Proceedings of the 5th WSEAS International Conference on Signal Processing, Istanbul, Turkey, May 27-29, 2006 (pp155-160)

Automatic Accentuation of Words for Slovenian TTS System

TOMAŽ ŠEF
Department of Intelligent Systems
Jožef Stefan Institute
Jamova 39, SI-1000 Ljubljana
SLOVENIA
http://dis.ijs.si/tomaz

Abstract: - The accentuation of unknown Slovene words represents a challenging task for automated solvers since in Slovenian, stress can be located on arbitrary syllables. Most words have only one stressed syllable, but there exist also words with no stress and words with more than one stress. Furthermore, different forms of the same word can be stressed differently. In this paper, we present a two level lexical stress assignment model for out of vocabulary Slovenian words used in our text-to-speech system. First, each vowel is determined, whether it is stressed or unstressed, and a type of lexical stress is assigned for every stressed vowel. Then, some corrections are made on the word level, according the number of stressed vowels and the length of the word. We applied a machine-learning technique (decision trees or boosted decision trees). The accuracy achieved by decision trees significantly outperforms all previous results. However, the sizes of the trees indicate that the accentuation in the Slovenian language is a very complex problem and a simple solution in the form of relatively simple rules is not possible.

Key-Words: - Text-to-Speech Synthesis, Natural Language Processing, Lexical Stress Assignment.

1 Introduction

Grapheme-to-phoneme conversion is an essential task in any text-to-speech system. It can be described as a function mapping the spelling form of words to a string of phonetic symbols representing the pronunciation of the word. A major interest of building rule based grapheme-to-phoneme transcription systems is to treat out of vocabulary words. Another applicability of storing rules is to reduce the memory amount required by the lexicon, which is of interest for hand-held devices such as palmtops, mobile phones, talking dictionaries, etc.

A lot of work has been done on data-oriented grapheme-to-phoneme conversion that was applied to English, and few other languages where extensive training databases exist [1]. Standard learning paradigms include error back-propagation in multi-layered perceptron [2] and decision-tree learning [3], [4]. Several studies have been published that demonstrates that memory-based learning approaches yield superior accuracy to both back-propagation and decision-tree learning [5].

Highly inflected languages are usually lacking for large databases that give the correspondence between the spelling and the pronunciation of all word-forms. For example, the authors [6] know of no database that gives orthography/phonology mappings for Russian inflected words. The pronunciation dictionaries almost exclusively list base forms. That is probably the main reason why data-oriented methods were not so popular and that only a few experiments were done for this group of languages.

2 Motivation

It is well known that the correspondence between spelling and pronunciation can be rather complicated. Usually it involves stress assignment and letter-to-phone transcription. In Slovenian language, in contrast to some other languages, it is straightforward to convert the word into its phonetic representation, once the stress type and location are known. It can be done on the basis of less than 100 context-dependent letter-to-sound rules (composed by well-versed linguists) with the accuracy of over 99 %. A crucial problem is the determination of the lexical stress type and position. As lexical stress in the Slovenian language can be located almost arbitrarily on any syllable in the word, it is often assumed to be "unpredictable".

The vast majority of the work on Slovenian lexical analysis went into constructing the morphological analyser [7]. Since the Slovenian orthography is largely based on phonemic principle, the authors of dictionaries do not consider it necessary to give the complete transcriptions of lexical entries. In the only electronic version of Slovenian dictionary, a lexical entry is represented by the basic word form with a mark for the lexical stress and tonemic accent, information regarding accentual inflectional type of the word, morphological information, eventual lists of exceptions and transcriptions of some parts of words. It is assumed that together with the very complex and extensive accentual schemes (presented as a free-form verbal descriptions that require formalization suitable for machine implementation), all the necessary information to predict the pronunciation of the basic word forms, their inflected forms and derivatives is given. The implemented algorithm has around 50,000 lines of a program code and together with the described dictionary allows correct accentuation of almost 300,000 lemmas. This represents several millions of different word forms. A morphological analyser, however, does not solve the problem of homographs with different stress placement and this problem requires stepping outside of the bounds of a separate word.

No dictionary can solve the "stress" problem for rare or newly created words. There exist some rules for Slovenian language, but the precision of those is not sufficient for good text-to-speech synthesis. Humans can (often) pronounce words reasonably even when they have never seen them before. It is that ability we wished to capture automatically in order to achieve better results. Therefore we introduce a two level model that applies the machine-learning methods for lexical stress prediction.

