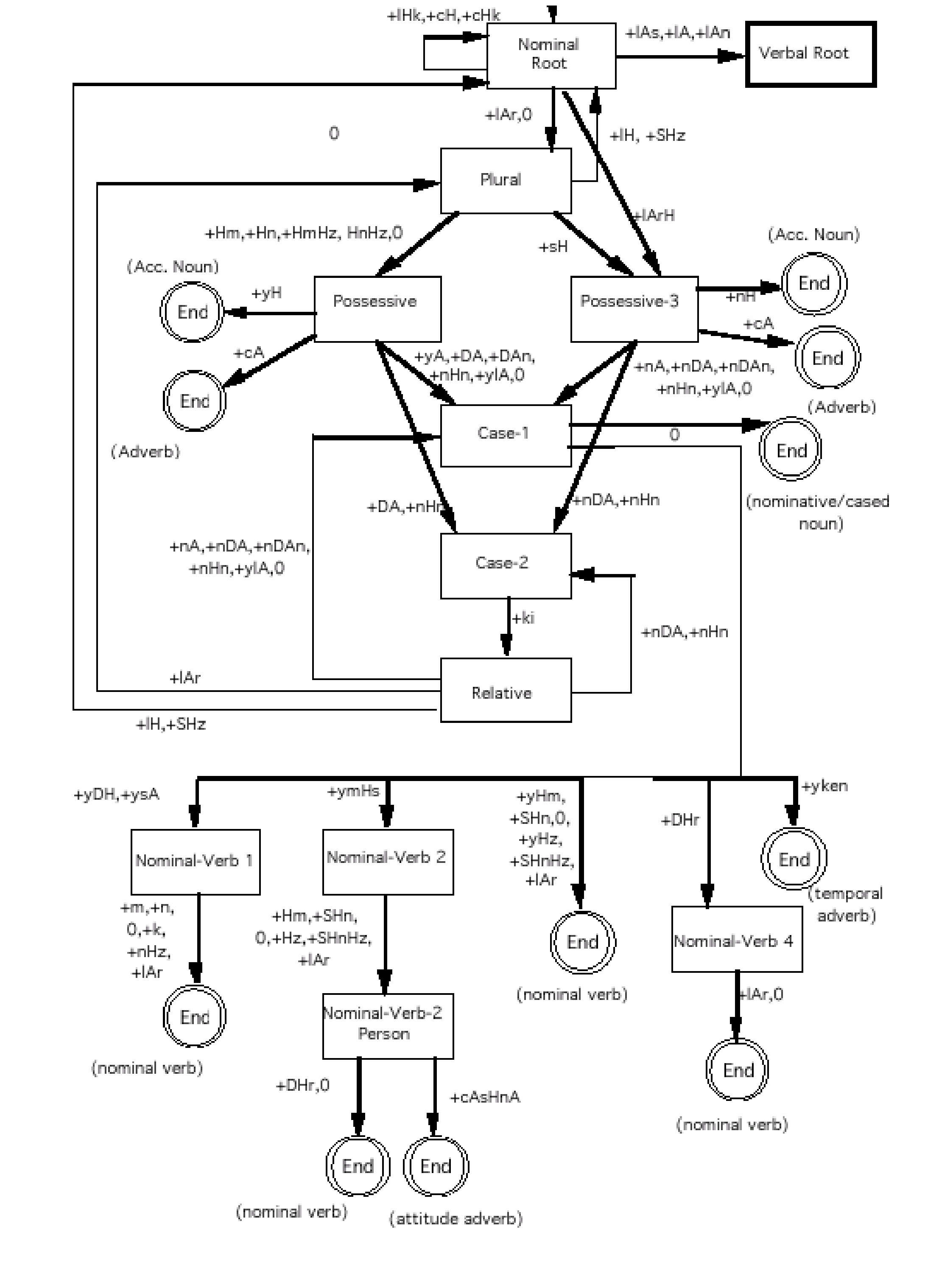

Figure 3.1. Finite state machine for nouns (Oflazer, 1994)

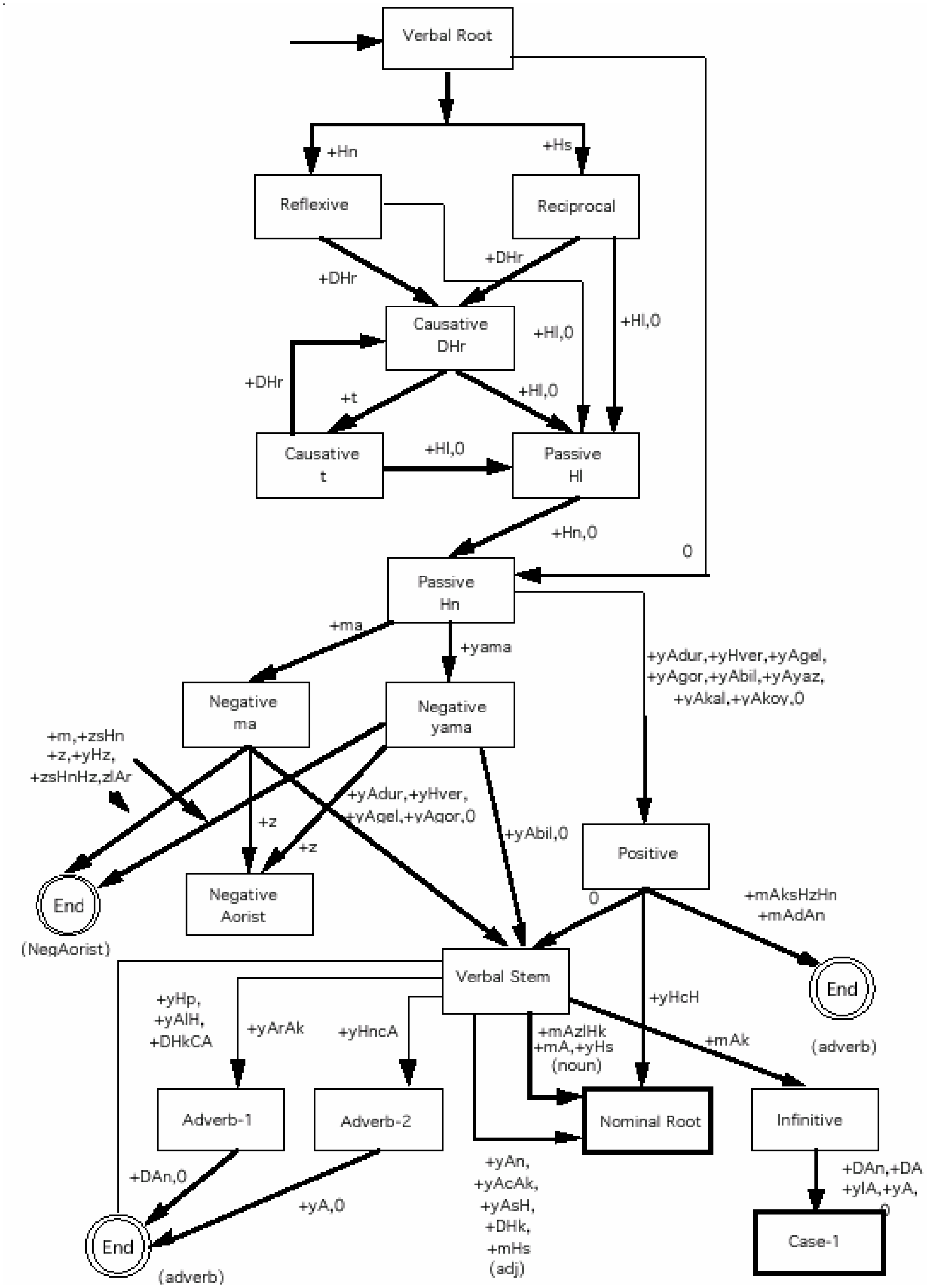

Figure 3.2. Finite state machine for verbs (Oflazer, 1994)

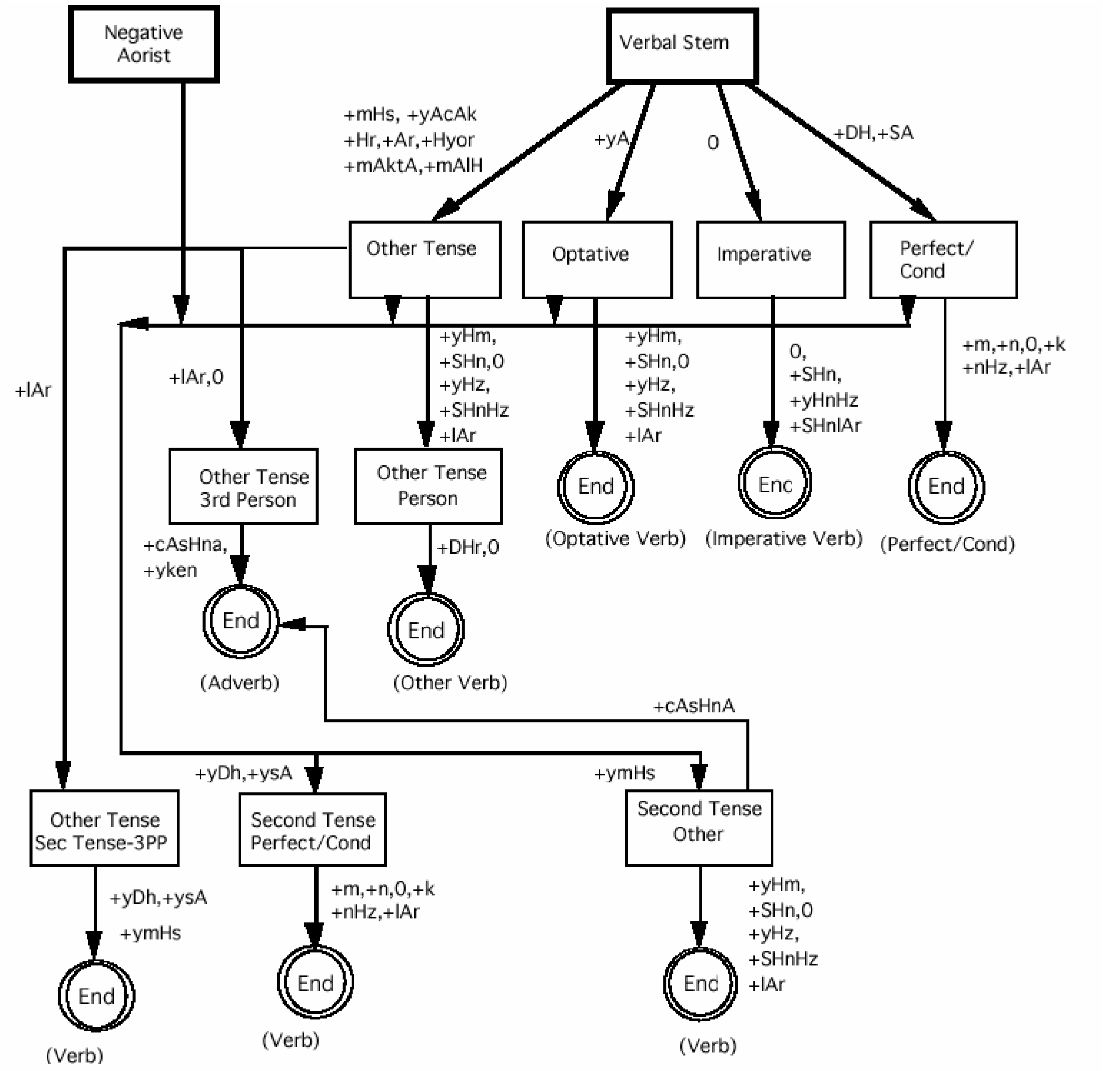

Figure 3.2. Finite state machine for verbs (continued) (Oflazer, 1994)

3.2.2. Modifications to the Parser

The parser had some problems and the main problem was the lack of a dictionary of Turkish stems. However, in order to parse texts, the parser needed a large lexicon of stems. A preliminary list of Turkish stems was integrated with the parser. Each stem was added to the Prolog source file (the lexicon) as a line describing its properties demanded by the parser. These properties are the character sequence of the stem, last letter, last vowel, category (whether verbal or nominal) and type (whether the surface form reflects an inflection).

If a stem was due to surface modifications (deletion or modification of the last letter, insertion of a new letter) with suffixation, the corresponding surface version of that stem was also added to the lexicon as a separate item. The last letter of a Turkish nominal stem, if that letter is a voiceless stop (ç, k, p, t), is in general modified to c, ğ, b or d when the stem is concatenated with a suffix beginning with a vowel. So for each of those stems we added another item to the lexicon with the same stem except that the last letter was modified to c, ğ, b or d.

The parser, integrated with the preliminary lexicon, was tested over a set of texts. Observing the parser results and undecomposed words, some erroneous and missing transitions describing inflectional suffixation were cured. In addition, the following refinements and corrections were done:

The last vowel of each morpheme is indicated since it determines the following suffixes according to the vowel harmony. But some stems, those introduced to Turkish from foreign languages, violated vowel harmony. For example the word “yar (lover)” is inflected as “yarim (my lover)”, instead of “yarım”. We did not have a complete list of those stems. Through examination of unparsed words in the test texts, we collected a number of those stems and defined their “last vowel” item according to the suffixes they take.

There are some special stems, of which the last consonant is doubled. For example the stem “hak” becomes “hakk+ım”, when concatenated with the suffix “ım”. Such stem modifications, which we came across in the test texts, were also introduced to the lexicon.

Observing the unparsed words, new stems were added to the lexicon. Also we have added some morphological derivations to the dictionary as single stems if the derivational suffix is added to few stems. Examples of such words are “toplum+sal”, “düş+ük”, “kadın+sı”, “düş+kün”, “ör+üntü”, “yün+ümsü”. We could add such derivational suffixes to the lexicon and to the finite state machine, but then we would overload the machine, since these suffixes are not added to every stem.

On the other hand, derivational suffixes that are added to large number of stems were added as new items to the lexicon and new transitions to the finite state machine. The derivational suffixes added to the parser are “+lA”, “Hm”, “GH”, “lAş”; examples of word forms are “ispat+la”, “bak+ım”, “sil+gi”, “güzel+leş”.

The decision whether to add a derivational suffix to the morphotactics is taken by the list of derivational suffixes and of stems that take these suffixes provided by N. Engin Uzun at Language, History and Geography Faculty at Ankara University. When the number of the stems taking that derivational suffix is large, it was added to the morphotactics.

N. Engin Uzun also provided us a list of Turkish stems of which the last vowel drops with suffixation. The following examples can be given for this type of stems:

- burun+un : burn+un
- fikir+i : fikr+i

For each of these stems, we added the corresponding modified stem as separate item to the lexicon.

After these refinements the parser could handle more than 90 per cent of the words in input texts. Most of the undecomposed words correspond to proper nouns and terms. There are some Turkish structures that could not be recognized by the parser, such as morphological forms of compound nouns. For example although “rengeyiği” is in the lexicon, the parser could not decompose the plural form “rengeyikleri”, since the morphotactics of compound nouns is a little bit different (Oflazer, 1994).

We should note that the morphological parser we have started to modify was a prototype of Çetinoğlu’s. Such problems with compound nouns were solved in (Çetinoğlu, 2001).

The final lexicon contained 29,652 items (29541 stems and 111 suffixes) with the modified versions of the stems. Lexical representations of the suffixes are given in

Appendix A. Since such a vocabulary could handle more than 90 per cent of the words in text inputs, we guarantee 90 per cent coverage of training text with a language model utilizing base units as morphemes.

3.2.3. Morphological Ambiguity

Turkish words are subject to morphological ambiguities, which can only be resolved in context. The morphological structure, hence the meaning of a single word can be multiple. For example the word “bakanının” can be parsed as concatenation of following morphemes and interpreted with the following meanings:

• bak+an+ı+nın	: of the one who look after him
• bak+an+ın+ın	: the one who look after you
• bakan+ı+nın	: of his (or its) minister
• bakan+ın+ın	: of your minister

For morphological disambiguation in Turkish, statistical models are proposed (Hakkani-Tür, *et al.*, 2000). They have reached a 95.07 per cent of accuracy in ambiguity resolution. They have used a trigram model of the inflectional groups of the words. The models, they have constructed, utilize the morphological structure of the previous words of the word in question.

The parser gives multiple parses for a single word. We did not attempt to disambiguate the parses; instead we selected randomly one of them. Actually, this is not an optimum solution from the speech recognition point of view. The solution could be to select the parse which gave minimum number of morphemes, hence which gave longer morphemes.

3.3. Turkish Syllables

Turkish spelling is assumed to be phonetic; i.e. all letters in a word are pronounced. Letters usually correspond to phonemes. There are a few exceptions. For example "geleceğim" is pronounced as "gelicem". The letter "ğ" is not usually pronounced, but causes an expansion of the vowel preceding it.

