LINGUISTIC MODELING FOR AUTOMATIC SPEECH RECOGNITION IN ROMANIAN LANGUAGE

Dan CIREŞAN[1], Cosmin CERNĂZANU[2]
"Politehnica" University of Timişoara
Vasile Pârvan Street, no 2, RO 300223 Timişoara
cdanc@cs.utt.ro

Abstract. *This paper experimentally deal with the problem of linguistic modeling for automatic speech recognition in Romanian language, with possible applications in automatic transcription of radio-tv news and in human-computer vocal dialogue automatic system. After a summary survey on theoretic bases of linguistic modeling, there will be described the developed and utilized software instruments for experimenting purpose. Next will be presented a series of experiments that, from our knowledge, represent the first evaluation of linguistic modeling possibility for automatic speech recognition in Romanian language using very large vocabularies (thousands of words). The results indicate the possibility of creating adequate linguistic models for the transcription of radio-tv news using newspaper text corpora collected from Internet, as well as the necessity of further research in linguistic modeling for human-computer vocal dialogue automatic system.*

Key words: *automatic recognition, linguistic models, vocabulary, perplexity*

Introduction

From the apparition of electronic computers, the interface between human and computer spectacularly evolved. At present, many types of interfaces exist, but most of them involve direct physical contact. Teleinterfaces category also contains speech recognition interfaces. Sustained efforts are made to develop an interface suitable to recognize the human speech.

Speech recognition is based on mathematical models, namely acoustic models and linguistic models (figure 1). Acoustic models are used for recognition at the level of acoustic units and linguistic models are used to complete the acoustic models in order to achieve word recognition.

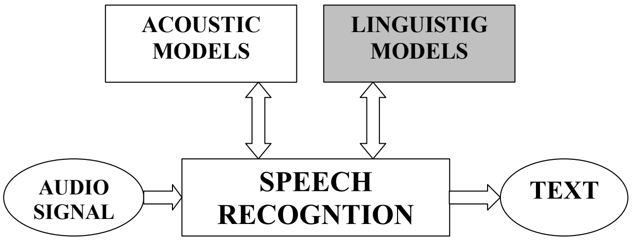


Figure 1. System for converting speech to text

This paper is organized as follows: Section 2 gives a very short survey on the results of mathematical theory of linguistic models. Perplexity will be defined and there will be presented the main estimation theories followed by the interpolating and reduction methods: linear, Witten-Bell and Good-Turing.

Section 3 describes the software instruments used for linguistic modeling, utilities for normalizing the text and the correspondent graphical interface.

Section 4 is the most important part of the paper. There the results of the experiments are presented and the use of linguistic models for automatic speech recognition in Romanian language is evaluated.

Finally, in section 5, the principal conclusions are drawn and, in section 6, the future research paths are enumerated.

Linguistic models

In this section, some aspects from mathematical theory of linguistic models will be presented insisting only on definitions and classifications. A detailed description, which is not the subject of this paper, can be found in [1], [2] and [7].

Automatic speech recognition (ASR) tries to recognize the words sequence $w_1...w_N$ from acoustic observations sequence $x_1...x_T$. That sequence $w_1...w_N$ is chosen for what the conditioned probability $p(w_1...w_N \mid x_1...x_T)$, i.e. the conditioned probability to observe the words sequence $w_1...w_N$ from the acoustic sequence $x_1...x_T$, is maximal. Also, $p(w_1...w_N)$ represent the probability of the words sequence $w_1...w_N$ to occur.

Linguistic models are stochastic models; their purpose is exactly the computing of the probability $p(W_1^N)$ for a words sequence $W_1^N = w_1...w_N$. Using the conditioned probability we get the decomposition: $\Pr(W_1^N) = \prod_{t=1}^{N} p(w_t \mid h_t)$, where $h_t = w_1,...,w_{t-1}$ is the *context* of word w_t. Theoretically, the prediction of the next word will be done with maximum probability if the context is composed of many words, but, practically, the context is formed from one, two or three words.

An n-gram is defined as a sequence of words $w_1...w_N$. In such an n-gram the context is $h_N = w_1...w_{N-1}$ and the predicted word is w_N.

To evaluate the quality of a linguistic model (LM), an entire recognition experiment must be run. Fortunately, the LM can be separately evaluated by measuring, for example, their capacity to predict words in a text. The most used measure of performance is the so-called *perplexity*.