3 Methodology

We use a two level lexical stress assignment model for out of vocabulary Slovenian words. In the first level we applied the machine-learning model (Decision Trees (DT) or boosted DT) to predict the lexical stress on each vowel (and consonant 'r'). In the second level the lexical stress of the whole word is predicted according the number of stressed vowels and the length of the word. If the model (of the first level) predicts more than one stressed vowel, one of them is randomly chosen. If the prediction of the lexical stress of a whole word is false, then typically two incorrect lexical stresses had been made: one on the right syllable (which is not stressed) and the other on the syllable incorrectly predicted to be stressed.

For the first step we generated a domain, were examples were vowels and consonant 'r'. The domain was separated into six domains, one for each vowel and consonant 'r'. For each vowel (and consonant 'r') we trained a separate model (DT and boosted DT) on learning set and evaluate on the corresponding test set. The error was then calculated for the level of syllable and word.

Our goals were as follows: (1) to predict the lexical stress and (2) to see whether there exist some relatively simple rules for stress assignment.

In our experiments we were focusing on accuracy of the models as well as on interpretability. Due to the measure of interpretability, the choice for DT method seems natural, since the tree models could be easily translated into rules.

4 Data

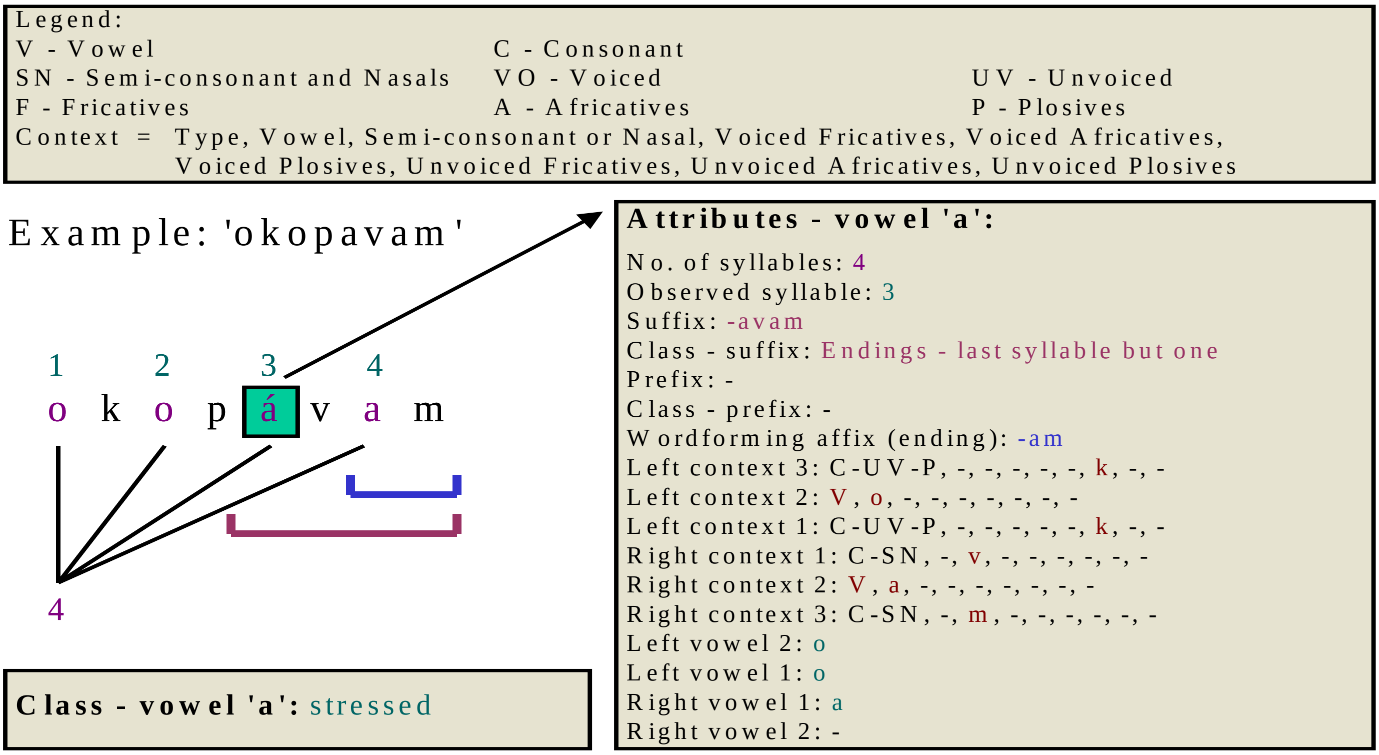
4.1 Data acquisition and preprocessing

The pronunciation dictionaries almost exclusively list base word forms. Therefore a new Slovenian machine-readable pronunciation dictionary was build. It provides phonetic transcriptions of approximately 600,000 isolated word-forms that correspond to 20,000 lemmas. It was build on the basis of the MULTEXT-East Slovene Lexicon [8]. This lexicon was supplemented with lexical stress marks. Complete phonetic transcriptions of rare words, that failed to get analysed by letter-to-sound rules, were also added. The majority of the work has been done automatically with morphological analyser [7]. The error was 0.2 percent. In slightly less than percent additional examination was recommended. Finally, the whole lexicon was reviewed by the expert.

For domain attributes we used 192,132 words. Multiplied instances of the same word-form with the same pronunciation, but with different morphological tags were removed. As the result we got 700,340 syllables (vowels). The corpus was divided in to training and test corpora. The training corpora include 140,821 words (513,309 vowels) and the test corpora include 51,311 words (51,311 vowels). The words (basic word forms, their inflected forms and derivatives) in the test corpora belong to