A Software Tool for Semi-Automatic Part-of-Speech Tagging and Sentence Accentuation in Serbian Language

Milan Sečujski*, Vlado Delić*

* Faculty of Engineering, University of Novi Sad
Trg Dositeja Obradovića 6, Novi Sad, Serbia
{secujski, vdelic}@uns.ns.ac.yu

Abstract

This paper presents a software tool for semi-automatic part-of-speech tagging, annotation of morphological categories and accentuation of texts in Serbian language. The software tool described in this paper is used for very efficient development of tagged text corpora in Serbian language since the accuracy of automatic POS tag and morphological category assignment is 87,2%. This result was obtained by testing the algorithm on a text containing 36692 words, and has turned out to be highly dependent on the type of text. The same algorithm for automatic POS tag and morphological category assignment can be included in text-to-speech systems, enabling correct accentuation of sentences, which, in turn, leads to fairly natural prosody. Within the test mentioned above, accent type and position were determined for each word based on automatically assigned POS tag, morphology-related information, as well as certain syntax cues, and correct accentuation assignment rate of 97,2% was achieved.

Programsko orodje za polavtomatsko oblikoskladenjsko označevanje in pripisovanje stavčnega poudarka v srbskem jeziku

V članku je prestavljeno programje za polavtomatsko oblikoskladenjsko označevanje, pripisovanje oblikoslovnih kategorij in mesta naglasa/poudarka besedilom v srbskem jeziku. V članku predstavljeno programsko orodje je uporabljeno za zelo učinkovit razvoj označenih besedilnih korpusov srbskega jezika; natančnost pripisovanja oblikoskladenjskih oznak je namreč okrog 88 %. Rezultat je bil dosežen s preizkušanjem algoritma na besedilnem korpusu velikosti 36.692 besed, izkazal pa se je za v veliki meri odvisnega od tipa besedil. Isti algoritem za avtomatsko oblikoskladenjsko označevanje je lahko vključen tudi v sisteme pretvorbe zapisanega v govorjeno besedilo, saj omogoča pravilen pripis stavčnega poudarka, ki vodi k precej naravni prozodiji. V zgoraj omenjenem preizkusu je bila na osnovi avtomatično pripisanih oblikoskladenjskih oznak pri ugotavljanju naglasnega tipa in mesta naglasa pri posamezni besedi dosežena natančnost 97,2 %.

1. Introduction

Current methods in language technology rely heavily on the use of large speech and text corpora. Text corpus collection and annotation are costly and time-consuming processes, and their perpetual necessity is the problem which every language community, especially the smaller ones, is facing. Most of the existing text corpora in Serbian language are not annotated. Exceptions include the Corpus of Serbian Language developed at the Institute for Experimental Phonetics and Speech Pathology in Belgrade, containing 11 million manually annotated words (Kostić, 2001), as well as the Serbian translation of George Orwell's "1984", the centrepiece of the MULTEXT-East resources for Serbian, containing about 90000 words (Krstev, Vitas, Erjavec, 2004). However, information related to accent type and position, essential for use of these corpora in high-quality text-to-speech synthesis and automatic speech recognition, is missing. Most of it could be recovered using an adequate dictionary, containing both morphology and accentuation information, however, some of the phenomena related to accentuation cannot be captured in this way (e.g. stressed vs. unstressed personal pronoun forms, falling accent shifting onto the preceding clitic etc.). This is why the AlfaNum team for development of speech technologies at the Faculty of Engineering in Novi Sad, Serbia, opted for developing a software tool for semi-automatic POS tagging and accentuation, with intention to use it for development of a large annotated text corpus.

2. Goal of the paper

This paper presents a possible solution for automatic assignment of POS tags as well as tags related to values of morphological categories and accent type and position to words in Serbian language. The algorithm explained in this paper is used for efficient development of annotated text corpus in Serbian, within a software tool developed for that purpose. Input text is tokenized and annotated automatically, and the visually intuitive software enables very efficient manual correction of errors. In this way a correctly annotated text corpus is developed semi-automatically. At the same time the comparison between initial and manually modified tags gives an estimate of the accuracy of the algorithm and points out the most frequent error types, enabling further improvement of the algorithm.