The rules that describe the syllable boundaries of Turkish words are simple. Each Turkish syllable can be uttered separately and this separate utterance resembles acoustically the same syllable uttered within a word. The group of phonemes (a vowel and consonants around them) in a word that are uttered immediately constitutes one syllable. Since Turkish is phonetic the orthographic rules for determining syllable boundaries are in accordance.

A Turkish syllable has exactly one vowel. This is an important property since every syllable in a word can be pronounced separately. The number of letters present in a syllable cannot be more than four. A syllable cannot start with two consonants. With these rules we can classify Turkish syllables as follows:

- Syllables with only one vowel, V ("a", "e")
- Syllables with one vowel in the beginning and one consonant at the end, VC ("ak", "ey")
- Syllables with one consonant in the beginning and one vowel at the end, CV ("pi", "bo")
- Syllables with a vowel in the middle and two consonants, CVC ("tik", "yor")
- Syllables with a vowel in the beginning and two consonants, VCC ("ark", "üst")
- Syllables with a consonant in the beginning, a vowel in the middle, and two consonants at the end, CVCC ("türk", "sarp")

Since Turkish has eight vowels and 21 consonants (the sound "ğ" is not a consonant in fact, however assuming it a consonant is more appropriate for the orthographic rules), the upper bound for the number of Turkish syllables is calculated to be 81,488. However, there are constraints for syllables of type 5 and 6. Not all sequences of two consonants can be the last two phonemes of a syllable. For example, "üst" and "kork" are syllables while "atg" and "inb" are not. N. Engin Uzun supplied us the consonant sequences that can be the last two phonemes of a syllable. There are 56 such sequences. With this constraint, the number of Turkish syllables reduces to 13,728.

We have implemented a syllable parser in MATLAB. The parser sets an initial number of syllables by counting the vowels in the input word. If the word starts with two consonants, which is the case for the words introduced to Turkish from foreign languages like “spor”, “tren”, it increments the number of syllables by one. It assigns the letter sequences to a syllable until it reaches a consonant preceding a vowel, then it assigns a new syllable starting with that consonant. When the number of syllables it parsed is equal to the predetermined number of syllables minus one, it assigns the last unprocessed letters to the last syllable.

The syllable parser assumes that the input words are Turkish words; i.e. it does not check whether the word is Turkish or not. Therefore it produces garbage when the input is not Turkish and the input word cannot be parsed to Turkish like syllables.

4. LANGUAGE MODELS PROPOSED

In this thesis, we are concerned with modeling single, isolated Turkish words for dictation tasks. This concern can also be viewed as a starting point for the solution of continuous speech recognition problem. The cues for the appropriateness of the base units (words, morphemes, stems, endings and syllables) for large vocabulary continuous speech recognition applications are present in the statistics we have derived from text corpus. However, we designed recognition systems for isolated Turkish words.

Our system can be illustrated as in Figure 4.1. All words are isolated from each other; no interword dependencies are taken into account. This system is inappropriate for continuous speech recognition, but can be used for dictation tasks where words are uttered with long silence intervals between them.

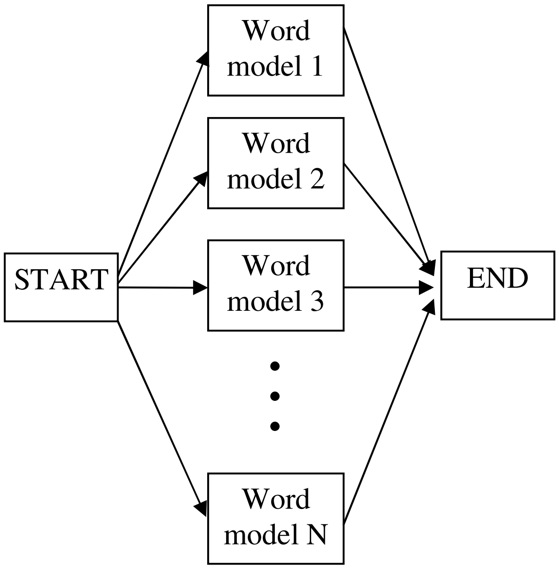

Figure 4.1. Word model

In this chapter, word based, morpheme based, stem-ending based and syllable-based models will be briefly described and motivation for the use of the latter three models will be given.

4.1. Word-based Model

Word-based system can be illustrated as in Figure 4.2. The problem reduces to simple isolated word recognition.

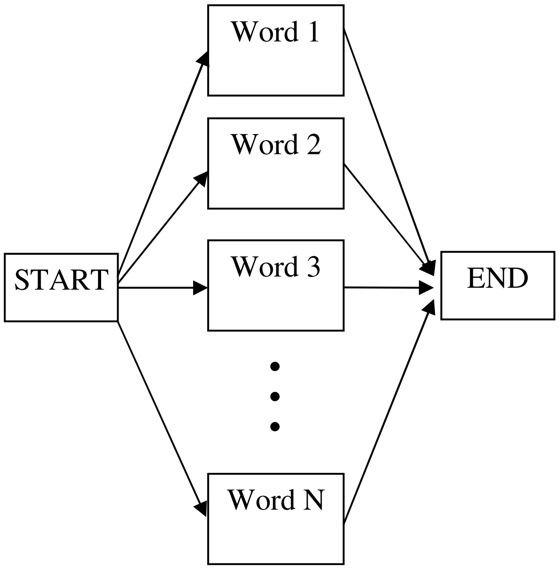

Figure 4.2. Word-based model

We have stated that longer language units would result in higher performance of the acoustic processor. Consider using Turkish words as base units of the recognizer and employing a trigram language model. In this case the dictionary size would be about a few hundred thousands of words to cover Turkish texts sufficiently, as we will show in Chapter 5. Still we would have out of vocabulary words, which are constructed through legal morphological rules and which sound more or less “typical”. We would need a huge training corpus to estimate the trigram probabilities.

The morphological productivity of Turkish makes it difficult to construct a word-based language model. Another important property of Turkish, the free word order, makes it even more difficult to benefit from the power of a N-gram language model. There are few constraints in the word order of Turkish. The word order is determined by semantic variations rather than strict syntactic rules. For example, all the following sentences, which contain the same words, are grammatically true:

- Çocuk bugün okula geldi. (The child came to school today.)
- Çocuk okula bugün geldi.
- Okula bugün çocuk geldi.
- Okula çocuk bugün geldi.
- Bugün çocuk okula geldi.
- Bugün okula çocuk geldi.

In the English sentence, the group of words, “child came to school” is strictly in that order. The concept of “word focus” determines the semantic interpretation of Turkish sentences. The emphasized word is the one that just comes before the word.

Converting the system to a continuous speech recognition system is straightforward. As words are used for the base units; N-gram models can be employed. However, as we stated, the word-base model demands a huge vocabulary. The word-based bigram and trigram language models yield high perplexities due to the free word order and very large vocabulary.

4.2. Morpheme-based Model

The morpheme-based language model utilizes morphemes as base units of the recognizer. The morpheme-based lattice is illustrated as in Figure 4.3. Words are modeled as a stem followed by suffixes in accordance with the morphotactics of Turkish and with the spelling rules. The transitions between morphemes are weighted with the bigram probabilities. The difference of the network from a bigram model is that all the transitions between morphemes are completely rule-based. The network has a weighted finite-state character.

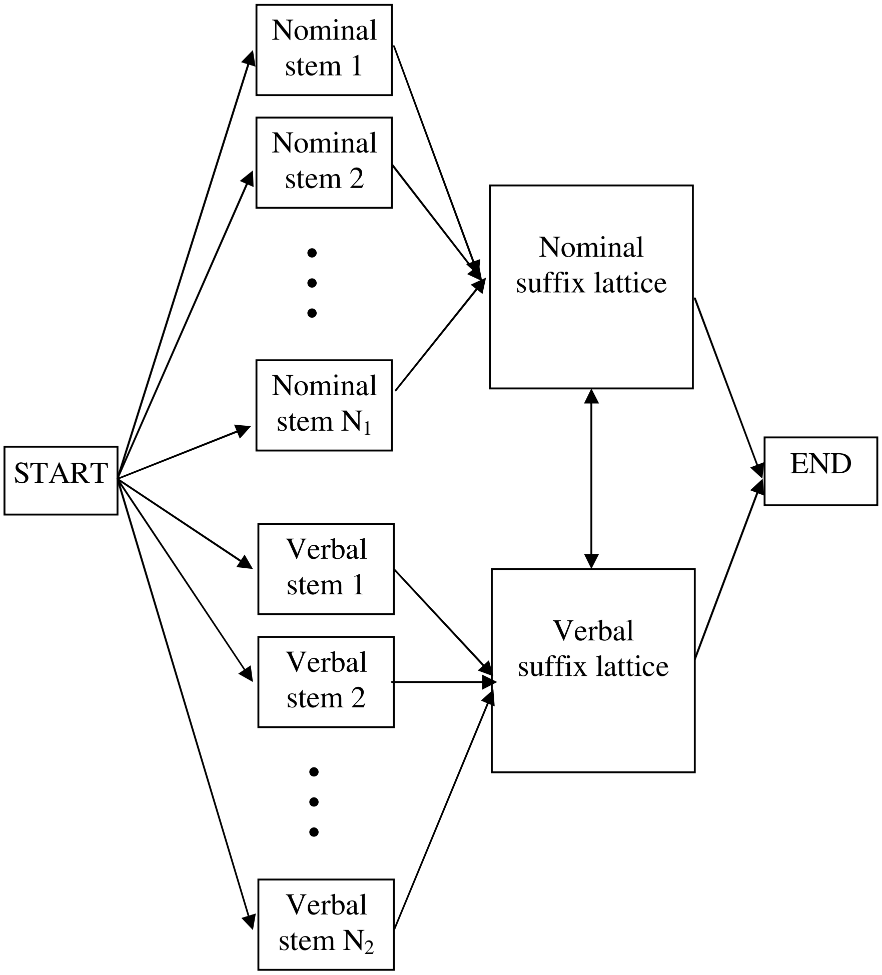

Figure 4.3. Morpheme-based word model

Stems fall into two groups. There are homonyms; a stem can be both verbal and nominal. These stems are linked with both verbal and nominal suffix lattices. There are transitions between verbal and nominal suffix lattices. In the lattice, units are represented with their surface realizations. The suffix lattices are constructed using the phonetic rules; i.e. all possible suffix sequences obtained through the network obey the phonetic rules.

The links between the stems and suffix lattices are also constructed using the phonetic rules. The suffixes that may follow a particular stem are determined by the last vowel and last phoneme of the stem. Once this mapping is defined, new stems can be added to the lattice automatically.

There appeared one problem in constructing the lattices in accordance with the phonetic rules: Some morphemes contained only one consonant. Since in the lattice the next unit is only dependent on the previous one, there appeared an ambiguity about the vowel harmony. For example, the stem “anne” takes the first possessive suffix as “anne+m”, and since the following suffixes are linked only to “+m”, there remains no reference vowel to decide on which surface realization should be selected. If genitive suffix “+Hn” were to be added, the word could be any of the forms as “anne+m+in”, “anne+m+ın”, “anne+m+un”, “anne+m+ün”, three of which are not legal. Each unit should be linked to only one surface realization of a particular lexical suffix. This cannot be done for suffixes that have no vowels.

The suffixes that contain only one consonant are as follows:

• +m	possessive, 1st person singular (added to nominal words)
• +n	possessive, 2nd person singular (added to nominal words)
• +n	reflexive (added to verbal words)
• +l	passive (added to verbal words)
• +t	causative (added to verbal words)
• +r	present tense (added to verbal words)
• +z	negative aorist (added to negation suffixes, “+mA”, “+yAmA”)

This problem is solved through merging such suffixes with all the previous morphemes (suffix or stem) they can follow. Then these merged suffixes are considered as single units. This solution resulted in a vocabulary growth; the number of stems and suffixes increased. For each nominal stem, we should add two more stems, and for each verbal stem, we should add four more stems to the vocabulary if the stem ends with a vowel.

There are a total of 559 units in the verbal and nominal suffix lattices including the merged ones. There appears a total of 10318 bigrams within the verbal and nominal suffix lattices. Each nominal stem is linked to 42 suffixes and verbal stems are linked to 65 suffixes. The weights between suffixes are determined by the bigram probabilities derived from training text corpus. If a bigram defined in the lattice does not exist in the training text, a small value (smaller than all the other bigrams that share the first morpheme) is assigned to that transition.

We have also constructed another lattice, where the bigrams that did not appear in the training data are not allowed. We will call this model Morpheme1 and the model described in the previous paragraph Morpheme2.

4.3. Stem-ending-based Model

The stem-ending based model is also a morphological model. It can be viewed as a solution to the short units of morpheme-based model and to the problem of very large vocabulary caused by the word based model. The base units of this model are stems and the group of suffixes following them, which are called "endings". Examples for words that are parsed to their stems and endings are as follows:

• gör+emediler	(they could not see)
• gör+meseydiniz	(if you had not seen)
• ev+imizdekiler	(those at our house)

Then the base units will be "gör", "ev", "emediler", "meseydiniz", "imizdekiler".

Stem-ending based modeling is proposed for agglutinative languages by (Kanevsky *et al.*, 1998). (Mengüşoğlu *et al.*, 2001) also proposed the stem-ending approach for modeling Turkish.

The stem-ending model needs an ending dictionary. There are two possible approaches to obtain an ending dictionary. Either the morphotactics can be employed to generate all possible endings that can follow nominal and verbal stems, and the dictionary built up with this method can also be used as a look-up table for morphological analysis.

Or we can rely on data and through morphological decomposition of a large text corpus we can construct a dictionary from the endings that appeared in the text.

One may not have all the endings of the language with the second approach if the text corpus is not large enough. However, if we let all verbal and nominal stems to take all the corresponding endings we would overload the system. Most of the endings do not follow certain stems, although this would be grammatically true, because of semantic properties of those stems. For example stems that are used mostly as adjectives tend to take different endings than stems used as nouns, although the two are nominal stems. There are cases where nominal stems will sound strange, and not likely to be used in Turkish, if they take some of the nominal endings.

The stem-ending based word model can be illustrated in Figure 4.4. Not all stems are connected to all endings. The connections are present if the corresponding bigrams appeared in the text corpus.

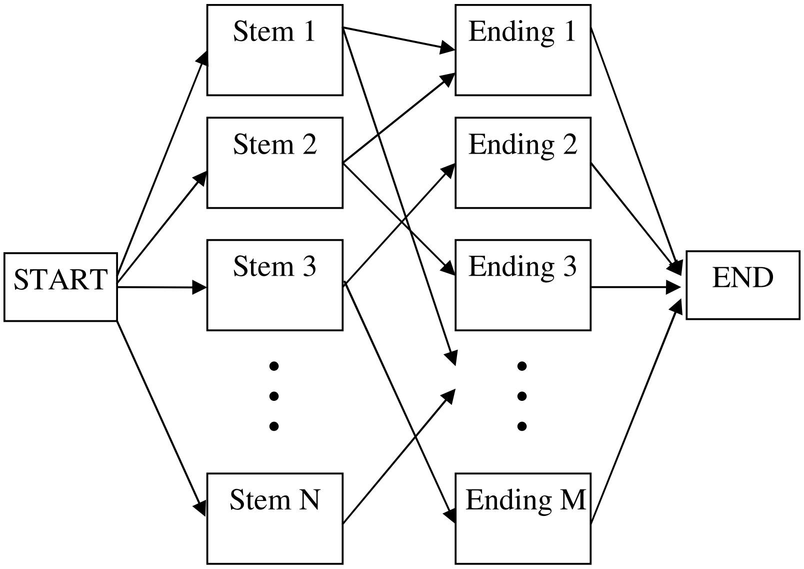

Figure 4.4. Stem-ending-based word model

4.4. Syllable-based Model

Syllable-based model is the simplest word model proposed as alternative to the word-based model. Once the training corpus is parsed into its syllables, the N-gram models can be employed as if syllables are words. The syllable-based word model is as in Figure 4.5.