different lemmas than the words in the training corpora. The entries in training and test corpora are thus not too similar. As unknown words are often the derivatives of the existing words in the pronunciation dictionary, the results obtained on the real data (unknown words in the text that is synthesized) would be probably even better than those presented in this paper. Another reason for that is the fact that unknown words are typically not the most common words and in general unknown words will have more standard pronunciations rather than idiosyncratic ones.

4.2 Data description

The training and test corpora were further divided by each vowel and consonant “r”. Thus we got six separated learning problems, one for each letter. The number of examples in each set is shown in Table 1. The class distributions are almost the same in learning and test sets, except for letter ”r”, where a small variance exists.



Legend:
V - Vowel
C - Consonant
SN - Semi-consonant and Nasals
VO - Voiced
UV - Unvoiced
F - Fricatives
A - Africatives
P - Plosives
Context = Type, Vowel, Semi-consonant or Nasal, Voiced Fricatives, Voiced Africatives, Voiced Plosives, Unvoiced Fricatives, Unvoiced Africatives, Unvoiced Plosives

Example: 'okopavam'

Class - vowel 'a': stressed

Attributes - vowel 'a':
No. of syllables: 4
Observed syllable: 3
Suffix: -avam
Class - suffix: Endings - last syllable but one
Prefix: -
Class - prefix: -
Wordforming affix (ending): -am
Left context 3: C-UV-P, -, -, -, -, -, k, -, -
Left context 2: V, o, -, -, -, -, -, -, -
Left context 1: C-UV-P, -, -, -, -, -, k, -, -
Right context 1: C-SN, -, v, -, -, -, -, -, -
Right context 2: V, a, -, -, -, -, -, -, -
Right context 3: C-SN, -, m, -, -, -, -, -, -
Left vowel 2: o
Left vowel 1: o
Right vowel 1: a
Right vowel 2: -

Fig.1: Attributes for the third vowel ('o') of Slovenian word "okopavam" (engl. "I spade up").

Table 1: Number of examples in learning and test sets

	A	**E**	**I**	**O**	**U**	**R**
Learning examples	142041	119227	116486	100295	28104	7156
Test examples	50505	47169	41156	35513	9870	2818

Table 2: Error of DT classifier and Boosted DT classifier

Min. examples in leaves	**1000**		**500**		**300**		**150**		**40**		**2**	
Method	**DT**	**Boosting**	**DT**	**Boosting**	**DT**	**Boosting**	**DT**	**Boosting**	**DT**	**Boosting**	**DT**	**Boosting**
A	19.6	13.9	16.5	10.6	16	9.8	13.7	8.6	10.9	7.1	9.5	6.9
E	24.4	21.8	22.9	19.5	22.3	16.8	20.2	14.3	19.1	11.6	16.1	11.7
I	20.3	16.4	19.2	14.4	17.6	13.0	15.9	11.4	13.9	10.5	11.3	9.5
O	15.3	13.4	15.4	12.3	13.7	11.4	13.3	10.3	11.1	9.1	10.0	8.4
U	20.0	12.8	18.3	12.8	18.0	12.1	14.1	9.7	11.7	8.1	10.6	7.8
R	44.1	28.2	27.5	23.0	28.6	24.4	20.5	23.9	22.1	13.1	14.2	11.9
Syllable	20.5	16.5	18.7	14.3	17.8	12.9	15.8	11.3	13.9	9.5	11.8	9.1
Word	37.4	30.1	34.2	26.0	32.4	23.5	28.8	20.5	25.3	17.3	21.5	16.5