The perplexity is defined as $PP = [p(w_1...w_N)]^{-1/N}$. After decomposing the probability and applying the logarithm we get: $\log(PP) = -\frac{1}{N}\sum_{n=1}^{N} \log p(w_n \mid h_n)$.

The perplexity can be viewed like a function with two arguments: a linguistic model and a sequence of text. From this point of view, there are two types of perplexity: training perplexity computed for the training text corpus and test perplexity corresponding to test corpus. The training perplexity measures how well the LM explains the training corpus and the test perplexity expresses the generalization capacity of the LM in forecasting words in a new text.

From the recognition viewpoint, the LM is said to reduce the number of possible words that can be chosen in recognition process. Therefore, the perplexity can be interpreted as the average number of alternative words that can be chosen in the recognition process. As a first approximation, the perplexity measures the difficulty of the recognition process: a lower perplexity represents a lower error rate.

Another important aspect is the size of the vocabulary. Because the number of n-grams is exponentially growing with rising of the vocabulary size, a large vocabulary will rise too much the computationally resources needed to estimate the n-grams probabilities. Even if we have a training corpus with large vocabulary we can reduce the vocabulary size by including in it only the words that appear at least of k times (cut offs), where we choose the k in such a manner to obtain a convenient vocabulary size. This procedure can be used to obtain the lowest out of vocabulary (OOV) rate, i.e. the probability to find unobserved unigrams. Therefore, the estimation of a test perplexity is usually computed with closed vocabulary, i.e. using only n-grams constructed with words from vocabulary.

Linguistic models usually use three types of distribution: the discrete distribution, the multinomial distribution and the symmetric distribution. The interpolation and the reducing of linguistic models are necessary because of the presence of unseen or very rare n-grams.

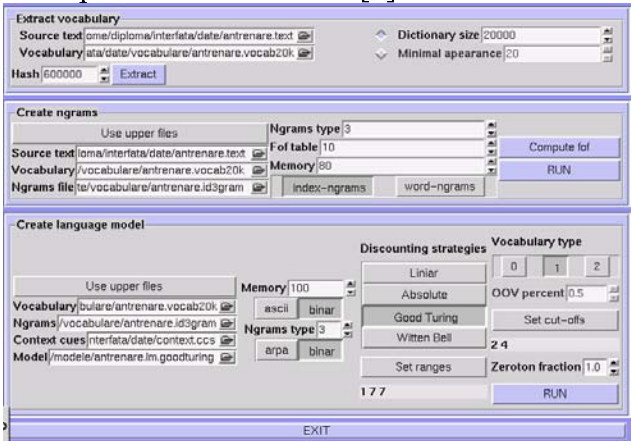
Having an n-gram hw, where h represent the history (the context), the parametric models for conditioned distribution $p(w \mid h)$ is, in general, obtained by combining two components: one for reduction and one for redistribution.

The reduction model is used for solving the unseen events estimation problem (Witten and Bell 1991). The probability of unseen words after a given context h must be estimated by reducing the estimated frequency of n-grams.

In this paper, only three methods of reduction will be used: Good-Turing reduction (Good 1953, Katz 1987), Witten Bell reduction (Witten and Bell 1991) and linear reduction (Placeway 1993).

Software instruments

To collect and normalize the data, two utilities were implemented: *prel* and *concat.* Also, many scripts were used and a graphical interface was built in TCL/TK (figure 2). The detailed description can be found in [7].



Extract vocabulary
Source text ome/diploma/interfata/date/antrenare.text
Vocabulary ata/date/vocabulare/antrenare.vocab20k
Hash 600000
Dictionary size 20000
Minimal apearance 20

Create ngrams
Source text loma/interfata/date/antrenare.text
Vocabulary /vocabulare/antrenare.vocab20k
Ngrams file te/vocabulare/antrenare.id3gram
Ngrams type 3
Fof table 10
Memory 80

Create language model
Vocabulary bulare/antrenare.vocab20k
Ngrams /vocabulare/antrenare.id3gram
Context cues nterfata/date/context.ccs
Model /modele/antrenare.lm.goodturing
Memory 100
Ngrams type 3
OOV percent 0.5
24
Zeroton fraction 1.0
177

Figure 2. Graphical interface

The actual modeling process was realized with the help of software tools from Carnegie Mellon University [4].