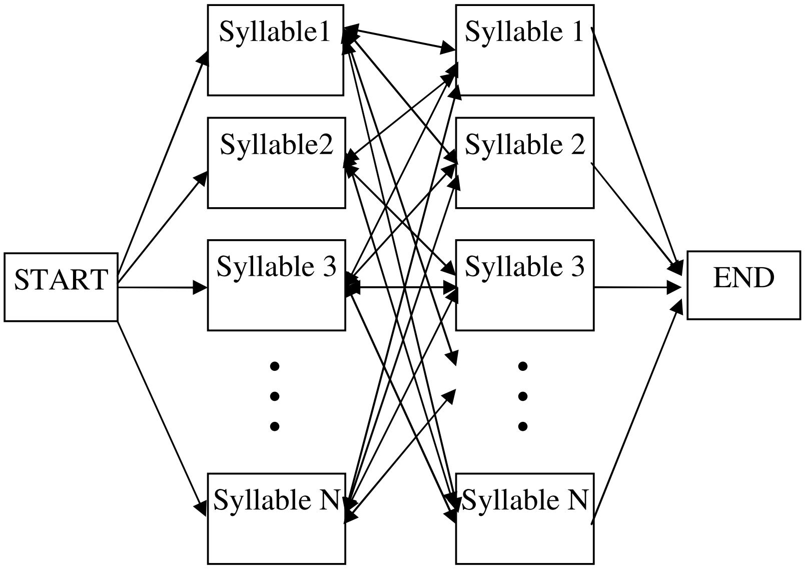

Figure 4.5. Syllable-based word model

As seen from Figure 4.5 the transitions are bilateral; i.e. syllables can follow each other with no constraints. The presence of the transitions between syllables is determined by the presence of the corresponding syllable bigram in the training corpus, and the weights of the transitions are the bigram probabilities.

Syllables are very short acoustic units; hence a syllable-based model does not promise high recognition rates. The previous two models had the advantage of employing stems as vocabulary units. It is possible to correct the recognizer errors by more sophisticated language processing modules if the stems are recognized correctly. However this is not the case for syllable-based case.

In addition, syllable based models do not enable detection of word boundaries as morpheme-based and stem-ending-based models can do. That is a crucial problem even for an isolated word recognition system.

5. STATISTICS OF CORPUS

5.1. Training Corpus

The lack of a large Turkish text corpus is an important drawback for research in Turkish language processing. There exists a text corpus at Bilkent University, which contains television news of TRT and some newspaper articles, but this corpus contains no more than one million words. Each researcher collects his own text corpus, mainly from newspaper articles (Çarkı *et al*., 2000), (Hakkani-Tür *et al*., 2000). The need for a large, standard Turkish corpus, containing text material from various domains is urgent.

We have collected text material to form a text corpus to construct statistical language models. The titles, authors, sources and domains of all the text material are given in Appendix B. The training corpus contains texts from literature (novels, stories, poems, essays), law, politics, social sciences (history, sociology, economy), popular science, information technology, medicine, newspapers and magazines.

Table 5.1 shows the number of words in the training data set (train1, train2,…, train9) we have prepared. Each training text is formed by adding new text data of approximately one million words to the previous training text; so for example train2 contains the text present in train1 and so on. All the training data we used in this thesis are of 9,142,750 words. Unfortunately, even for bigrams, this data size is not sufficient for effective language modeling. This fact will become clear in the following sections, when we look at the improvement in bigram models with respect to data size. The improvement does not seem to saturate except for the syllable-based model.

Words in each training text are morphologically decomposed, using the morphological parser we have mentioned in Chapter 3 to form morpheme-tokenized and stem-ending-tokenized training texts. The morpheme-tokenized training texts will be denoted as mor_train1, mor_train2,…,mor_train9 and the stem-ending-tokenized training texts will be denoted as se_train1, se_train2,…, se_train9. Table 5.1 shows the number and percentage of morphologically decomposed words in each training text. From Table 5.1 we can see that 93 per cent of the words were morphologically decomposed. Morphologically undecomposed words were left in their places in the morpheme-tokenized and stem-ending-tokenized training texts. They were considered as words with one morpheme (or words with no ending).

Table 5.1. Number of words and number of morphologically decomposed words in the training data set

	Number of words	Number of decomposed words	Number of undecomposed words	Percentage of decomposed words (%)	Percentage of undecomposed words (%)
train1	1,116,500	1,048,355	6,8145	93.9	6.1
train2	2,236,118	2,095,224	140,894	93.7	6.3
train3	3,232,674	3,023,178	209,496	93.5	6.5
train4	4,257,519	3,979,298	278,221	93.5	6.5
train5	5,341,227	4,986,375	354,852	93.4	6.6
train6	6,222,790	5,814,392	408,398	93.4	6.6
train7	7,328,388	6,858,267	470,121	93.6	6.4
train8	8,349,917	7,796,326	553,591	93.4	6.6
train9	9,142,750	8,476,678	666,072	92.7	7.3

In morpheme-tokenized texts the lexical descriptions of the suffixes appeared as tokens instead of their surface realizations. Since the surface realizations of suffixes are totally determined by the particular word they are added, they will have no contribution to the statistics of the language. Most of the lexical descriptions are in English, not to mix with other Turkish tokens (stems or undecomposed words). So in the morpheme-tokenized texts stems, lexical descriptions of suffixes and undecomposed words appear as tokens.

For stem-ending-tokenized texts it is more appropriate to use surface realizations of endings. In stem-ending-tokenized texts there exists stems, their endings and undecomposed words as tokens.

The training texts are also parsed into their syllables to form syllable-tokenized training texts, which are denoted as syl_train1, syl_train2,…, syl_train9. The syllable parser, we described before, assumes that the input word is a Turkish word; it does not check if the outcome syllables are legal Turkish syllables. Therefore the syllable parser produces strings, that are not Turkish syllables, for input words from foreign languages. We did not eliminate that "noise" from the text, assuming that such strings will appear rarely compared to correct Turkish syllables.

5.2. Test Data

In addition to training corpus, we constructed a test data set with the text material not present in any of the training texts. Table 5.2 shows the titles and domains of test texts.

Table 5.2. Titles and domains of the test texts

	Title	Domain
test1	Collection of essays	literature
test2	Transcription of TRT news	news
test3	Constitution of Turkish republic (1982)	law
test4	"Varolmanın dayanılmaz hafifliği", Translation of novel by Milan Kundera	literature
test5	"Seçme hikayeler", Stories by Ömer Seyfettin	literature
test6	"Lamiel" Translation of novel by Stendal	literature
test7	Collection of articles on psychiatry	medical
test8	"Siyasal Anılar ve Sosyal Demokrasinin Öykusu", Political memories by Cezmi Kartay	politics
test9	"Yeni Türkiye", Sociological research by Jean Deny	sociology
test10	Collection of articles on history	history
test11	Milliyet newspaper	newspaper
test12	Newspaper articles by Murat Belge	newspaper articles
test13	"Dergibi Dergisi", Literature magazine	literature
test14	Collection of articles on economy	economy

We have constructed a single test text of size about one million words by merging all the test texts given in Table 5.2. The perplexity and out of vocabulary rate measures were carried out with respect to this single test text. The language models obtained from training text train9, which contains all the training data, were tested on the test texts separately to compare the results with respect to different domains.

Table 5.3 gives the number of words in the test texts. Table 5.3 also gives the number and percentage of morphologically decomposed words in the test texts. For test7, test9, and test11 the percentage of morphologically decomposed words fall below 90 per cent. Total size of the test text is 1,091,804 words, 93 per cent of which is morphologically decomposed.

Table 5.3. Number of words and number of morphologically decomposed words in the test data set

	Number of words	Number of decomposed words	Number of undecomposed words	Percentage of decomposed words (%)	Percentage of undecomposed words (%)
test1	50,088	49,007	1,081	97.8	2.2
test2	91,451	82,929	8,522	90.7	9.3
test3	18,210	17,923	287	98.4	1.6
test4	64,331	60,809	3,522	94.5	5.5
test5	20,949	20,438	511	97.6	2.4
test6	22,904	21,544	1,360	94.1	5.9
test7	26,289	23,083	3,206	87.8	12.2
test8	82,047	76,615	5,432	93.4	6.6
test9	26,018	23,347	2,671	89.7	10.3
test10	58,951	55,278	3,673	93.8	6.2
test11	305,801	274,549	31,252	89.8	10.2
test12	80,641	76,615	4,026	95.0	5.0
test13	185,636	178,615	7,021	96.2	3.8
test14	58,488	56,240	2,248	96.2	3.8
total	1,091,804	1,016,992	74,812	93.1	6.9

5.3. Number of Distinct Tokens

The number of distinct tokens in the training text is an important figure, since it gives the minimum vocabulary size that will cover 100 per cent of the training data. The distinct tokens are not necessarily Turkish words; they can be both Turkish and foreign proper nouns, abbreviations, scientific and technical terms, wrongly spelled Turkish words, etc. In general, a dictionary that contains the words and proper nouns is used to eliminate wrongly spelled words, but we did not have such a dictionary of all Turkish words, all proper nouns, all scientific and technical terms. We could restrict the training text to have only those words that could be morphologically decomposed, but then we would miss the proper nouns and terms that are commonly used.

5.3.1. Word-tokenized Training Data

Table 5.4 gives the number of tokens, the number of distinct tokens and the number of new distinct tokens introduced with text addition, in the word-tokenized training texts, where the tokens correspond to words.

Table 5.4. Number of tokens (words), number of distinct tokens and number of new distinct tokens in the word-tokenized training data set

	Number of tokens (words)	Number of distinct tokens (distinct words)	Number of new distinct tokens (distinct words)
train1	1,116,500	144,227	144,227
train2	2,236,118	211,345	67,118
train3	3,232,674	268,845	57,500
train4	4,257,519	308,519	39,674
train5	5,341,227	346,277	37,758
train6	6,222,790	367,513	21,236
train7	7,328,388	396,889	29,376
train8	8,349,917	425,773	28,884
train9	9,142,750	457,684	31,911

As expected the number of distinct words are very high. For train1 containing about one million words, the number of distinct words is 144,227 and with train9 of about nine million words, this number increased up to 457,684. For each one million word added to the training corpus, approximately 40 thousand new words are introduced as an average.

The increase is due to two reasons: First reason is addition of new proper nouns, terms, wrongly spelled words, etc. The second reason is the addition of new word forms, of which the stems were not new to the previous training data. The amount of contribution of the morphological productivity can be figured out by looking at the number of new words in morpheme-tokenized training data.

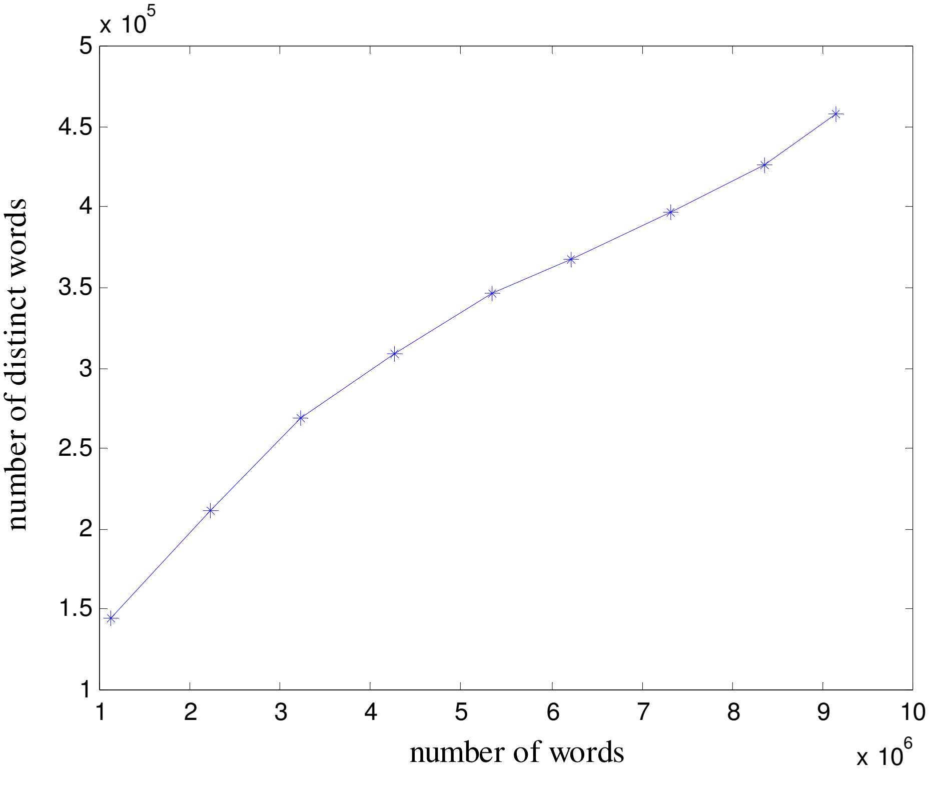

Figure 5.1. Number of distinct tokens (words) versus number of words in the word-tokenized training data set

Figure 5.1 shows the number of distinct tokens versus the number of words in the training data. As the data size is increased the amount of new word addition tends to drop.

This drop can be seen in Figure 5.1 as a drop of acceleration of distinct token increase. From train8 to train9 we observe an increase in the amount of new word addition since the additional text (added to train8 to form train9) contained medical text, which contains a large amount of medical terms.

5.3.2. Morpheme-tokenized Training Data

Table 5.5 gives the number of tokens in morpheme-tokenized training data. In this case tokens correspond to morphemes. Table 5.5 also gives the number of words in the corresponding word-tokenized training data as to estimate the average number of morphemes per word. This number came out to be 2.1, i.e. Turkish words have approximately two morphemes (a stem plus a suffix) on average. Nearly 50 per cent of the tokens in morpheme-tokenized training texts correspond to suffixes.

Table 5.5. Number of tokens (morphemes) and number of morphemes per word in the morpheme-tokenized training data set

	Number of words	Number of tokens (morphemes)	Morphemes per word
mor_train1	1,116,500	2,293,389	2.1
mor_train2	2,236,118	4,645,786	2.1
mor_train3	3,232,674	6,691,962	2.1
mor_train4	4,257,519	8,798,233	2.1
mor_train5	5,341,227	11,004,527	2.1
mor_train6	6,222,790	12,815,300	2.1
mor_train7	7,328,388	15,120,844	2.1
mor_train8	8,349,917	17,144,315	2.1
mor_train9	9,142,750	18,726,131	2.0

Table 5.6 shows the number of words having 1, 2, …, 10 morphemes and their percentage to the number of morphologically decomposed words in mor_train9. Words having one, two or three morphemes are typical, on the other hand words with more than five morphemes are seldom seen (eight in one thousand words). 37 per cent of

morphologically decomposed words are used in their root form. Figure 5.2 shows the histogram of number of morphemes per word.

Table 5.6. Number of words having 1, 2,…, 10 morphemes

Number of morphemes	**1**	**2**	**3**	**4**	**5**
Number of words	3,141,698	2,438,823	1,912,926	701,916	210,156
Percentage	37.1	28.8	22.6	8.3	2.5
Number of morphemes	**6**	**7**	**8**	**9**	**10**
Number of words	55,333	12,447	3,088	274	17
Percentage	0.7	0.1	0.04	0.003	0.0002

There are no words with more than 10 morphemes in the training data we used. We counted 17 words with 10 morphemes; following are two examples of them:

- bil+gi+le+n+dir+il+eme+me+si+ne
- kon+uş+la+n+dır+ıl+ma+ma+sı+na

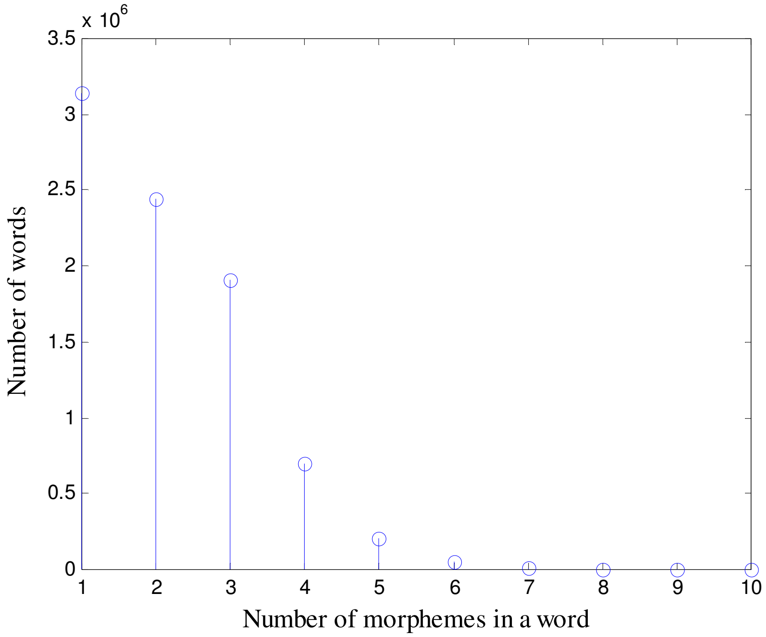

Figure 5.2. Histogram of number of morphemes per word

Table 5.7 gives the number of tokens, the number of distinct tokens and the number of new distinct tokens in morpheme-tokenized training data. As the training data is increased by one million words, approximately 13 thousand new tokens are introduced on average. The new proper nouns, terms, wrongly spelled words, etc. that appeared in the word-tokenized training text also appeared in morpheme-tokenized training texts. When compared with the distinct token increase in the word-tokenized training data we can deduce that for each one million text addition there are 27 thousand new words on average resulting from the morphological productivity of Turkish.