Table 3: Error of grammatical rules [10] on vowels, syllable and word.

	A	**E**	**I**	**O**	**U**	**R**	**Syllable**	**Word**
Grammatical rules	22.8	29.4	22.7	24.0	22.9	33.6	24.7	47.9

Each example is described by 66 attributes including class, which represents type of lexical stress. Its values are 'Unstressed', 'Stressed-Wide', 'Stressed-Narrow', 'Unstressed-Reduced_Vowel', and 'Stressed-Reduced_Vovel'. The factors that corresponds to remaining 65 attributes, are:

- the number of syllables within a word (1 attribute),
- the position of the observed vowel (syllable) within a word (1 attribute),
- the presence of prefixes and suffixes in a word and the class they belong to (4 attributes),
- the type of wordforming affix (ending) (1 attribute) and
- the context of the observed vowel (grapheme type and grapheme name for three characters left and right from the vowel, two vowels left and right from the observed vowel) (58 attributes).

Self-organizing methods for word pronunciation start with the assumption that the information necessary to pronounce a word can be found entirely in the string of letters, composing the word. In Slovenian language placement of lexical stress also depends upon morphological category of the word. It is believed that the string of letters cannot be sufficient to predict placement of the stress. So, to pronounce words correctly we would need the access to the morphological class of a given word. A part of speech information is available in our TTS system with a standard POS tagger even for unknown words but it is not too much reliable for the time being due to the lack of morphologically annotated corpuses (only around 100.000 words). Another reason that we did not include that information into our model (although that would be easy to implement) was a requirement to reduce the size of the lexicon for usage in the hand-held devices. Besides, some morphological information is included in the wordforming affix (ending) of the word and in present prefix and/or suffix of the word.

We achieved better results and more compact models if we represent the context of the observed vowel with the letter type (whether the letter is a vowel or consonant, a type of consonant, etc.) rather then letter itself [9]. The letter itself is indicated in one of the attributes that describe separate letter types (for example, the attribute for a letter type 'vowel' can contain following values: 'a', 'e', 'i', 'o', 'u', '-' (not a vowel)). An example is presented in Fig.1.

5 Experiments

On the six domains, which correspond to five vowel and consonant 'r', we apply DT and boosted DT, as implemented in See5 system [11]. The evaluation was made on separated test sets.

The pruning parameter was minimum examples in leaves. We compare DT classifier with the boosted DT classifier for each value of pruning parameter, which varied between 2 and 1000 minimum examples in leaves. The results are presented in Table 2, Fig. 2, Fig. 3 and Fig. 4. We can see that the lowest error was achieved by boosting and almost no pruning (minimum 2 examples in leaves).

As can be expected the machine-learning methods outperform the grammatical rules [10] (shown in Table 3). The error on the level of word of almost un-pruned boosted DT is reduced for 31.4 percent. We can also observe that even the error on a level of word on highly pruned trees is lower than error on grammatical rules. If we observe the error for each letter, in case of pruned trees (min. 1000 examples in leaves) the results of DT are slightly better, except for the letter 'r'. But the results of boosted DT, pruned with the same parameter, show noticeable improvement in accuracy (error was reduced for 5.4 to 10.6 percent). When we look at the accuracy for boosted DT with min. 2 examples in leave, the error reduction is 15.1 to 21.7 percent.

Another observation was that the DT error is similar to the error of boosted DT for all vowels. However, this is not the case for letter 'r'. During testing different pruning parameters we noticed an interesting anomaly for letter 'r'. Although the error increasing by pruning, it is slightly jumping up and down. If the pruning parameter was set to 400 min. examples in leaves, the error was lower (20,2 %) then if we varied this value between 300 and 500. This could be due to the fact, that this domain has significantly lower number of cases or that the distribution in testing set is slightly different then the distribution of the learning set. In other domains the error is monotony decreasing.