Experiments

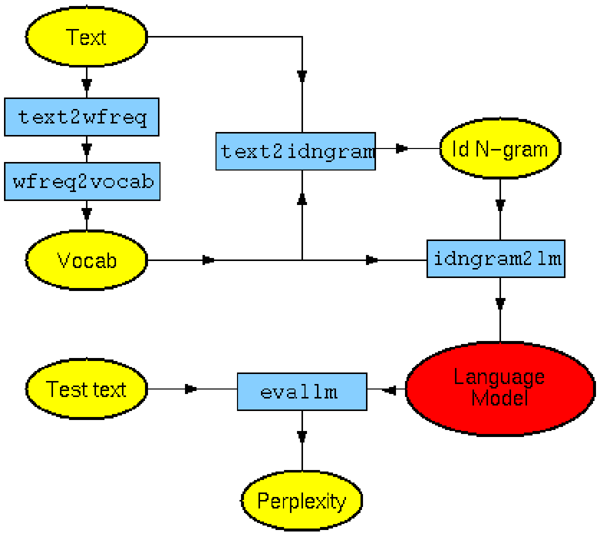


Figure 3. The creation and evaluation phases of a linguistic model [4]

Theoretically, an immense data quantity (of GB order) is essentially for building an ASR model. All this data should be composed of speech transcriptions from various sources to entirely cover the human communication spectrum.

Practically, gathering so much data quantities requests too much time, so alternative methods are used, i.e. Internet.

The experiments will study the possibility to build linguistic models in Romanian language by testing their dependability for ASR, based on measuring their perplexities when they are used on test corpora.

The experiments were made with software tools from Carnegie Mellon University Statistical Language Modeling [4] and tools made by us [7]. A graphical interface was build to facilitate the experimenting process.

The phases of the experiments are:

1. Collecting the text
2. Normalizing the collected texts
3. Splitting the text in two corpora: training corpus and test corpus
4. Extracting useful information from these corpora (vocabulary dimension, number of words)
5. Building the files with id-n-grams
6. Building linguistic models for diverse values of their parameters
7. Evaluating the models by measuring the perplexity over test corpora
8. Drawing conclusions by comparing the results.

The steps for building a linguistic model can be seen in figure 3.

For the conducted experiments, newspaper and magazine archives and manually recorded and transcribed dialogues have been used.

The text sources can be divided in three categories, by the acquisition mode:

- News papers and magazines archives collected from Internet [5]:

  Magazine „22"
  Years 1996-1999
  Size: 8,5 MB
  Address: *http://www.dntb.ro/22/*

  Newspaper „Cronica română"
  No: 1-467
  Size: 40 MB
  Address: *http://domino.kappa.ro/e-media/cronica.nsf*

Newspaper „Dimineața”
No: 1-442
Size: 38,5 MB
Address: *http://domino.kappa.ro/e-media/dimineata.nsf*

- Human computer dialogues
  Size: 45kB
  Source: [5] and [6].
- Transcriptions of radio-tv news collected from Internet
  Size: 0,9MB
  Source: Jurnal TVR1
  Address: *www.tvr.ro/jurnal*

Specific to all these is the fact that they were collected over the years, and that the news is changing every day.

Normalizing the texts

The normalizing task consists of processing the collected texts by removing all unwanted symbols (the text will contain only space separated words) and by adding diacritics, if needed. The normalization is necessary because the data collected from Internet or manually gathered may contain other type of information but texts (i.e. html files contain information that refers to alignment, fonts, and these information are harmful for the scope of this paper).

The normalization was done by transforming the source files to atf files (by segmentation followed by identification of words based on a morpho-syntactic dictionary and then lexically disambiguated)[5]. After that the atf files were converted with our *prel* [7] program to plain text files.

Preparing the training and the test corpora

From newspapers archive, 97% has been allocated to training corpus and 3% to test corpus (randomly chosen). This test corpus will be denoted as test corpus *newspaper*. We also create a test corpus from these data to have a reference when we will test the models on data from other sources.

The text corpus composed of radio-tv news will be denoted as test corpus *news*, and the text corpus composed of human-computer dialogues will be denoted as test corpus *dialogue*. Normalization of these corpora was made using the *grep* command to extract the human part of the text followed by the application of *prel* for including the diacritics and the context identifiers.

Experiment on training corpus

In this case, we are interested in: number of words, size of the vocabulary, cover degree of corpus by different sizes of vocabularies and n-grams number afferent to each vocabulary.

The data obtained after experimenting are presented in table 1.