Table 5.7. Number of tokens (morphemes), number of distinct tokens and number of new distinct tokens in the morpheme-tokenized training data set

	Number of words	Number of tokens (morphemes)	Number of distinct tokens (distinct morphemes)	Number of new distinct tokens
mor_train1	1,116,500	2,293,389	34,830	34,830
mor_train2	2,236,118	4,645,786	52,538	17,708
mor_train3	3,232,674	6,691,962	68,986	16,448
mor_train4	4,257,519	8,798,233	81,671	12,685
mor_train5	5,341,227	11,004,527	92,505	10,834
mor_train6	6,222,790	12,815,300	98,464	5,959
mor_train7	7,328,388	15,120,844	105,347	6,883
mor_train8	8,349,917	17,144,315	114,805	9,458
mor_train9	9,142,750	18,726,131	134,727	19,922

The amount of new distinct token addition to mor_train6 and mor_train7 is relatively small compared to other training texts. This situation would show itself as a reduction in perplexity with respect to self, for mor_train6 and mor_train7; a result which we will discuss later.

Figure 5.3 gives number of distinct tokens versus number of words in the morpheme-tokenized training data. The increase in the number of distinct tokens from mor_train8 to mor_train9 is more abrupt than in word-based case. This proves our assumption that most of the new distinct words in word-based case is due to new forms of words, whose stems appeared in smaller training texts.

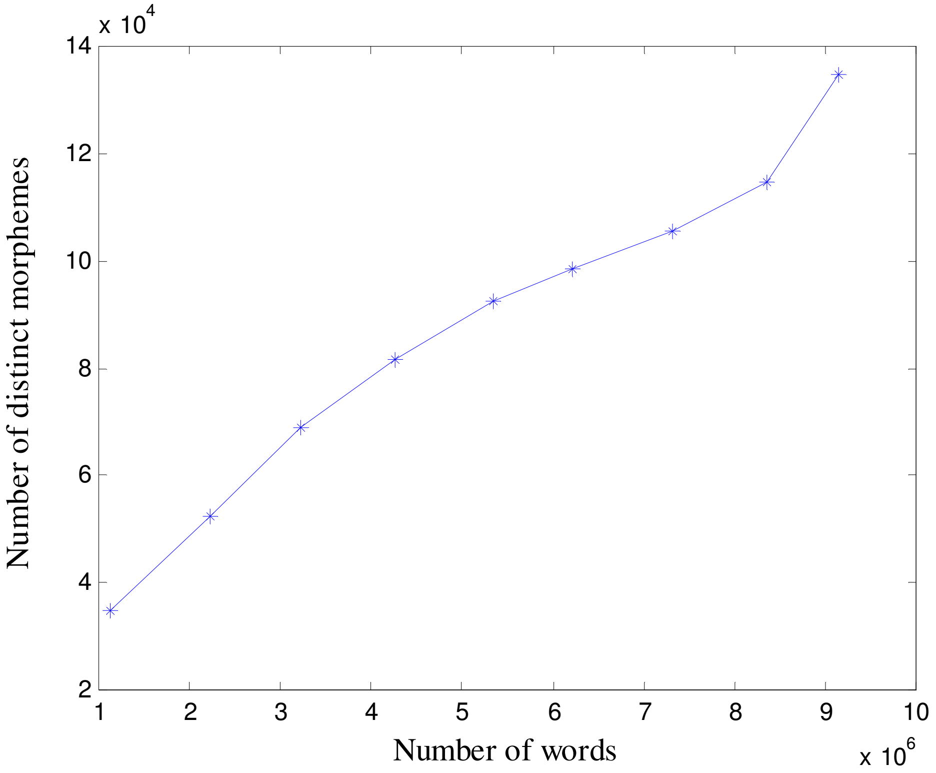

Figure 5.3. Number of distinct tokens (morphemes) versus number of words in the morpheme-tokenized training data set

5.3.3. Stem-ending-tokenized Training Data

Table 5.8 gives the number of tokens, number of distinct tokens, number of new distinct tokens in the stem-ending-tokenized training data set, where tokens correspond to stems and endings. Table 5.8 also gives the number of words in the corresponding word-tokenized training data set. The number of tokens per word came out to be 1,6 for stem-ending-based model. This means that 40 per cent of the words in the training data have no endings; they are either in root form or could not be decomposed by the morphological parser. 60 per cent of the words in the training corpus were parsed into two parts; i.e. their stems and endings.

Table 5.8. Number of tokens (stems and endings) and number of distinct tokens in the stem-ending-tokenized training data set

	Number of words	Number of tokens	Number of distinct tokens	Number of new distinct tokens
se_train1	1,116,500	1,785,933	51,634	51,634
se_train2	2,236,118	3,580,201	74,971	23,337
se_train3	3,232,674	5,159,056	96,121	21,150
se_train4	4,257,519	6,787,249	111,641	15,520
se_train5	5,341,227	8,497,906	125,178	13,537
se_train6	6,222,790	9,892,048	132,610	7,432
se_train7	7,328,388	11,654,171	141,815	9,205
se_train8	8,349,917	13,253,194	153,028	11,213
se_train9	9,142,750	14,477,050	173,904	20,876

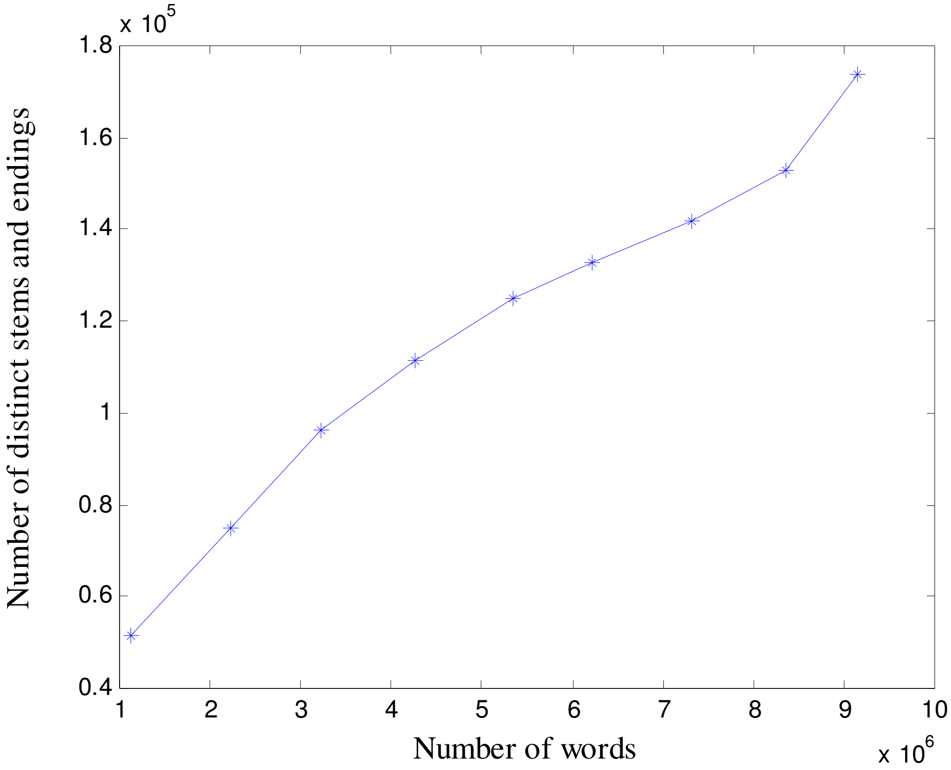

Figure 5.4. Number of distinct tokens (stems and endings) versus number of words in the stem-ending-tokenized training data set

Figure 5.4 gives the number of distinct tokens in the stem-ending-tokenized training data versus the number of words. As in the morpheme-based case, there exist fewer new distinct tokens in texts se_train6 and se_train7 and there is an abrupt change going from se_train8 to se_train9 due to new medical terms introduced to last training text.

For each one million words text addition, there appear approximately 15 thousand new tokens for stem-ending-tokenized training texts. That number was 40 thousand for word-tokenized training texts and 13 thousand for morpheme-tokenized training texts. With addition of new text data, new endings arise. When we compare with the morpheme-tokenized case we can conclude that two thousand new tokens corresponds to those new endings. This conclusion is verified by Table 5.9, which gives number of distinct endings and number of new distinct endings in the stem-ending-tokenized training data set.

In se_train1, which contains one million words, there exist 17,368 distinct endings. This number continues to grow up to 40,503 endings with se_train9. These large numbers demonstrate the morphological productivity of Turkish.

Table 5.9. Number of distinct endings and number of new distinct endings in the stem-ending-tokenized training data set

	Number of distinct endings	Number of new distinct endings
se_train1	17,368	17,368
se_train2	23,124	5,756
se_train3	27,962	4,838
se_train4	30,889	2,927
se_train5	33,711	2,822
se_train6	35,234	1,523
se_train7	37,636	2,402
se_train8	39,459	1,823
se_train9	40,503	1,044

Figure 5.5 gives the number of distinct endings versus the index of the stem-ending-tokenized training text. It can be seen that although the number of new distinct tokens

tends to drop with increasing data size, the increase continues. The number of distinct tokens do not saturate for nine million words of training data.

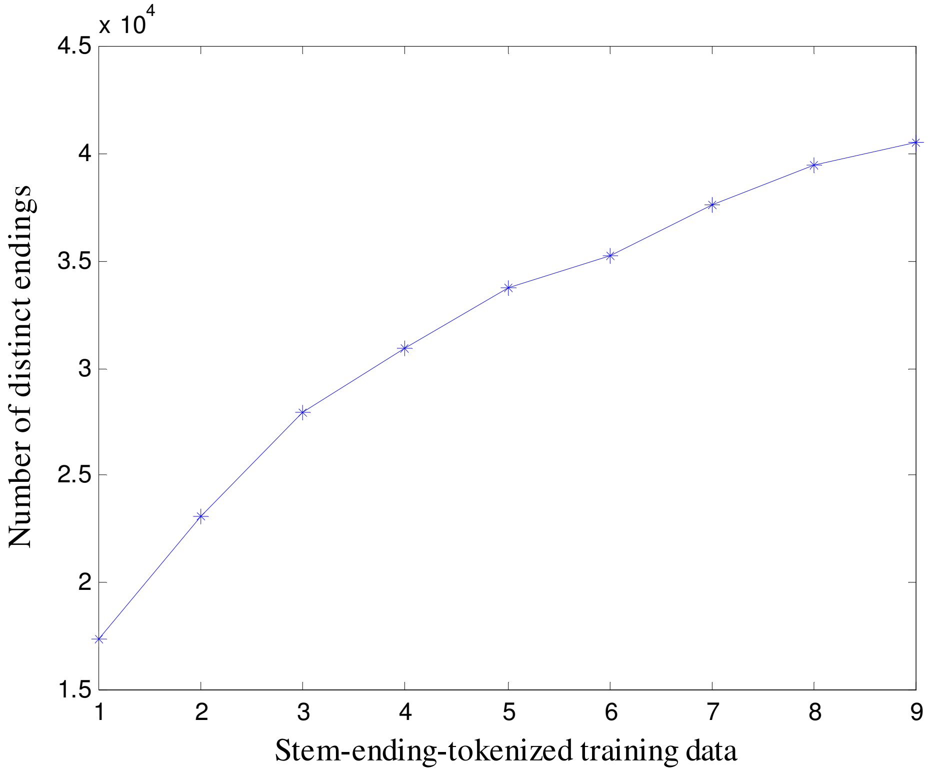

Figure 5.5. Number of distinct endings in the stem-ending-tokenized training texts

5.3.4. Syllable-tokenized Training Data

Table 5.10 shows the number of tokens (syllables), the number of distinct tokens present in each syllable-tokenized training text. We have mentioned that there are strings that should not be considered as "Turkish syllables" in the syllable-tokenized texts. A portion of distinct tokens corresponds to such strings. However, the number of distinct syllables is low as expected. We have estimated the maximum number of Turkish syllables as 13,728 with constraints on the last two consonants and 81,488 without constraints. There exist 12,148 different syllables, some of which are garbage, in the whole training text, lower than our estimation.

Table 5.10. Number of tokens (syllables) and number of distinct tokens in the syllable-tokenized training data set

	Number of words	Number of tokens (syllables)	Number of distinct tokens (distinct syllables)	Number of new distinct tokens
syl_train1	1,116,500	3,002,773	5,328	5,328
syl_train2	2,236,118	6,013,139	6,789	1,461
syl_train3	3,232,674	8,763,941	8,212	1,423
syl_train4	4,257,519	11,547,629	8,881	669
syl_train5	5,341,227	14,546,782	9,803	922
syl_train6	6,222,790	17,007,067	10,096	293
syl_train7	7,328,388	20,121,936	10,490	394
syl_train8	8,349,917	22,839,472	11,155	665
syl_train9	9,142,750	25,124,334	12,148	993

From Table 5.10, which also gives the number of words in the corresponding word-tokenized texts, we can estimate the number of syllables per word as 2.7. Figure 5.6 gives the number of distinct syllables versus number of words.

Word-based, morpheme-based and stem-ending-based models are comparable with each other since they both contain “words”, whether in root form or not, in their lexicon. On the other hand syllable-based model is not comparable with them; syllables are more similar to phonemes. Properties of syllables can be described in acoustic terms rather than linguistic. Since they are very short units, they can easily cover the language with very small “vocabularies” as phonemes do.

Small number of tokens, hence small vocabulary, makes the syllable-based model attractive for speech recognition task; however syllables, like phonemes, are due to acoustic confusability. Their acoustic properties are highly influenced by the acoustic context they are in.

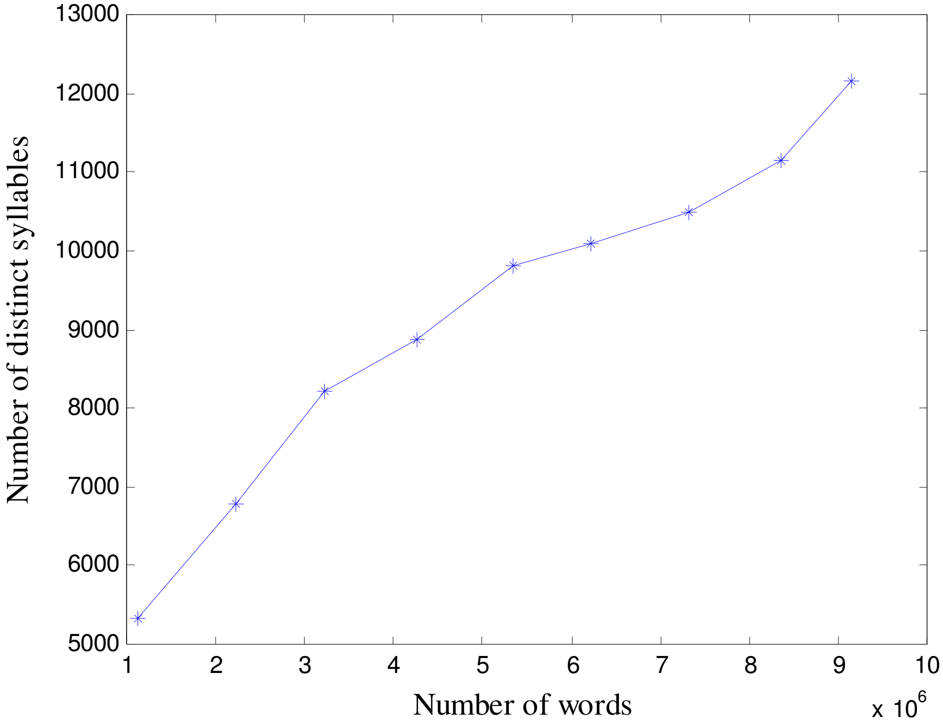

Figure 5.6. Number of distinct tokens (syllables) versus number of words in the syllable-tokenized training data set

5.3.5. Comparison of Unit Selection with respect to Number of Distinct Tokens

Figure 5.7 shows the number of distinct tokens versus number of words for all the four base-units. The curves for morpheme-based and stem-ending-based cases resemble each other. The slope difference between the curves corresponds to the new endings introduced to stem-ending-tokenized texts; the endings whose morphemes are not new to morpheme-tokenized training texts. However the slope difference between the curves for word-based case and morpheme-based case (or stem-ending-based case) is drastic. As we have mentioned, the average slope for word-tokenized texts is about 40 thousand while it is 13 thousand for morpheme-tokenized texts and 15 thousand for stem-ending-tokenized case.