When we were reducing the pruning parameter we could see the accuracy is increasing, which means that pruning doesn’t improve the accuracy. It could be due to many factors, such as anomalies in the data, missing information, provided by attributes, etc. With pruning DT the error increased the most for letter 'r'.

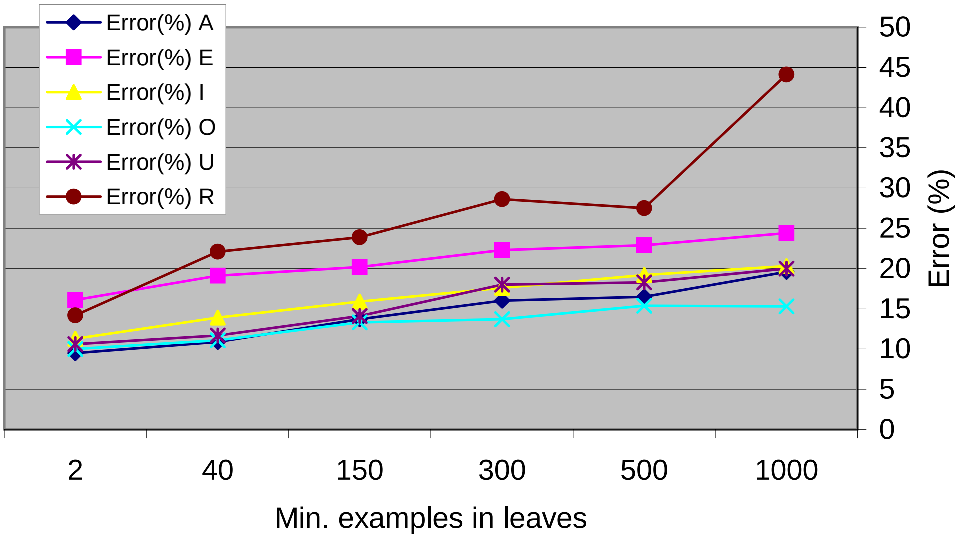


Fig.2: DT error on different pruning parameters.

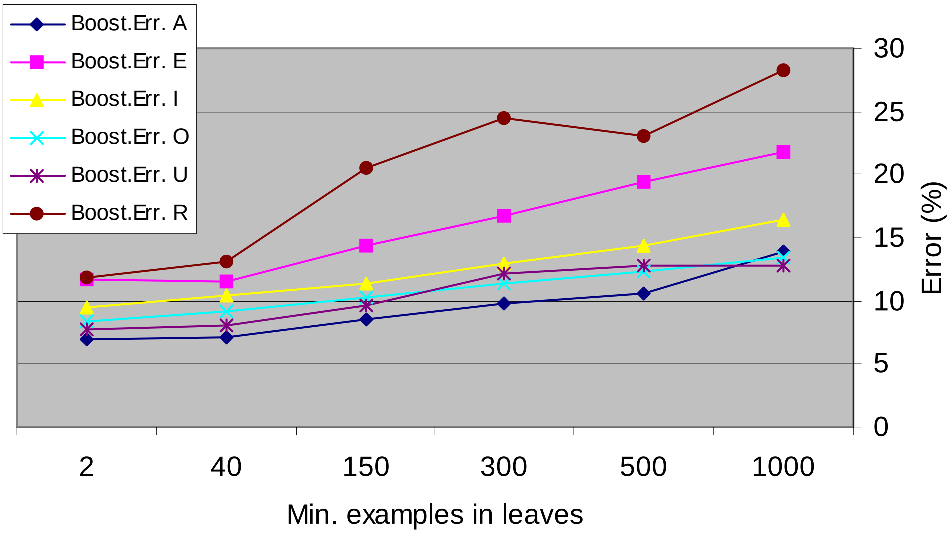


Figure 3: Boosted DT error on different pruning parameters.