Table 1. The influence of the size of vocabulary on number of n-grams and on number of Out Of Vocabulary (OOV) rate

Vocabulary size (kwords)	Number of bigrams	Number of trigrams	Number of 4-grams	Training OOV (%)
1	132k	732k	1,83M	35.00
2	251k	1,2M	2,60M	27.34
5	501k	1,98M	3,52M	17.52
10	756k	2,55M	4,01M	11.10
20	1,02M	2,98M	4,28M	6.12
40	1,25M	3,23M	4,40M	2.79
65	1,36M	3,32M	4,43M	1.35

Table 2. The influence of the elimination limits and of the order of n-grams on models size and on perplexity

Cut offs		**Model size**	**Perplexity**
Bigram	**Trigram**		
0	0	24,1 MB	324,1
1	1	7,5 MB	342,97
2	2	5,09 MB	370,04
5	5	3,27 MB	433,54
10	10	2,56 MB	505,06
20	20	2,18 MB	599,66
50	50	1,96 MB	756,30
100	100	1,89	876,01

Because the number of n-grams grows rapidly for large vocabulary and the generation of a single model took over 20 minutes on a 500MHz computer, we tried to build linguistic models with elimination limits (cut offs) for n-grams (i.e. the n-grams that appear less than or equal to a number of times are removed). In table 2 are shown the results obtained with linguistic models created on a 20 kwords vocabulary using the Good-Turing reduction method.

As we expected, though the computing times and the sizes of the models have diminished, the effect on perplexity was negative, and it grew significantly. Consequently, we gave up the cut offs use.

The data from the experiment (text corpora, vocabularies, id-n-grams files and the linguistic models) occupy around 2GB.

The training corpus size is 5062404 words.

Conclusions:

- The lowest OOV rate (1,35%) is obtained for the biggest size of the vocabulary (65 kwords). Further measurements were done and the size of the vocabulary that completely covers the training corpus is 137 kwords, from which approximately 16000 appear at least 20 times.
- The size of trigrams based model is significantly bigger than the size of the bigrams based model (there are 2,4 more trigrams than bigrams), but the number of quadrigrams is only 1,3 bigger than the number of trigrams. An explanation to this fact is that the number of words in a sentence is not much more than four.
- Only large vocabularies (over 20kwords) are good for building linguistic models
- Till now, the best n-grams order cannot be established

Experiments on test corpus *newspaper*

The experiments on this corpus are meant to offer a reference for the next experiments.

As the test corpus and the training corpus both contain texts from identical sources, the results on this corpus will be used as a limit to what we wish to tend our experiments made on the other test corpora.

The test corpus *newspaper* contains 136287 words and the vocabulary size is 8786 words.

Table 3. The influence of the order of n-grams and of the size of vocabulary on perplexity

	Perplexity for a linguistic model based on			
Vocabulary size (kwords)	Bigrams	Trigrams	Quadri-grams	Test OOV (%)
1	129,18	103,78	103.28	34,95
2	161,50	130,01	130.65	28,02
5	225,81	182,33	184.02	18,73
10	297,70	242,53	246.06	12,48
20	391,90	321,89	359.01	7,34
40	489,84	405,30	409.53	3,93
65	545,77	453,09	457.50	2,48

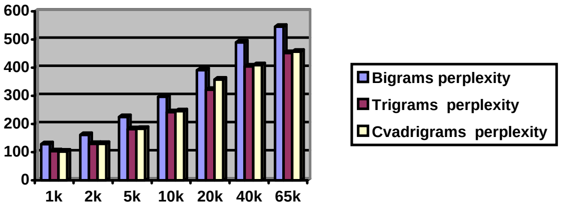


Figure 4. The influence of the order of n-grams and of the size of vocabulary on perplexity

From table 3 and from figure 4 it can be observed that from the three n-grams order that was experimented, the trigrams offer the best linguistic models and that the model based on cuadrigrams is not so good because the training corpus is too small. The LM based on bigrams is also not so good, because on bigrams level the possibilities of continuation after a context made by a single word are to numerous (this fact is specific to all languages of Latin origin, languages that present many inflections).

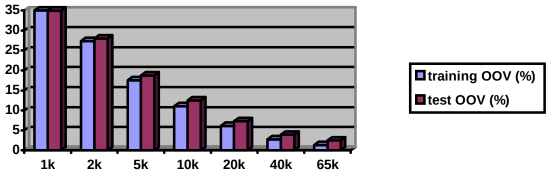


Figure 5. The influence of vocabulary size on OOV rate

As we expected, the test OOV is close to training OOV because both corpora contain text from the same sources (table 1, table 3 and figure 5).