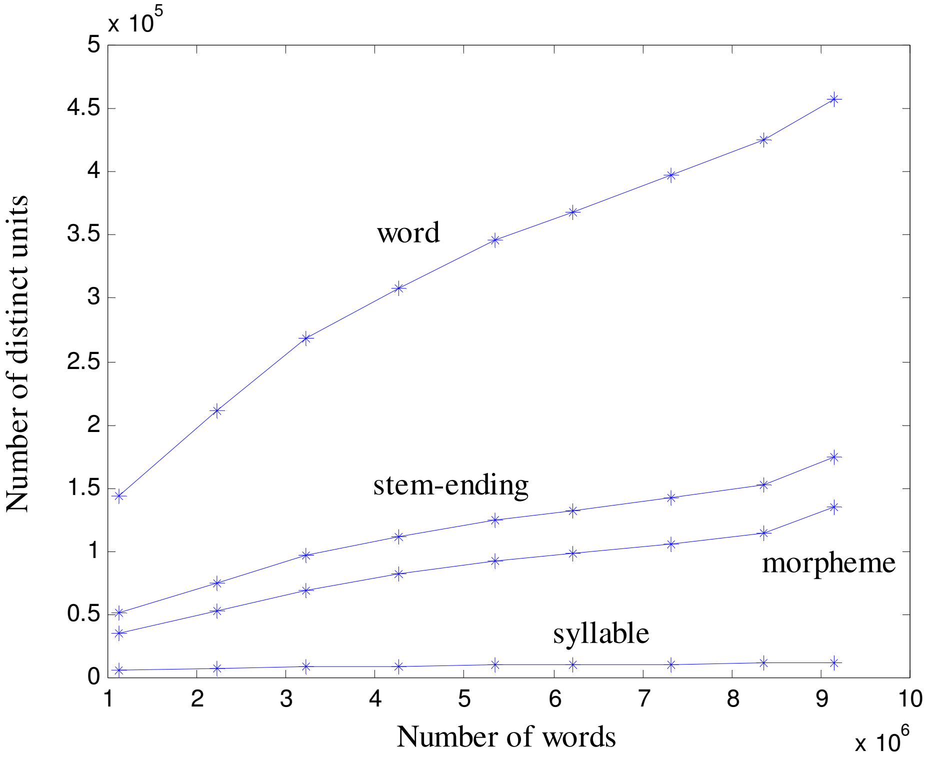

Figure 5.7. Number of distinct tokens versus number of words for training data set

The number of tokens and number of distinct tokens in the whole training for all units are reproduced in Table 5.11. The data size, counted in number of tokens, increases as the unit size becomes smaller. The problem of data sparseness can be overcome with large data size and small number of distinct tokens; therefore better estimations of bigram probabilities are possible with base units as morphemes and syllables.

Table 5.11. Number of tokens and number of distinct tokens in training data

	number of tokens	number of distinct tokens
word-tokenized	9,142,750	457,684
morpheme-tokenized	18,726,131	134,727
stem-ending-tokenized	14,477,050	173,904
syllable-tokenized	25,124,334	12,148

5.4. Coverage

It is impractical to construct a vocabulary from all the distinct words (or tokens) appeared in the training corpus. Rather smaller sizes of vocabularies are selected, items of which are frequent enough to cover a large portion of the text data. The most frequent units of a training corpus are selected as vocabulary units then this vocabulary is tested both on the training and test texts to see what percentage of the text it covers.

The percentage of out of vocabulary words is an important figure. Any word that is uttered by the speaker, and is not present in the vocabulary encounters a recognition error. Hence the vocabulary size to be selected depends on the application and the recognition error tolerance. Some applications, like radiology dictation, require smaller size of vocabularies, since the probability of uttering out of vocabulary words is too small. On the other hand general purpose dictation machines demand larger vocabularies containing tens of thousands of words.

In the following sections we have investigated the coverage with the increasing vocabulary size; where the items of vocabularies are the base-units we have selected. The maximum vocabulary size is limited to 60,000 items because of computational limitations. Training corpus used to select the most frequent words is train9, which contained about nine million words. The test data, used to measure coverage, was described in Section 5.2 and it contained about one million words.

5.4.1. Coverage with Words

Table 5.12 gives the percentage of coverage of word-tokenized training and test text versus vocabulary size. Despite the large number of distinct words present in training data, 100 most frequent words could cover more than 20 per cent of text data and, 2000 words are enough to cover more than 50 per cent of the training and test texts. However, much larger vocabularies are needed to approach 90 per cent coverage. With 60,000 most frequent words, 90 per cent of the training text and 88 per cent of the test text could be covered.

Table 5.12. Percentage of coverage of word-tokenized training and test data versus vocabulary size

Vocabulary size (in words)	Percentage of coverage (%)	
	train9	test
10	10.45	10.00
20	13.32	12.96
50	17.91	17.54
100	22.44	22.18
500	36.37	35.95
1000	43.95	43.30
2000	52.08	51.15
3000	57.00	55.94
5000	63.31	62.11
10000	71.72	70.42
20000	79.58	78.22
30000	83.75	82.32
40000	86.44	84.88
50000	88.36	86.74
60000	89.81	88.18

The out of vocabulary words in both training and text texts tend to be long words; i.e. words with relatively large number of morphemes. Words that have common stems and fewer morphemes are able to enter the vocabulary since they appear more frequent. All the morphological forms of frequent stems, present in the training corpus, do not appear in the vocabulary. For example the vocabulary of 60,000 words could not contain the words “iyileşiyorum (I am getting better)”, “duracaktır (It will definitely stop)”, “edermişçesine (as if doing)”, “sorunsuzdu (it was without problem)”, although their stems “iyi (good)”, “dur (stop)”, “et (do)”, “sorun (problem)” are very frequent. It can be intuitively said that, the words themselves should be recognized by a general purpose Turkish dictation system. The stem “gör (see)” is one of the most frequent Turkish stems; however its 1,360 morphological forms out of 1,775 are not in the vocabulary of size 60,000.

5.4.2. Coverage with Morphemes

Table 5.13 shows the percentage of coverage of morpheme-tokenized training and test text versus vocabulary size. With only 10 morphemes nearly 25 per cent of the training data is covered. With 50 morphemes half of the data is covered and 99 per cent coverage is reached with 20,000 morphemes.

Table 5.13. Percentage of coverage of morpheme-tokenized training and test data versus vocabulary size

Vocabulary size (in morphemes)	Percentage of coverage (%)	
	mor_train9	Test
10	24.18	23.97
20	34.51	33.74
50	51.88	51.71
100	61.60	61.43
500	79.94	79.74
1000	86.63	86.26
2000	91.81	91.51
3000	94.11	93.86
5000	96.20	96.05
10000	97.91	97.68
20000	98.85	98.59
30000	99.19	98.88
40000	99.38	99.01
50000	99.50	99.11
60000	99.60	99.18

We have stated in Section 5.3.2 that nearly 50 per cent of the morpheme-tokenized training data is composed of suffixes. Since we have only 111 suffixes represented in lexical level, a vocabulary containing those suffixes guarantee 50 per cent coverage. A vocabulary containing suffixes and the most frequent stems of Turkish solves the problem with the word-based case. Such a vocabulary would cover all the morphological word forms of the stem “gör (see)”.

5.4.3. Coverage with Stems and Endings

In Table 5.14, the percentage of coverage of stem-ending-tokenized training and test data with respect to vocabulary size is given. The stem-ending model demands large vocabulary (50,000 units) compared to morpheme-based model to obtain 99 per cent coverage. This is due to the large number of endings, in the order of tens of thousands. Morpheme-based model handles these endings with only 111 suffixes. However, once the endings appear in the vocabulary, the words whose stems and endings are present in the dictionary can be covered by that vocabulary.

Table 5.14. Percentage of coverage of stem-ending-tokenized training and test data versus vocabulary size

Vocabulary size (in stems and endings)	Percentage of coverage (%)	
	se_train9	test
10	11.74	11.54
20	16.89	16.49
50	25.56	25.25
100	34.18	33.91
500	59.63	59.37
1000	70.80	70.44
2000	80.37	79.92
3000	85.03	84.72
5000	89.83	89.60
10000	94.35	94.13
20000	97.03	96.71
30000	98.00	97.64
40000	98.50	98.08
50000	98.81	98.34
60000	99.03	98.50

We have stated that 40 per cent of the words in the training corpus have no endings; so endings constitute 37.5 per cent of the stem-ending-tokenized text.

5.4.4. Coverage with Syllables

Table 5.15 gives the percentage of coverage of syllable-tokenized training and test data with the vocabularies of given size. As expected small number of syllables could easily cover the text. With only 2,000 syllables, 99 per cent coverage is achieved. A speech recognizing system that utilizes syllables should build a vocabulary containing not more than 2,000 syllables. Vocabulary items that have arisen due to wrong parsing of foreign words should be cleaned up, and those foreign words, if demanded by the application, should be introduced to the vocabulary as a whole.

Table 5.15. Percentage of coverage of syllable-tokenized training and test data versus vocabulary size

Vocabulary size (in syllables)	Percentage of coverage (%)	
	syl_train9	test
10	16.96	16.67
20	27.28	27.2
50	46.97	47.24
100	61.69	61.82
500	90.75	90.58
1000	97.41	97.35
2000	99.55	99.54
3000	99.83	99.81
5000	99.95	99.92
10000	99.99	99.95
20000	100	99.96

5.4.5. Comparison of Unit Selection with respect to Coverage

Figure 5.8 and Figure 5.9 give the percentage of coverage of training and test data versus vocabulary size curves for the models proposed. The vocabulary size is in log scale.

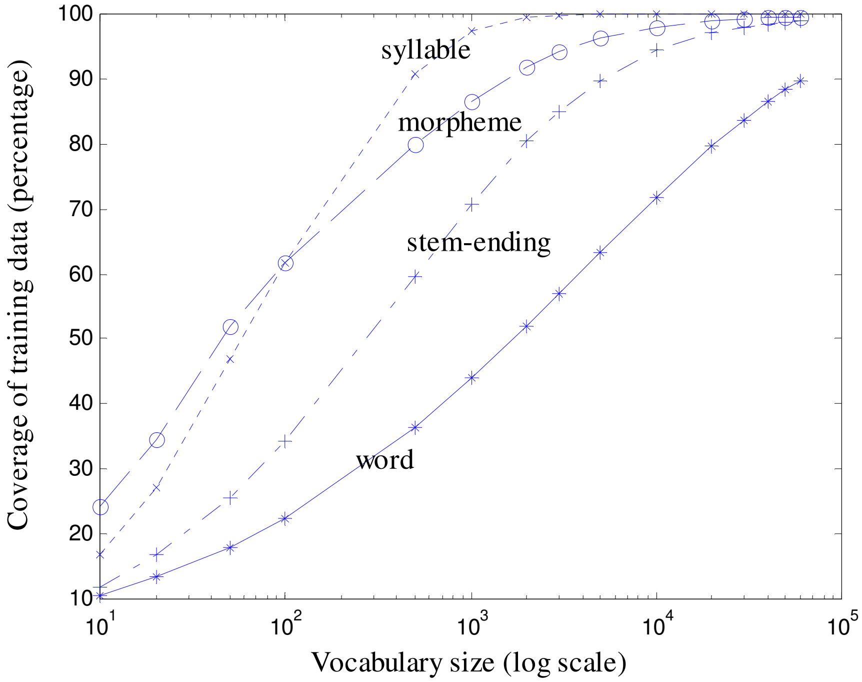

Figure 5.8. Coverage of training data in percentage with respect to vocabulary size

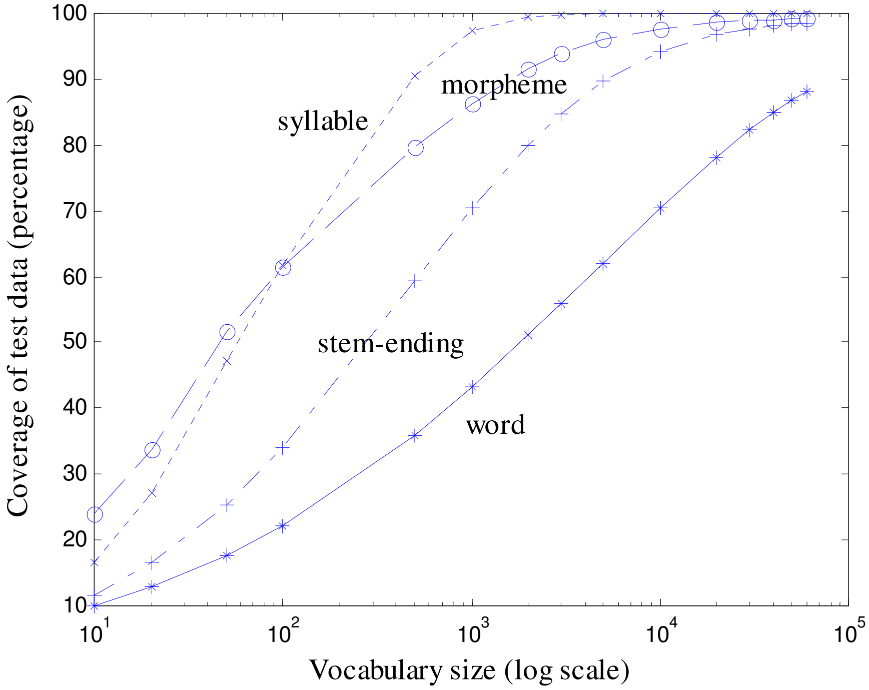

Figure 5.9. Coverage of test data in percentage with respect to vocabulary size

The syllable-based model has the steepest curve; i.e. the coverage increases fast with respect to small number of syllables. For vocabulary sizes smaller than 100, the morpheme-based model achieves higher coverage. As we have stated the most frequent suffixes and a few most frequent stems ("ve", "o", "bir") could cover 50 per cent of the morpheme-tokenized text. After vocabulary size 100, syllable-based model has highest coverage.

5.5. Bigram Models

We have utilized bigram models for speech recognition tests in this thesis. Therefore a detailed analysis of bigram models is made. We have obtained bigram language models for training texts of increasing data size, as described in Section 5.1. Perplexities and entropies are calculated with respect to self and test texts. Percentage of bigrams hit (percentage of bigrams in the test data, that were also present in the training data) is also given. The language model derivations, perplexity, entropy and percentage of bigrams calculations were done with the CMU Statistical Language Modeling Toolkit (Clarkson *et al.*, 1997). For all the models, except for the syllable-based case, vocabularies of size 60,000 items are used.

5.5.1. Word-based Bigram Model

Perplexity depends on the vocabulary size as well as the distribution of bigrams. If the probability distribution of the next word given a particular word is close to uniform distribution the model yields a high perplexity. This is the case for Turkish words because of the free word order property. Another factor is the size of the training corpus. When the data is sparse, the bigram frequencies will be close to each other, hence the distribution will be close to uniform. Larger training corpus will catch the regularities of the language better, and the perplexity (uncertainty about what will be the next word) will drop.

Another important factor, which is related with the training data size, is the percentage of bigrams hit in the test data. If a bigram in the test data did not occur in the training data then the probability of the second word is determined by its unigram probability. Using unigram probabilities instead of bigrams causes higher uncertainty.

Table 5.16 gives the perplexity and entropy values of word-based bigram language models obtained from word-tokenized data set with respect to self and test data. The perplexity values with respect to test data are very high, in the order of 10 thousands. However, the perplexities with respect to test data, drop significantly with the increasing training data size.

On the other hand, perplexities calculated with respect to self, the training data the model was obtained from, increase with the increasing training data size. This is due to new bigrams occurring with addition of new text. The number of words that can follow a particular word increases as new text data is introduced.

Table 5.16. Perplexities of bigram language models obtained from word-tokenized training data set with respect to self and test data

	Bigram perplexity and entropy with respect to self	Bigram perplexity and entropy with respect to test
train1	229.36 (7.84 bits)	4878.63 (12.25 bits)
train2	286.94 (8.16 bits)	3688.83 (11.85 bits)
train3	359.10 (8.49 bits)	3448.56 (11.85 bits)
train4	381.55 (8.58 bits)	3218.83 (11.65 bits)
train5	399.20 (8.64 bits)	3041.47 (11.57 bits)
train6	408.22 (8.67 bits)	2858.87 (11.48 bits)
train7	424.49 (8.73 bits)	2778.78 (11.44 bits)
train8	433.68 (8.76 bits)	2641.14 (11.37 bits)
train9	452.06 (8.82 bits)	2529.63 (11.30 bits)

Figure 5.10 gives the plots of perplexities with respect to self and test data versus the training data index. The improvement of the model from train1 to train2 is significant, then the improvement starts to decrease. However we observe no saturation sign of the improvement with the increasing size of the training data. The perplexities with respect to self and test data tend to approach each other. We assume they will meet at an asymptote when the training data is large enough as to estimate the correct probabilities.