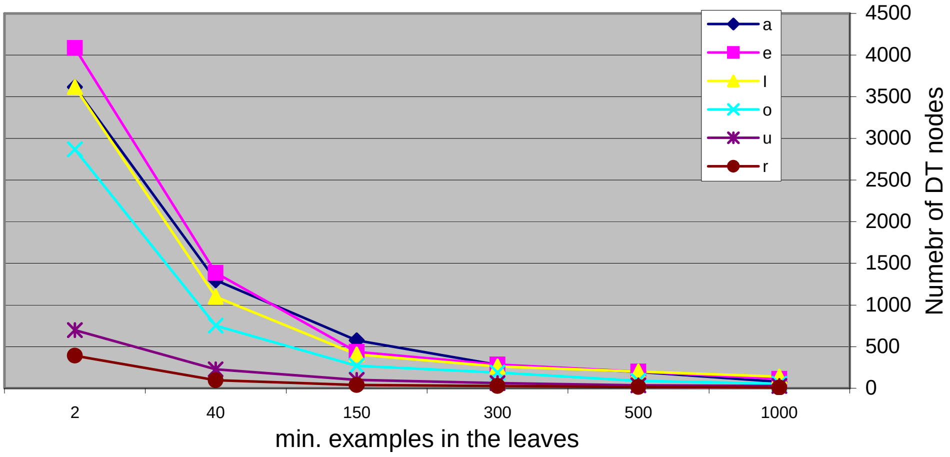


Figure 4: Sizes of DT from Table 3.

Regarding interpretability of the models, we could point out some of the characteristics of the DT: (1) they are very wide, (2) very deep, (3) the top nodes of the trees stays the same, when the trees are growing. Regarding (1) and (2) we can say, that the DT are very large and hard to interpret, so with these attributes no simple rules can be extracted. The reason why the DT are so wide is in choice of attributes. Majority of attributes are discrete and have many values. For all vowels the structure of the trees is stable. The top nodes stay the same. The exception is again the letter 'r'.

6 Conclusion

In this paper we apply the machine learning technique – decision trees on data upgraded Slovenian dictionary in order to: (1) improve the accuracy of defining the lexical stress mark and (2) to establish whether exists some relatively simple rules for accentuation.

The results show that both machine-learning techniques, DT and boosted DT, reduced error of grammatical rules for 26 to 31 percent on the level of word.

Based on our experiments we can conclude that for our data set pruning did not improve the performance. For all tested letters pruning actually increase the error for a couple of percent. We can notice difference in error behavior on letter 'r', which is slightly jumping up and down.

Maybe the most interesting observation is considerable reduction of error in boosted DT. For instance the error on almost un-pruned DT was with boosting reduced for 16 to 27 percent.

Since DT have a large number of nodes and are also very wide, straightforward interpretation was not possible. The pruned trees were substantially smaller, but their error compared to the grammatical rules was only slightly better. The size and structure of the trees indicates that simple or relatively simple rules for lexical stress assignment cannot be constructed on this set of attributes.

References:

[1] W. M. P. Daelemans, A. P. J. van den Bosch, Language-Independent Data-Oriented Grapheme-to-Phoneme Conversion, *Progress in Speech Synthesis*, Springer, pp. 77-89, 1996.

[2] T. J. Sejnowski, C. S. Rosenberg, Parallel networks that learn to pronounce English text, *Complex Systems*, 1:145-168, 1987.

[3] T. G. Dietterich, H. Hild, G. Bakiri, A comparison of ID3 and backpropagation for English text-to-speech mapping, *Machine Learning,* 19:5-28, 1995.

[4] A. Black, K. Lenzo, V. Pagel, Issues in Building General Letter to Sound Rules, *3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 77-80, 1998.

[5] B. Busser, W. Daelemans, A. van den Bosch, Machine Learning of Word Pronunciation: The Case Against Abstraction, *Sixth European Conference on Speech Communication and Technology (Eurospeech'99)*, Budapest, Hungary, 2123-2126, 1999.

[6] R. Sproat (ed.), *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Kluwer Academic Publishers, 1998.

[7] T. Sef, Text Analysis for the Slovenian Text-to-Speech Synthesis system, *PhD Thesis*, Faculty of Computer and Information Science, University of Ljubljana, 2001.

[8] T. Erjavec, N. Ide, The MULTEXT-East Corpus, *First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30, 1998.

[9] T. Sef, M. Gams, “Data Mining for Creating Accentuation Rules”, *Applied Artificial Intelligence*, Taylor & Francis, 18:395-410, 2004.

[10] J. Toporisic, *Slovene Grammar*, Zalozba Obzorja, Maribor, 1984.

[11] J. R. Quinlan, “Induction of Decision Tress”, *Machine Learning*, 1:81-106, 1986.