The reduction method influence on perplexity can be observed from table 4, and we can see, that whatever would be the n-grams order, the

Good-Turing reduction method lead to the smallest perplexity, regardless the size of vocabulary. One other observation is that the perplexity increases with the growing size of vocabulary.

Table 4. The influence of the order of n-grams, of the reduction method and of the size of vocabulary on perplexity

	Bigrams			Trigrams			Cuadrigrams		
Reduction method	Vocabulary size			Vocabulary size			Vocabulary size		
	20k	40k	65k	20k	40k	65k	20k	40k	65k
Linear	483	546	615	392	499	560	535	518	581
Good-Turing	391	489	545	321	405	453	359	409	457
Witten-Bell	423	495	556	340	439	496	405	469	527

Experiments on test corpus *news*

After experimenting on test corpus *newspaper* and obtaining results that can be used like an etalon, we will test the LM on tests corpora made from different sources from that of the training corpus for testing the LM prediction capacity on various domains.

Table 5. The influence of the order of n-grams and the size of vocabulary on perplexity

	Perplexity for a linguistic model based on			
Vocabulary size (kwords)	Bigrams	Trigrams	Quadri-grams	Test OOV (%)
1	99,45	73,95	72,55	37,82
2	124,27	91,55	91,28	30,50
5	175,79	129,3	140,71	20,19
10	228,93	170,61	176,02	13,64
20	294,35	221,18	229,67	8,47
40	381,84	290,53	300,18	4,72
65	418,73	319,35	329,48	3,56

The test corpus *news* is composed of transcriptions of radio-tv news collected from Internet. The size of the corpus is 52356 words and the vocabulary is made of 4514 words. After we have studied the comportment of the models created for different sizes of vocabulary on test corpus *news*, we observed that in this case too the trigrams offer the optimal solution for a linguistic model (table 5).

The OOV rate for this test corpus is slightly better because of the different domain sources of the training and test corpora (table 5).

We have done studies on reduction methods on perplexity (table 6).

Table 6. The influence of the order of n-grams, of the reduction method and of the size of vocabulary on perplexity

	Bigrams			Trigrams			Cuadrigrams		
Reduction method	Vocabulary size			Vocabulary size			Vocabulary size		
	20k	40k	65k	20k	40k	65k	20k	40k	65k
Linear	323	425	471	273	362	402	297	390	432
Good-Turing	294	381	418	221	290	319	229	300	329
Witten-Bell	299	390	431	233	312	347	252	334	370

For all the three types of n-grams, the perplexity was smaller (20%) than the minimum perplexity measured on test corpus *newspaper*. These results can be explained by the fact that the news is written with more care and it has an elaboration pattern.

The conclusion for this test corpus was that the created linguistic models are adequate to automatic transcription of the radio-tv news and that the best performance is obtained when the Good-Turing method is used.

Experiments on *dialog* test corpus

The purpose of these experiments is to test the possibility of using created models for automatic transcription of dialogs. The *dialog* test corpus is composed of human-computer dialogs which were registered and then transcribed. From these dialogs only the human speech was kept, because only this is relevant for our purpose.

The text corpus contains 7833 words and the vocabulary is composed of 243 words.

Unlike the test corpus *newspaper*, the perplexity of the test corpus *dialogues* is raising continuously with the n-grams order, regardless the size of vocabulary, here existing no more a local minimum for trigrams. This increase of the perplexity can be explained by the great test OOV. Therefore, for test corpus *dialogues*, the bigrams represents the best solution (table 7).

Table 7. The influence of the n-grams order and of vocabulary size on perplexity

	Perplexity for a LM based on			
Vocabulary size (kwords)	Bigrams	Trigrams	Quadri-grams	Test OOV (%)
1	497,01	495,01	537,35	42,51
2	782,44	840,31	888,82	35,47
5	824,94	896,81	1083,10	33,80
10	1295,46	1366,99	1417,05	26,99
20	1605,28	1706,93	1830,25	25,05
40	2133,50	2307,00	2427,01	20,89
65	2256,90	2430,30	2711,70	19,88

The test OOV is greater than training OOV because the two corpora are made from very different sources (figure 6).

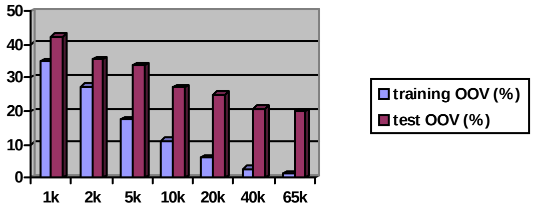


Figure 6. The influence of vocabulary size on OOV rate

For easing the comparing task, we put in the figures 7, 8 and 9 the minimum perplexities measured on test corpus *newspapers*.