3. On Serbian language

Owing to significant dependence of natural language processing techniques on target language, some attention should be given to general features of Serbian language before proceeding to specific details of the algorithm.

Serbian language is an Indo-European, South-Slavic language, with 10 million speakers in Serbia (11 million world-wide) (Grimes, 1996). Like other Slavic languages, it exhibits some interesting features that prove challenging to natural language processing technologies.

It exhibits a high degree of inflection – a complete overview of grammatical categories is too complex to be presented here, but the fact that a complete declension of adjectives consists of seven grammatical cases, three genders and two numbers, including suppletive forms for neutral plural as well as separate forms for dual/paucal, can serve as an illustrative example. Derivation with the use of prefixes and suffixes is also quite common, and the word order has significant freedom.

Complexity of morphology in Serbian language makes the dictionary size extremely large, causing well-known sparse data problems to statistically oriented language models based on *N*-grams. This fact together with the significant freedom of word order makes basic *N*-gram based models an unlikely choice for practical use in POS tagging. Since in languages with relatively free word order the information needed for accurate POS tagging lies in morphological categories of the words rather than in word order, a strategy aiming at accurate POS tagging should include a grammatically controlled search in sentence parsing.

Regarding the use of this algorithm within text-to-speech systems, some other features of Serbian language should be taken into account. The accentuation system of Serbian language is rather complex. Serbian is a tonal language, meaning that words possess inherent pitch patterns. Accented syllables are termed either rising or falling, and contain a long or a short vowel. Traditional notation in grammars and dictionaries combines these two features using four accent marks. Complexity of the accentuation system in Serbian aggravates the task of automatic accent assignment since the number of options is much greater than the number of syllables. It is, however, very important that it be as accurate as possible, since in a tonal language such as Serbian many minimal word pairs that differ only in accent can be found. Such minimal pairs can exist both within the inflection of a single lemma, such as in *žèna* (*woman*, n. nom. sg.) vs. *žéna* (*woman*, n. gen. pl.) as well as across different lemmas, such as in *céne* (*price*, n. nom. pl.) vs. *cêne* (*appreciate*, v. pres. 3. pers. pl.)). The problem is further complicated by the fact that vowel timbre can also vary with accent type, and that errors in accent assignment can introduce vowel timbre errors in synthesized speech.

Although lexical accent type does not depend on syntax in general, in some situations grammatical status may be insufficient for correct accent assignment. The most notable cases are when a clitic has a nonclitic homographic alternant, as is the case with some personal pronouns (*nas*, *vas*), when a falling accent shifts onto the preceding clitic (in modern language this can happen when negative particle *ne* is added to a verb form), as well as within certaint frequent collocations.

4. Tagging algorithm

POS tagging process relies on a dictionary containing more than 80,000 lemmas. Since there are many irregularities in inflection of Serbian words (especially nouns and verbs), all the inflected forms were included in the dictionary as separate entries, each of them containing lemmatization and morphology information, as well as information regarding accent type and position. Another reason for including all inflected forms as separate entries was the fact that accent can vary along with inflections of the same word, and that those variations are predictable only to a certain extent. However, the number of entries in such a dictionary exceeds 3 millions. The number of possible combinations of different parts-of-speech and values of morphological categories used in the dictionary exceeds 1200. Whereas tagset size affects the accuracy of automatic tag assignment (Manning, Schütze, 1999), it is still important to keep the tagset complex enough so as to avoid losing important information that may be needed for a particular application such as accentuation of unknown texts (Hladká, 2000).