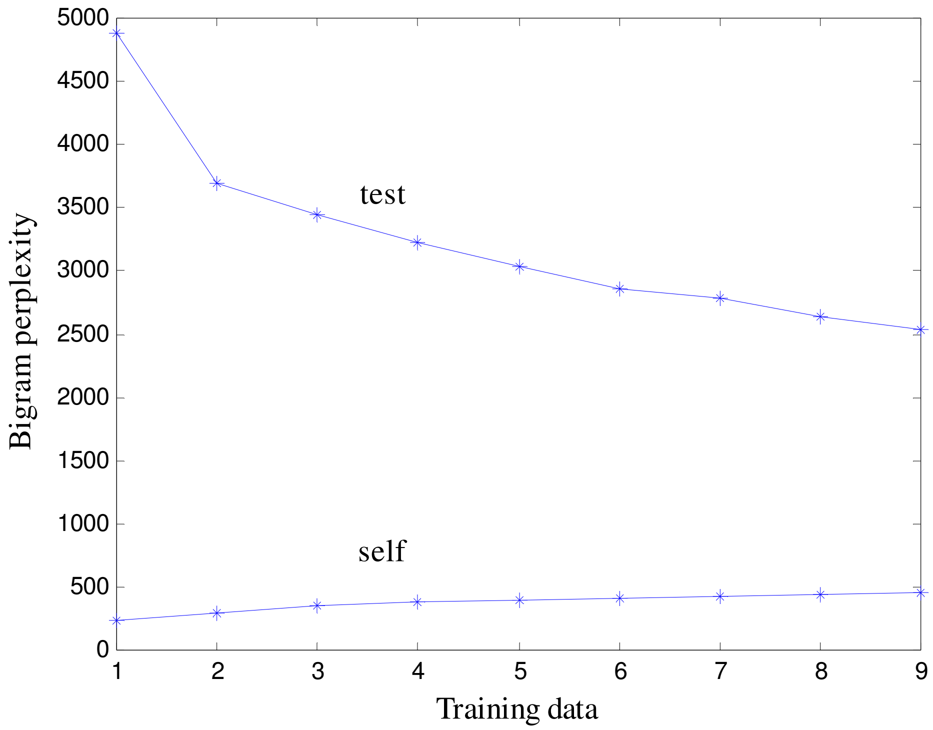

Figure 5.10. Perplexities of bigram language models obtained from word-tokenized training data set with respect to self and test data

Table 5.17. Percentage of bigrams hits in the test data with the word-based bigram language models obtained from the word-tokenized training data set

	Bigrams hits in the test data (percentage)
train1	43.48
train2	50.60
train3	53.76
train4	56.30
train5	58.41
train6	59.99
train7	61.41
train8	62.69
train9	63.95

Table 5.17 shows the percentage of bigrams hit in the test data. On average 50 per cent of the bigrams in the test text did not occur in the training data.

5.5.2. Morpheme-based Bigram Model

Table 5.18 gives the perplexities and entropies of the bigram language models obtained from morpheme-tokenized training data set with respect to self and test data. The perplexity values are much lower than the word-based model. That is due to the morphotactics of Turkish. Most of the bigrams appearing in the training data correspond to the transitions in the verbal and nominal finite state machines. The bigrams that are not defined by the Turkish morphotactics appeared at the word boundaries. The first token corresponds to the last morpheme of a word and the second token corresponds to the stem of the following word.

Table 5.18. Perplexities of bigram language models obtained from morpheme-tokenized training data set with respect to self and test data

	Bigram perplexity and entropy with respect to self	Bigram perplexity and entropy with respect to test
mor_train1	72.03 (6.17 bits)	148.46 (7.21 bits)
mor_train2	74.88 (6.23 bits)	132.54 (7.05 bits)
mor_train3	78.95 (6.30 bits)	127.56 (7.00 bits)
mor_train4	80.35 (6.33 bits)	124.34 (6.96 bits)
mor_train5	80.80 (6.34 bits)	121.96 (6.93 bits)
mor_train6	80.39 (6.33 bits)	119.39 (6.90 bits)
mor_train7	80.12 (6.32 bits)	118.75 (6.89 bits)
mor_train8	82.05 (6.36 bits)	116.53 (6.86 bits)
mor_train9	82.84 (6.37 bits)	115.57 (6.85 bits)

Figure 5.11 gives the plot of perplexity values versus the training data index. The difference between perplexities with respect to self and test data is not as large as in word-based case. The test perplexities show saturation sign with the increasing data size.

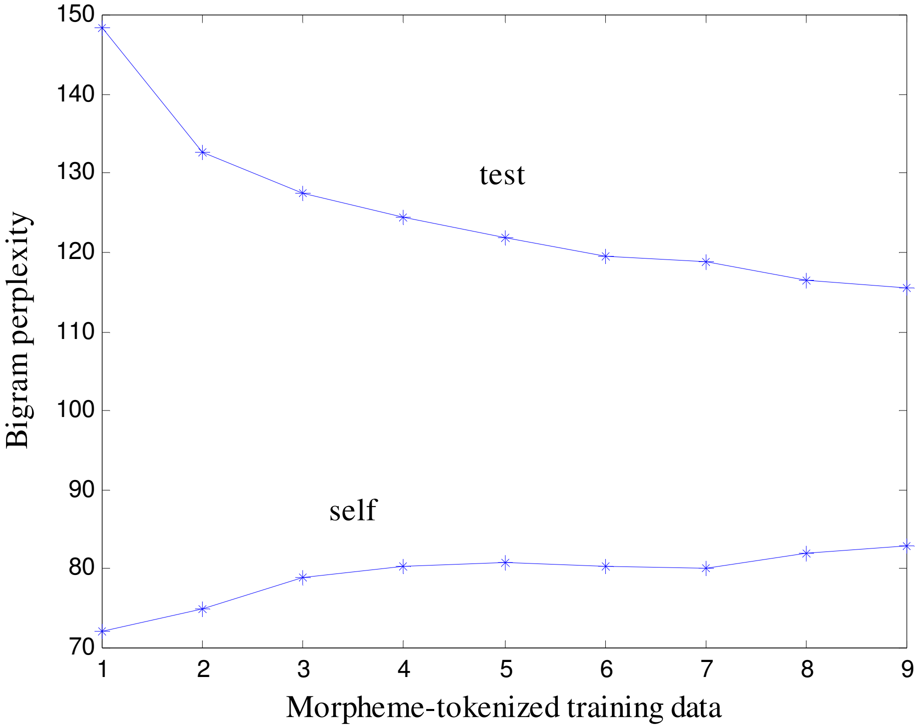

Figure 5.11. Perplexities of bigram language models obtained from morpheme-tokenized training data set with respect to self and test data

Table 5.19. Percentage of bigrams hits in the test data with the bigram language models obtained from the morpheme-tokenized training data set

	Bigrams hits in the test data (percentage)
mor_train1	85.16
mor_train2	88.84
mor_train3	91.10
mor_train4	92.04
mor_train5	92.74
mor_train6	93.22
mor_train7	93.55
mor_train8	93.95
mor_train9	94.21

Table 5.19 gives the percentage of bigrams hit in test text. The percentages of bigrams hit are very high; i.e. 90 per cent of the morpheme pairs in the test data has also appeared in the training data. However the existence of a particular bigram is not sufficient. The number of occurrences should be “high” enough to estimate the correct bigram probabilities. That is why the bigram perplexity of the morpheme-based model continue to improve with increasing size of training data.

5.5.3. Stem-ending-based Bigram Model

Table 5.20 gives the perplexity values of stem-ending-based bigram language models with respect to self and test data. The values are high compared to the morpheme-based model. The bigrams in stem-ending-tokenized text is of three kinds: The pair is composed of either a stem following another stem or a stem following an ending or an ending following a new stem. Since the number of endings that can follow a word is high the bigram perplexity is higher than the morpheme-based case. With the increasing number of stems and endings in the training data, the bigram perplexities with respect to self also increase.

Table 5.20. Perplexities of bigram language models obtained from morpheme-tokenized training data set with respect to self and test data

	Bigram perplexity and entropy with respect to self	Bigram perplexity and entropy with respect to test
set_train1	121.61 (6.93 bits)	509.90 (8.99 bits)
set_train2	137.29 (7.10 bits)	410.63 (8.68 bits)
set_train3	151.29 (7.24 bits)	383.02 (8.58 bits)
set_train4	155.91 (7.28 bits)	366.06 (8.52 bits)
set_train5	157.90 (7.30 bits)	353.31 (8.46 bits)
set_train6	158.07 (7.30 bits)	339.39 (8.41 bits)
set_train7	159.79 (7.32 bits)	333.76 (8.38 bits)
set_train8	163.43 (7.35 bits)	322.45 (8.33 bits)
set_train9	166.38 (7.38 bits)	315.65 (8.30 bits)

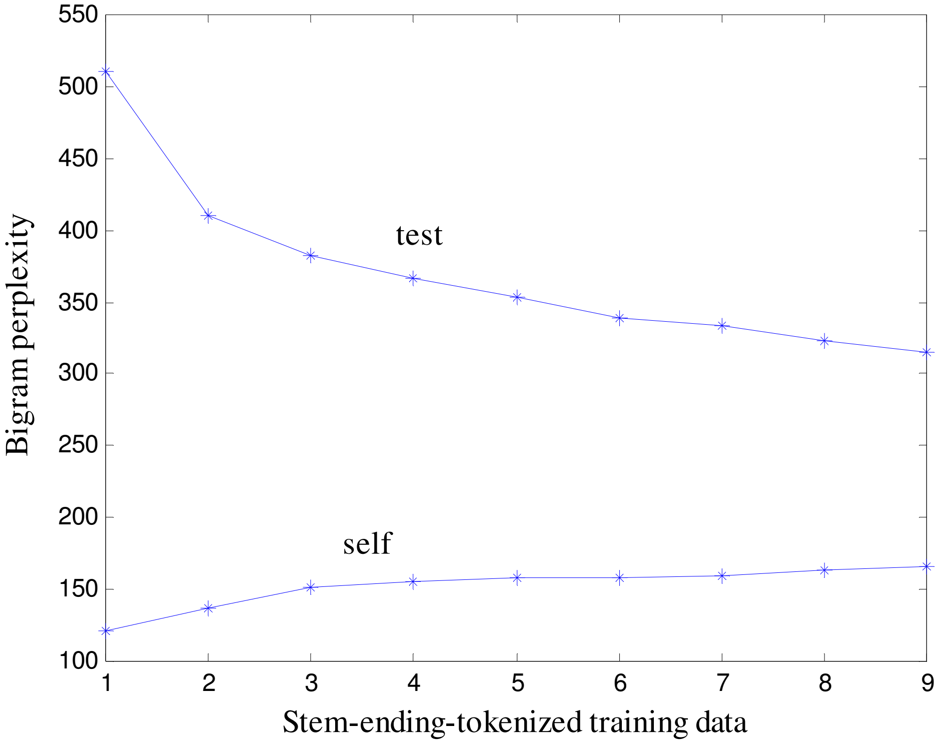

Figure 5.12. Perplexities of bigram language models obtained from stem-ending-tokenized training data set with respect to self and test data

Table 5.21. Percentage of bigrams hits in the test data with the bigram language models obtained from the stem-ending-tokenized training data set

	Bigrams hits in the test data (percentage)
se_train1	68.81
se_train2	77.41
se_train3	80.10
se_train4	81.86
se_train5	83.24
se_train6	84.25
se_train7	85.07
se_train8	85.89
se_train9	86.51

Figure 5.12 shows the plots of the perplexity values versus training data index. As with the word-based and morpheme-based case the values approach to each other. The improvement is still more significant for the high training data size compared to the morpheme-based case.

Table 5.21 gives the percentages of bigram hits in the stem-ending-tokenized test data. This value grows from 69 per cent to 87 per cent with the increase of training data size since new endings are seen and modeled.

5.5.4. Syllable-based Bigram Model

Table 5.22 gives the perplexity values with respect to the syllable-tokenized self and test data. The vocabulary sizes are also included to the table. Since the vocabulary sizes are too low compared to the previous three models' (60,000) the perplexity values obtained from syllable tokenized training data are not comparable with others'. The perplexities are small due to the small vocabulary sizes. Another factor is that the syllables in frequent words also occur frequently.

Table 5.22. Perplexities of bigram language models obtained from syllable-tokenized training data set with respect to self and test data

	Vocabulary size	Bigram perplexity with respect to self	Bigram perplexity with respect to test
syl_train1	5328	56.23 (5.81 bits)	78.78 (6.30 bits)
syl_train2	6789	56.02 (5.81 bits)	71.10 (6.15 bits)
syl_train3	8212	57.22 (5.84 bits)	69.40 (6.12 bits)
syl_train4	8881	58.08 (5.86 bits)	68.33 (6.09 bits)
syl_train5	9803	57.72 (5.85 bits)	67.43 (6.08 bits)
syl_train6	10096	56.87 (5.83 bits)	66.44 (6.05 bits)
syl_train7	10490	56.36 (5.82 bits)	66.43 (6.05 bits)
syl_train8	11155	57.64 (5.85 bits)	65.56 (6.03 bits)
syl_train9	12148	58.62 (5.87 bits)	65.42 (6.03 bits)

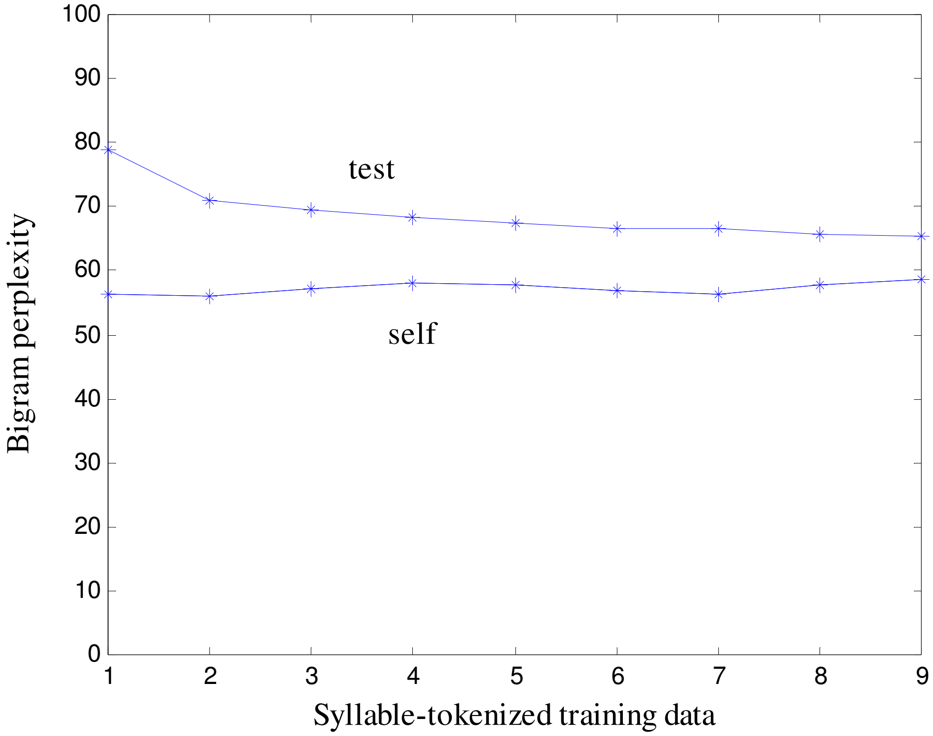

Figure 5.13. Perplexities of bigram language models obtained from syllable-tokenized training data set with respect to self and test data

Table 5.23. Percentage of bigrams hits in the test data with the bigram language models obtained from the syllable-tokenized training data set

	Bigrams hits in the test data (percentage)
syl_train1	94.99
syl_train2	96.91
syl_train3	97.57
syl_train4	97.92
syl_train5	98.25
syl_train6	98.42
syl_train7	98.53
syl_train8	98.68
syl_train9	98.80

Figure 5.13 shows the plots of perplexities versus training data index. Perplexity values with respect to self and test data are already close to each other and little improvement is observed with the increasing size of training data. In Table 5.23 percentages of bigrams hit in the test data are given. This value also did not improve much with the increasing training data size.

5.6. Statistics with respect to Test Texts

Table 5.24 gives the out of vocabulary units appeared in each test text, where the vocabularies for word-based, morpheme-based and stem-ending-based models are of size 60,000 units. The syllable-based model has a vocabulary of 12,000 units. The vocabularies are constructed with the most frequent units appearing in the training data.