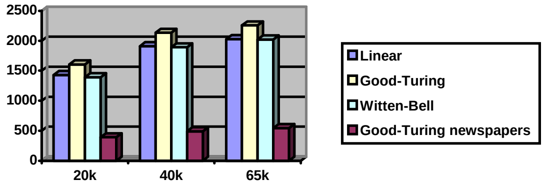


Figure 7. The influence of the reduction method and of the vocabulary size on perplexity for a LM based on bigrams

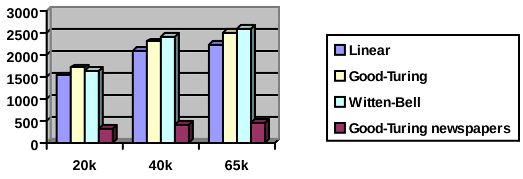


Figure 8. The influence of the reduction method and of the vocabulary size on perplexity for a LM based on trigrams

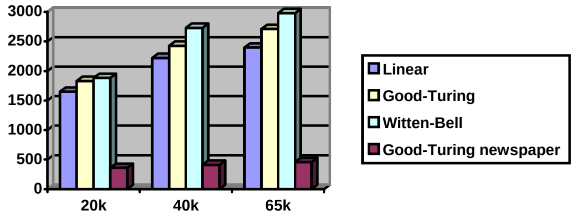


Figure 9. The influence of the reduction method and of the vocabulary size on perplexity for a LM based on quadrigrams

For the present test corpus the linear reduction method was the best, no matter what the n-grams order was used (figures 7, 8 and 9).

For all three cases, the perplexities were much bigger (4 times) than the minimum perplexity measured on the test corpus *newspaper*. These results were expected considering the fact that the sources for the two corpora were completely different. The conclusion is that the linguistic models created so far cannot be directly utilized to transcribe dialogs and they need improvements.

Conclusions

In this paper, we experimentally treated the problem of linguistic modeling for automatic speech recognition in Romanian language, with possible application in automatic transcription of radio-tv news and in human-computer vocal dialogue automatic systems.

The software instruments created for this purpose significantly decrease the experimenting time, by integrating facilities to automatically normalize the data, as well to creating and evaluating the linguistic models.

The conducted experiments represent, from what we know, the first evaluation of the possibility to use linguistic models for automatic speech recognition in Romanian language using very large vocabulary (on the order of tens of thousands of words). From the results of the experiments, there can be drawn important observations on computing power needed and on optimal parameters for creating useful linguistic models. Thus, Good-Turing reduction method and trigrams lead to linguistic models that guarantee a good prediction rate.

The results of the experiments show that it is possible to build linguistic models adequate for automatic transcription of radio-tv news using *newspaper* text corpus collected from Internet, but, at the same time, the necessity of further research on linguistic models for human-computer vocal dialog automated systems.

Future research

We see two directions for future research:

The first consists in improving the development medium by:

1. Integration of instruments for automatic collecting and normalization the data
2. Adding more parameters for creating linguistic models
3. Tools for editing the vocabularies

The second is referring to a new type of experiments:

1. Improving the training corpus by collecting new data from other domains
2. Creating LM usable for human-computer automatic speech recognition
3. Creating and testing linguistic models based on words roots considering the fact that Romanian language present many inflections
4. Integration with other recognition methods.

References

[1] Renato de Mori. (1998) *Spoken Dialogues with Computers.* Chapter 6 – Training of Acoustic Models, Chapter 7 – Language Modeling

[2] Ney H., Martin S., Wessel F. (1997) *Corpus-Based Methods in Language and Speech Processing* – Chapter 6: Statistical Language Modeling

[3] Ousterhout John K. (1994) *Tcl and the Tk Toolkit*

[4] Clarkson P., Rosenfeld R. *Statistical Language Modeling Using The CMU-Cambridge Toolkit,* Eurospeech 1997.

[5] Bohuş, D. (2000) *A Web-based Text Corpora Development System. In Proceedings Second International Conference on Language Resources and Evaluation*, Athena, Greece.

[6] Munteanu, C., (1999) *Mediu pentru dezvoltarea sistemelor de dialog prin metoda Vrăjitorului din Oz. Diploma project.*

[7] Cireşan Dan, *www.cs.utt.ro/~cdanc*