Table 5.24. Percentage of out of vocabulary units in test data set

Percentage of out of vocabulary units				
	Word-based model	Morpheme-based model	Stem-ending-based model	Syllable-based model
test1	15.97	0.66	1.39	0.01
test2	12.04	1.49	2.39	0.09
test3	6.02	0.04	0.20	0.00
test4	14.84	1.20	2.12	0.02
test5	15.81	0.72	1.23	0.01
test6	15.95	1.85	2.75	0.34
test7	13.34	1.12	2.15	0.38
test8	8.47	0.51	1.21	0.00
test9	12.79	1.01	1.60	0.02
test10	9.13	0.68	1.23	0.02
test11	11.47	0.92	1.61	0.03
test12	8.35	0.37	0.79	0.01
test13	14.85	0.67	1.38	0.02
test14	6.74	0.30	0.65	0.01

We have the highest number of out of vocabulary words in test1, however with the latter three models, this test text did not happen to contain more out of vocabulary units compared to other test texts. The situation is the same for test5. Test1 and test2 are texts from literature and they do not contain much terms and proper nouns. The high out of vocabulary rate is due to not frequent morphological forms of common stems, and they were resolved with morpheme-based and stem-ending-based models.

We can see from 5.24. that the relative amount of out of vocabulary words are more or less preserved with morpheme-based and stem-ending-based models; however for syllable-based case, there are too little differences among the texts.

Table 5.25. Perplexities of bigram language models with respect to test data set

Bigram perplexity				
	Word-based model	Morpheme-based model	Stem-ending-based model	Syllable-based model
test1	3276	139	426	85
test2	2553	94	271	57
test3	1195	73	188	43
test4	2975	110	328	67
test5	5410	165	502	92
test6	2931	119	343	80
test7	2230	94	283	60
test8	1907	88	234	51
test9	3014	160	420	76
test10	1453	85	227	51
test11	2634	140	342	73
test12	1870	107	257	58
test13	3588	140	424	80
test14	1821	70	200	42

Table 5.25 gives the bigram perplexities with respect to each test text. There are large differences among the word-based model’s perplexity values with respect to different

texts. On the other hand, although the ordering is more or less preserved, the differences are not as much with the other three models.

6. RECOGNITION EXPERIMENTS

6.1. The Recognizer

The recognizer used for speech recognition experiments in this thesis is developed by GVZ Speech Technologies Company. The properties of the recognizer are as follows:

- It is trained with speech data of only one speaker. The training speech data is composed of 7,433 phrases, corresponding to 16,287 words.
- Speech sampling rate is set to 16 KHz.
- Triphone models are used.
- The HMM's (Hidden Markov Models) were trained with forward-backward algorithm.
- The recognizer accepts bigram language models.

6.2. Measures of Recognition Performance

We have used Word Error Rate (WER) and Phoneme Error Rate (PER) measures to analyze the speech recognition performances of the models. Word Error Rate is simply the percentage of erroneously recognized words to the number of words to be recognized, as in (6.1).

$$\text{WER} = \frac{N_e}{N_t} \times 100 \tag{6.1.}$$

where N_e is the number of erroneously recognized words and N_t is the number of total words to be recognized.

Phoneme error rate is the average percentage of number of phoneme deletions, plus number of phoneme insertions, plus number of phoneme substitutions to the total number of phonemes in the reference word.

6.3. Recognition Experiments

Recognition experiments have been done with small vocabulary sizes for each model.

6.3.1. First Experiment

First experiment is done with the bigram models obtained from the train1; i.e. the training corpus containing one million words. We have used the 458 different morphological forms of the Turkish stem “gör (see)”, one of the most frequent verbal stems. The vocabulary for word-based model contains only these 458 words. The vocabulary of the morpheme-based model contains 69 different stems and 558 suffixes represented at the surface level. The vocabulary of stem-ending-based model contains 23 different stems and 4398 endings. For the syllable-based model we have constructed a vocabulary with the most frequent 1,000 syllables of the training corpus.

The vocabulary sizes, word error rates and phoneme error rates that were obtained from each model are given in Table 6.1. The results for morpheme-based and syllable-based model are discouraging. However the stem-ending model yielded comparable performance to the word-based case although it is the model with the highest vocabulary size. That is because, stem-ending-based model puts more constraints to the possible unit sequences and has longer base units compared to the morpheme-based and syllable-based models.

Table 6.1. Recognition performances for the first experiment

	Vocabulary size	WER (%)	PER (%)
Word-based	458	22	5.3
Morpheme-based	627	75	25.7
Stem-ending-based	4,412	36	8.5
Syllable-based	1,000	83	27.9

6.3.2. Second Experiment

The second experiment was performed with the same words to be recognized as in the first experiment; i.e. the 458 morphological forms of the stem “gör”. This time, the training corpus of size nine million words is used to extract the bigrams and bigram probabilities. The word-based model was not tested for recognition.

The morpheme-based lattice had 271 stems and 558 suffixes in its vocabulary. We have defined two morpheme-based model named as “morpheme1” and “morpheme2”. The model morpheme1 allowed only the bigrams that appeared in the training corpus while morpheme2 allowed all the bigrams possible due to the morphotactics. The bigrams that did not appeared in the training corpus were assigned a small probability. Morpheme1 contained 12,381 bigrams, while morpheme2 contained 20,729 bigrams.

The stem-ending-based lattice had 192 stems and 13,011 endings. Only the bigrams that appeared in the training corpus were allowed. The lattice contained 73,072 bigrams.

The syllable-based model had a vocabulary constructed from the most frequent 1,500 syllables obtained from the training corpus. The number of bigrams were 81,465.

Table 6.2 summarizes the properties of the models, and gives WER and PER of the models. Table 6.2 also gives the recognition speed with each model.

Table 6.2. Recognition performances for the second experiment

	Vocabulary size	Number of bigrams	Recognition speed	WER (%)	PER (%)
Morpheme1	830	12,381	High	60	17
Morpheme2	830	20,729	High	69	22
Stem-ending-based	13,203	73,072	Low	31	8
Syllable-based	1,500	81,645	Low	90	35

Once more, we see that the stem-ending-based model gives the best performance. On the other hand, the recognition speed is very low (more than one minute per word) compared to the morpheme-based models. The morpheme-based model, especially the Morpheme1, seems to give better results with the increasing training data size.

6.3.3. Third Experiment

The third experiment was performed with the same lattices as in the second experiment. This time, the words to be recognized were selected from a paragraph. The number of words is 99, and the WER and PER obtained from each model is shown in Table 6.3.

Table 6.3. Recognition performances for the third experiment

	WER (%)	PER (%)
Morpheme1	50	27
Morpheme2	56	33
Stem-ending-based	42	22
Syllable-based	92	57

7. CONSLUSION

7.1. Conclusion

We have searched for base units alternative to words for constructing a language model for large vocabulary Turkish speech recognition. The alternative base units are defined to be morphemes, stems and endings and syllables. We have extracted some statistics to measure the appropriateness of the base units to speech recognition tasks. We have also performed some preliminary speech recognition experiments. Table 7.1 summarizes the properties of the alternative base units together with the words.

Table 7.1. Comparison of the base units

	Words	Morphemes	Stems and endings	Syllables
Vocabulary size (covering 90 per cent of the training text data)	60,000	2,000	5,000	500
Number of distinct tokens (in nine million words)	460,000	140,000	170,000	12,000
Decomposition difficulty	-	difficult	difficult	Easy
Meaning possession	meaningful	meaningful	meaningful	meaningless
Detection of word boundaries	enabled	enabled	enabled	disabled
Sensitivity to context	high	medium	medium	low
Bigram perplexity (with respect to test text)	2530	116	316	65
Complexity of model (in terms of complexity of lattices)	low	high	low	low
Order of N-grams for continuous speech recognition	low	high	low	high
Acoustic confusability (in terms of unit length)	low	high	low	very high
Recognition performance	-	low	medium	very low
Recognition speed	-	low	high	medium

Large amount of text data is necessary for effective language modeling. Lack of a Turkish text corpus is an important handicap for research on statistical language modeling. The Turkish texts we have collected can be a contribution to the efforts of corpus construction.

7.2. Future Work

It should be stated that the size of the training text data are far from sufficient even for bigram language modeling. Therefore next step should be enlargement of the text corpus. Also the training text corpus should be balanced; i.e. should be equally distant from various domains for modeling general English. Our corpus does not contain enough text material from all the domains; rather we have collected what we found on Web.

Enlargement of text corpus is also crucial for trigram language modeling. All the units proposed in this thesis, even the syllable-based model, will yield better recognition performances if efficient trigram modeling is done. Trigram models should be constructed and utilized on further Turkish speech recognition systems.

We have stated that the major handicap of parsing words is the increase of acoustic confusabality. Therefore a decompose-than-merge strategy can be applied to the morpheme-based model. We can merge short morphemes into larger ones in order to search for a solution between morpheme-based model and stem-ending-based model.

Another method can be a mixed model that selects the most frequent words in their undecomposed forms, the most frequent endings and the merged morphemes as vocabulary units. These units should be placed in the lattice in a way that does not violate the Turkish grammar.

A stem clustering method can be applied to reduce the number of bigrams. Through examination of bigram statistics, stem cathegories that share similar bigrams can be extracted. Then the stems of the same cathegory can be connected to the same group of morphemes (or endings) with the same bigram probabilities.

The syllable-based model failed in the recognition experiments. However if a seperate HMMs can be constructed for each syllable, the recognition performance can increase. This can also be done by training pentaphone models since Turkish syllables cannot have more than four phonemes. This improvement together with trigram modeling can yield better recognition performances of a system utilizing syllable-based language model.

APPENDIX A: LEXICAL REPRESANTATION OF SUFFIXES

Following list gives the lexical representations of Turkish suffixes that are used in the morphological parser. Most of the suffixes are named in English not to mix with the Turkish stems that share the same surface represantation.

ablative
ablative3
accusative
accusative3
adj_dig
adj_dik
adjective_suffix_lH
adjective_suffix_sHz
adv_dan
adverb_suffix_cA
amazlig
amazlik
attitude_adverb
casina_a
causative_dir
causative_t
dative
dative3
definiteness
derivational_cH
derivational_cHg
derivational_cHk
derivational_lHg
derivational_lHk
dikca

perso2p
person1p
person1p_k
person1p_k2
person1p_lHm
person1s
person1s_m
person2p
person2p_neg
person2p_niz
person2s
person2s_n
person2s_neg
person3p
person3p_neg
plural
possesive1p
possesive1s
possesive2p
possesive2s
possesive3
possesive3s
reflexive
relative
relative_suffix_ki

genitive
gi_der
if_suffix
iken_e
im_der
la_der
la_tense_yor
lAs
locative
locative3
madan
mak
maksizin
mazlig
mazlik
negative_aorist
negative_ma
negative_ma_tense_yor
negative_yama
negative_yama_tense_yor
noun_ma
optative
p2_pl
p2_plural_iz
p3_plural
passive_Hl
passive_Hn
past_hikaye
past_rivayet
perso1p
perso1s
sana_rica
sanica_rica
tense_ar
tense_di
tense_ir
tense_makta
tense_mali
tense_mis
tense_sa
tense_yor
ya_kadar
yabil
yacag
yacak
yadur
yagel
yagor
yakal
yakoy
yali
yamadan
yan
yarak
yasi
yayaz
yici
yinca
yip
yis
yiver

APPENDIX B: TEXT DATABASE

Following list gives the properties of the text material that constitute the text corpus. The title, author, domain and size of each text are indicated. If the text material is obtained from the Web, the Web site is also indicated.

- "Çağdaş Tiyatroda Aydın Sorunu", Handan Salta, research (theatre), http://www.altkitap.com, 15,175 words
- "Bal-Ayı", Ergun Kocabıyık, stories (literature), http://www.altkitap.com, 17,183 words
- "Belki de Gerçekten İstiyorsun", Murat Gülsoy, stories (literature), http://www.altkitap.com, 21,249 words
- "Konuşmayan Adam, Yaşamı Kapsamayan Bir Anlatı", Özge Baykan, essays (literature), http://www.altkitap.com, 9,978 words
- "Bir Laboratuar Romansı", Adnan Kurt, essays, http://www.altkitap.com, 23,480 words, 46,801 words
- "Cazname I", Tunçel Gürsoy, essays, http://www.altkitap.com, 46,801 words
- "Bir Hasta Sahibinin Hastane Günlüğü", Doğan Pazarcıklı, essays (literature), http://www.altkitap.com, 16,271 words
- "Kumcul, Bir Karabasan", İbrahim Yıldırım, essays (literature), http://www.altkitap.com, 5,519 words
- "Rüzgara Karşı II", Ömer Madra, essays, http://www.altkitap.com, 37,537 words
- "Sivil Toplum, Düşünsel Temelleri ve Türkiye Perspektifi", Ayşenur Akpınar Gönenç, Research (politics), http://www.altkitap.com, 35,404 words
- "Günümüz Basınında Kadın(lar)", Leyla Şimşek, research, http://www.altkitap.com, 35,807 words
- Collection of essays, (literature), 50,092 words
- "Alacakaranlıkta & Tonio Kröger", Thomas Mann, novel (literature), http://ekitap.kolayweb.com, 25,848 words
- "Alice Harikalar Ülkesinde", Lewis Carroll, novel (literature), http://ekitap.kolayweb.com, 19,263 words

- “Anadolu'da Arkeobotani (Tarihe Işık Tutan Tarım Ürünleri)”, Mark Nesbit, research (science, archeobotany), http://ekitap.kolayweb.com, 2,305 words
- “Artistler ve Modeller”, Anais Nin, story (literature), http://ekitap.kolayweb.com, 1,167 words
- “Bizans”, Ferenc Herczeg, tragedy (literature), http://ekitap.kolayweb.com, 17,099 words
- “Bozkırda”, Maksim Gorki, stories (literature), http://ekitap.kolayweb.com, 17,363 words
- “Değirmenimden Mektuplar”, Alphonse Daudet, novel (literature), http://ekitap.kolayweb.com, 35,304 words
- “Deney”, Mustafa Yelkenli, story (science-fiction), http://bilimkurgu.cjb.net, 1,731 words
- “Başkasının Karısı & Namuslu Hırsız”, Dostoyevski, stories (literature), http://ekitap.kolayweb.com, 24,910 words
- “Apartman I”, Emile Zola, novel (literature), http://ekitap.kolayweb.com, 22,841 words
- “Apartman II”, Emile Zola, novel (literature), http://ekitap.kolayweb.com, 25,818 words
- “Çin Gezisi (Beijing'te Dört Gün)”, Ergün Aydalga, travel notes (literature), http://www.aydalga.gen.tr, 2,832 words
- “Genç Werther'in Acıları”, Johann Wolfgang Goethe, novel (literature), http://ekitap.kolayweb.com, 30,006 words
- “Bir Ejderha ile Savaşan Bilgisayar'ın Hikayesi”, Stanislav Lem, story (science-fiction), http://ekitap.kolayweb.com, 1,716 words
- “Haksız Yönetime Karşı & Tembellik Hakkı”, Henry D.Thoreau&Paul Laforge, (politics), http://ekitap.kolayweb.com, 17,531 words
- “Hatay'ın Kurtuluşu İçin Harcanan Çabalar”, Tayfur Sökmen, research (history), http://ekitap.kolayweb.com, 29,929 words
- “Jules Amcam”, Guy de Maupassant, stories (literature), http://ekitap.kolayweb.com, 28,079 words

- “Kitaptan E-kitaba, Teknoloji ve Yaşam Kültürü Dergisi”, journal article (technology), http://www.microsoft.com/turkiye/mslife/aralik00/teknoloji.htm, 1,841 words
- “Konuşmalar”, Konfüçyüs, dialogues (philosophy), http://ekitap.kolayweb.com, 22,641 words
- “Küçük Prens”, Antoine de Saint-Exupery, novel (literature), http://arzudurukan.www9.50megs.com, 10,542 words
- “Yoksul Çalgıcı”, Franz Grillparzer, stories (literature), http://ekitap.kolayweb.com, 12,616 words
- “Macbeth”, Shakespeare, play (literature), http://ekitap.kolayweb.com, 14,618 words
- “Menekşe Yakınlığı”, Ümit Oktay Yılmaz, poems (literature), http://ekitap.kolayweb.com, 4,137 words
- “Metafizik Üzerine Konuşma”, Leibniz, (philosophy), http://ekitap.kolayweb.com, 24,118 words
- “Hastalık Hastası”, Moliere, play (literature), http://ekitap.kolayweb.com, 14,418 words
- “New York'u Nasıl Sevdi”, O'Henry, stories (literature), http://ekitap.kolayweb.com, 26,640 words
- “Penguenler Adası”, Anatole France, novel (literature), http://ekitap.kolayweb.com, 53,115 words
- “Doğudaki Hayalet”, Pierre Loti, memories (literature), http://ekitap.kolayweb.com, 22,222 words
- “Erzurum Yolculuğu&Byelkin'in Öyküleri”, Puşkin, stories (literature), http://ekitap.kolayweb.com, 29,958 words
- “Ertelenen Aşk”, Ray Brudbury, story (science-fiction), http://ekitap.kolayweb.com, 2,845 words
- “Sevginin Gül Kokusu”, Gülsüm Güven, stories, http://ekitap.kolayweb.com, 1,222 words
- “Stuttgart Cücesi”, Eduard Mörike, tales (literature), http://ekitap.kolayweb.com, , 27,905 words
- “Sürbeyan”, Kayhan Belek, poems (literature), http://ekitap.kolayweb.com, 2,481 words

- “Taras Bulba”, Gogol, stories, http://ekitap.kolayweb.com, 31,824 words
- “Atatürk'le Konuşmalar”, Mustafa Baydar, biography (history), http://ekitap.kolayweb.com, 23,048 words
- “Üç Kısa Oyun: Sicilya Turunçları, Aptal, Ağzı Çiçekli Adam”, Luigi Pirandello, plays (literature), http://ekitap.kolayweb.com, 9,954 words
- “Nutuk”, Mustafa Kemal, autobiography (history) http://www.adk.boun.edu.tr/ataturk/kendi_kaleminden/nutuk/nutuk.htm, 123,299 words
- “Öyküler-masallar”, Ergun Aydalga, stories and tales (literature) http://www.aydalga.gen.tr, 51,183 words
- Collection of science-fiction stories (science-fiction), 14,751 words
- Collection of TV news, (news), Bilkent Text Corpus, 91,455 words
- “T.C. Anayasası, 1982”, Turkish constitution (law), 18,214 words
- “Devrim Yazıları”, Babeuf, (politics), http://ekitap.kolayweb.com, 11,375 words
- “Varolmanın Dayanılmaz Hafifliği”, Milan Kundera, novel (literature), 64,335 words
- Collection of stories by Ömer Seyfettin, stories (literature), 20,954 words
- “Şeyh Bedrettin Destanı”, Nazım Hikmet, poems (literature), 6,274 words
- “Lamiel”, Stendhal, novel (literature), http://ekitap.kolayweb.com, 22,936 words
- Collection of articles from psychiatry, articles (medicine, psychiatry), http://lokman.cu.edu.tr/psychiatry/default.htm, 26,379 words
- Collection of psychiatry course notes, course notes (medicine, psychiatry), http://lokman.cu.edu.tr/psychiatry/default.htm, 86,131 words
- Collection of psychiatry course notes, course notes (medicine, psychiatry), http://lokman.cu.edu.tr/psychiatry/default.htm, 43,808 words
- “Cumhuriyet öyküleri”, collection, stories (literature), http://www.kultur.gov.tr, 67,722 words
- Collection of Turkish legends, legends (literature), http://www.kultur.gov.tr, 145,494 words
- "Bilim ve Aklın Aydınlığında Eğitim Dergisi", magazine archive, stories, essays, articles, poems, http://www.meb.gov.tr, 99,246 words
- "Milli Eğitim Dergisi", collection of journal articles, articles (studies on education), http://www.meb.gov.tr, 315,505 words

- “Bilim ve Aklın Aydınlığında Eğitim Dergisi", magazine archive (literature), http://www.meb.gov.tr, 102,327 words
- “Osmanlı İmparatorluğu”, collection, research (history), http://www.kultur.gov.tr, 13,182 words
- “Aydınlık”, magazine archive (news, politics), http://www.aydinlik.com.tr ,133,984 words
- “Aydınlık”, magazine archive (news, politics), http://www.kultur.gov.tr, 94,663 words
- “Araf Dil-Düşünce Dergisi”, magazine archive, (literature, art, philosophy, language, politics), http://www.araf.net/dergi/, 587, 947 words
- “Evrensel Kültür Dergisi”, magazine archive (literature, art, politics) http://www.evrenselbasim.com/ek/index.asp, 323,642 words
- “Dış Politika, Kültür ve Tarihte Araştırma Dergisi”, magazine archive (politics, history), http://www.arastirma.org/, 59568
- Collection of essays on literature (literature), http://www.ykykultur.com.tr/kutuphane/sayi4/4kat4.htm, 25,665 words
- Collection of stories, (literature), http://www.basarm.com.tr/yayin/index.html, 230,891 words
- Collection of stories by Cezmi Ersöz, (literature), http://www.cezmiersoz.net 37,097 words
- “Onlar Başlattı”, Fahrettin Macit, novel (literature), http://www.basarm.com.tr/yayin/index.html, 149,662 words
- Collection of political articles, articles (politics, history), www.turan.tc, 81,407 words
- Collection of essays on Alevi culture, (culture), http://www.alevibektasi.com, 193,544 words
- Collection of political articles by members of Turkish parliement, articles (politics), http://www.tpb.org.tr/makaleler.htm, 11,934 words
- “Demokrasinin Kilit Taşı Anılar”, Nermin Çiftçi, memories (politics, history) http://www.basarm.com.tr/yayin/index.html, 53,598 words
- “Ortadoğu Su Sorunlari ve Türkiye”, Özden Bilen, research (politics), http://www.basarm.com.tr/yayin/index.html, 23,649 words

- "Siyasal Anılar ve Sosyal Demokrasinin Öyküsü", Cezmi Kartay, memories (politics, history), http://www.basarm.com.tr/yayin/index.html, 82,051 words
- "AB Ulusal Programı 1. cilt", government program, (government report) http://www.byegm.gov.tr, 165,850 words
- "Siyasi Partiler ve Demokrasi", symposium, (politics), http://www.basarm.com.tr/yayin/index.html, 34,602 words
- "Türkiye için Nasıl Bir Seçim Sistemi", Hikmet Sami Türk and Erol Tuncer, research (law), http://www.basarm.com.tr/yayin/index.html, 42,679 words
- "AB Ulusal Programı 2.cilt", Government program, (government report) http://www.byegm.gov.tr, 121,220 words
- "Avrupa Güvenlik Şartı", International aggreement (government report), http://www.byegm.gov.tr, 5,641 words
- "Basın Kanunu", (law), http://www.byegm.gov.tr, 4,292 words
- "Basın Kartı Yönetmeliği", (law), http://www.byegm.gov.tr, 6,591 words
- "Paris Şartı", International aggreement (government report), http://www.byegm.gov.tr, 6,835 words
- "Türkiye'nin Güçlü Ekonomiye Geçiş Programı", Government program (government report, economy), http://www.byegm.gov.tr, 4,832 words
- Collection of articles of social sciences, research (social sciences, economy), http://www.makaleler.8m.com, 102,996 words
- "Bilanço", Tevfik Çavdar, research (economy), Gelenek Yayınevi, 33,686 words
- "Bir Yılbaşı Öyküsü", Dudintsev, story (literature), Gelenek Yayınevi, 9,006 words
- "Sosyalizm Yolunda İnadın ve Direnci Adı: Kıvılcımlı", Emin Karaca, biography (biography, history), Gelenek Yayınevi, 29,377
- "Yeraltı Dünyadan Başka Bir Yıldız Değildi", Emin Karaca, research (history), 32,324 words
- "Küba Sokaklarında", Evren Madran, travel notes (literature), Gelenek Yayınevi, 41,731 words
- "Omurgayı Çakmak", Ali Mert, essays (literature, politics), Gelenek Yayınevi, 66,022 words
- "Sosyalist İktidar Dergisi, Seçme Yazılar", collection of political articles, (politics, social sciences), Gelenek Yayınevi, 8,0731 words

- “Şostokoviç, Hayatı ve Eserleri”, biography (music, history), Gelenek Yayınevi, 87,024 words
- “Sosyalizm ve Din”, collection of articles, (social sciences), Gelenek Yayınevi, 32,117 words
- “Devekuşu Rosa”, Yusuf Ziya Bahadınlı, novel (literature), Gelenek Yayınevi, 44,568 words
- “Radikal 2 dergisi”, magazine archive, (various domains: politics, music, art, literature, etc.), http://www.radikal.com.tr, 865,409 words
- “OTP ve Türkiye”, Chamber of Agriculture Engineers, report on economy (economy), http://www.tmmobzmo.org.tr/kutuphane.html, 11,108 words
- “Tahkim”, Chamber of Agriculture Engineers, report on economy (economy), http://www.tmmobzmo.org.tr/kutuphane.html,12,082 words
- “Ekonomik İstikrar, Büyüme ve Yabancı Sermaye”, articles on ecomony (economy), http://www.tcmb.gov.tr/, 74,145 words
- “Euro El Kitabı”, articles on ecomony (economy), http://www.tcmb.gov.tr/, 70,004 words
- “Aksiyon Dergisi”, journal archive, (news, politics), www.aksiyon.com.tr, 652,798 words
- “İnsan Hakları Koordinatör Üst Kurulu’nun Çalışmaları”, (government report), http://www.basbakanlik.gov.tr/yayinlar/yayinlar.htm, 187,949 words
- “Gelenek Dergisi”, journal archive, research (politicics, social sciences), http://www.gelenekyayinevi.com/gelenek_dizi.asp, 421,726 words
- “30 Ağustos Hatıraları”, memories (history) http://ekitap.kolayweb.com, 19,026 words
- “Sanal”, Ali Akdoğan, story (science-fiction), http://ekitap.kolayweb.com, 1,008 words
- “Gözlük”, Ali Akdoğan, story (science-fiction), http://ekitap.kolayweb.com, 1,595 words
- “Xar’ın Gezegeni”, Altuğ Gürkaynak, story (science-fiction), http://ekitap.kolayweb.com, 1,277 words
- “Ölümsüz Antikite”, Hikmet Temel Akarsu, novel (literature), http://ekitap.kolayweb.com, 54,539 words

- “Baz İstasyonları”, Burak Dağıstanlı, research (technical), http://ekitap.kolayweb.com, 11,047 words
- “Ben Gidersem Ay Sendeler”, Adnan Durmaz, poems (literature), http://ekitap.kolayweb.com, 7,924 words
- “Bir Teknoloji Öyküsü, COMMODORE’ün Yükseliş ve Çöküşü”, Bahadır Akın, research (technical), http://ekitap.kolayweb.com, 3,524 words
- “Mükemmel Konser”, Cüneyt Uçman, story (science-fiction), http://ekitap.kolayweb.com, 630 words
- “Danabaş Köyünün Öyküsü”, stories (literature), Celil Memmedguluzade, http://ekitap.kolayweb.com, 28,001 words
- “Türkiye Cumhuriyeti Devrim Yasaları” B. M. Erüreten, (history, politics), http://ekitap.kolayweb.com, 23424 words
- “Avrupa ile Asya ArasIndaki Adam, Gazi Mustafa Kemal”, Dagobert Von Mikusch, biography (biography, history), http://ekitap.kolayweb.com, 66,835 words
- “Gılgamış Destanı”, legend (literature), http://ekitap.kolayweb.com, 16,438 words
- “Gözlemci” , story (science-fiction), http://ekitap.kolayweb.com, 1,117 words
- “Knulp”, Hermann Hesse, novel (literature), http://ekitap.kolayweb.com, 21,692 words
- “Uğursuz Miras”, E.T.A Hoffmann, http://ekitap.kolayweb.com, 24,021 words
- “İlk Son ve En İlginç Yolculuk”, Hasan Özkan, story (science-fiction), http://ekitap.kolayweb.com, 1,861 words
- “Marie Grubbe”, Jens Peter Jacobsen, novel (literature), http://ekitap.kolayweb.com, 61,904 words
- “Yeni Türkiye”, Jean Deny, research (social sciences), http://ekitap.kolayweb.com, 26,026 words
- “27 Mayıs'tan 12 Mart'a”, Nadir Nadi, research (social sciences), http://ekitap.kolayweb.com, 35,135 words
- “Gökdelen”, Serkan Turan, story (science-fiction), http://ekitap.kolayweb.com, 1,907 words
- “Tristan ve Iseut”, tales (literature), http://ekitap.kolayweb.com, 28,696 words
- “M.Ö. 2. Binyılın 2. Yarısında Asur'un Anadolu'ya Yayılımı”, Umut Devrim Eryarar, research (history), http://ekitap.kolayweb.com, 2,912 words

- "Birinci Balkan Savaşı", Yusuf Hikmet Bayur, research (history), http://ekitap.kolayweb.com, 72,995 words
- Collection of speeches of Turkish presidents, (politics, formal speech), http://www.cankaya.gov.tr, 196,484 words
- "Avukatlık Kanunu", (law), http://www.aciksayfa.com, 51,539 words
- "Açık Sayfa Hukuk Dergisi", archive of lawyers' magazine, (law, politics), http://www.aciksayfa.com, 12,044 words
- Collection of poems, poems (literature), http://www.basarm.com.tr/yayin/index.html, 106,501 words
- "Alman İdeolojisi", Marx & Engels, (philosophy), 24,967 words
- "Komünist Manifesto", Marx & Engels, (politics) 14,584 words
- "Akademya Dergisi", articles on language and thought (philosophy, social sciences), http://www.geocities.com/akademyayadogru/makkonulist.htm, 78,488 words
- Collection of articles from Hacettepe University, History Department , (history), 58,960 words
- "Milliyet Gazetesi", newspaper archive (news), 305,805 words
- Collection of newspaper articles of Eruğrul Özkök, http://www.hurriyet.com, 129,086 words
- Collection of newspaper articles of Murat Belge, http://www.radikal.com.tr, 80,645 words
- Collection of newspaper articles of Perihan Mağden, http://www.radikal.com.tr, 24,719 words
- Collection of newspaper articles of Oktay Ekşi, http://www.hurriyet.com, 89,682 words
- Collection of short stories, http://www.efsaneler.com, 74,145 words
- Dergibi Dergisi", essays, poems, stories (literature), http://www.dergibi.gen.tr, 185,641 words
- "Cumhuriyet Gazetesi", newspaper archive (news), 415,830 words
- "Hürriyet Gazetesi", newspaper archive (news), http://www.hurriyet.com, 210,496 words
- Collection of articles on banking, (economy), http://www.tcmb.gov.tr, 87,603 words

- Collection of articles on labor, (economy), http://www.ceterisparibus.net/calisma.htm, 58,492 words
- "Hürriyet Gazetesi, Bilim Eki", (popular science), http://www.hurriyet.com, 78,136 words
- Collection of articles on science, engineering and information technologies, (science and engineering), http://www.makaleler.8m.com, 134,573 words
- Collection of course notes on pediatry, (medicine, pediatry), http://lokman.cu.edu.tr, 282,025 words

REFERENCES

Ando, R. K. and L. Lee, 2000, "Mostly Unsupervised Statistical Segmentation of Japanese: Applications to Kanji", *ANLP First Conference of the NAACL*, pp. 241-248.

Carter, D., J. Kaja, L. Neumeyer, M. Rayner, W. Fuilang and M. Wirén, 1996, "Handling Compound Nouns in a Swedish Speech-Understanding System", *in Proceedings of ICSLP 96*, Philadelphia, USA.

Chomsky, N., 1969, "Quine's Empirical Assumptions", In Davidson, D. and J. Hintikka, (Eds.), *Words and objections. Essays on the Work of W. V. Quine*, pp. 53-68. D. Reidel, Dordrecht.

Clarkson, P. R. and R. Rosenfeld, 1997, "Statistical Language Modeling Using the CMU-Cambridge Toolkit", *Proceedings of EUROSPEECH 97*, Rhodes, Greece.

Çarkı, K., P. Geutner, and T. Schultz, 2000, "Turkish LVCSR: Towards Better Speech Recognition for Agglutinative Languages", *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-2000*, Istanbul, Turkey, June.

Çetinoğlu, Ö., 2001, "A Prolog Based Natural Language Processing Infrastructure for Turkish", M.S. Thesis, Boğaziçi University.

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

Jelinek, F., 1997, *Statistical Methods for Speech Recognition*, The MIT Press, London, England.

Jurafsky, D. and J. H. Martin, 2000, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, New Jersey.

Kanevsky, *et al*., 1998, “Statistical Language Model for Inflected Languages”, US patent No: 5,835,888,1998.

McCrum, R., W. Cran and R. McNeil, 1992, *The Story of English*, Penguin, New York.

Mengüşoğlu, E. and O. Deroo, 2001, “Turkish LVCSR: Database preparation and Language Modeling for an Agglutinative Language”, *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-2001 Student Forum,* Salt Lake City, May.

Oflazer, K., 1994, “Two-level Description of Turkish Morphology”, *Literary and Linguistic Computing*, Vol. 9, No. 2.

Rosenfeld, R., 2000, “Two Decades of Statistical Language Modeling: Where Do We Go From Here?”, *Proceedings of the IEEE*, Vol. 88, pp. 1270-1278, August.

Shannon, C.E., 1948, “A Mathematical Theory of Communication”, *Bell System Technical Journal*, Vol. 27, pp. 379-423 and 623-656.