4.1. Tagging procedure step by step

POS tagging procedure can be divided into several steps. After the initial tokenization of the input text, the words are looked up in the dictionary and a list of all possible POS and morphologic category values that correspond to given inflectional forms is created. In languages with poor inflection, tags usually contain only POS information, whereas in highly inflective languages tags usually contain much more information. The next step consists of context analysis, which considers a word in its context and seeks to determine its tag given the possible tags of neighbouring words. The result of context analysis is a list of words with their corresponding tags, as well as accentuation pattern, which is even more important from the point of view of speech synthesis.

Each of the steps listed above is wrought with difficulties. To begin with, some of the words may not be found in the dictionary, as is the case with many proper names and words including nonstandard affixes. Therefore strategies for assigning correct lemmatization and morphology information must be defined. Some of the strategies for overcoming that problem include making analogies based on standard prefixes and suffixes and rhyming. For example, having failed to find the infrequent word *podleteti* (*v*. to fly under sth.), the system searches the dictionary sorted in “rhyming order” and very soon comes upon the infinitive verb *uleteti* (*v*. to fly into sth.). Knowing that both *pod-* and *u-* are standard verb prefixes, the system will conclude (correctly) that *podleteti* is also an infinitive verb. Such a procedure is not entirely error-free, but it performs well in practice, and provides us not only with morphologic information but with accent type and location as well, since words derived in the same way are likely to possess the same accentuation pattern.

The input data for context analysis consist of a list of possible tags of all words in the sentence. As it would be impossible to consider all tag combinations separately, an algorithm similar to dynamic programming is used, keeping the number of partial hypotheses under control.

Let us consider a sentence $W = w_1 w_2 ... w_N$. Each of the words w_i has a corresponding tag list:

$$T_i = \{t_{i1}, t_{i2}, ... t_{iN_i}\}, \qquad (1)$$

and its actual tag t_i is one of the t_{ij}, $j = 1, 2, ... N_i$. Initially only the hypotheses of length one are considered, containing only the first word of the sentence:

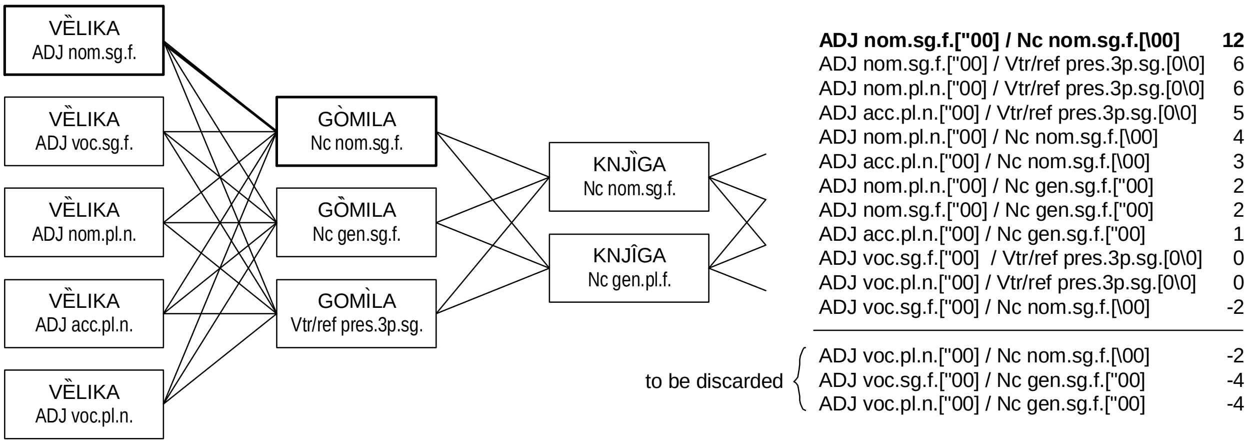
VȄLIKA
ADJ nom.sg.f.
VȄLIKA
ADJ voc.sg.f.
VȄLIKA
ADJ nom.pl.n.
VȄLIKA
ADJ acc.pl.n.
VȄLIKA
ADJ voc.pl.n.
GÒMILA
Nc nom.sg.f.
GȌMILA
Nc gen.sg.f.
GOMÌLA
Vtr/ref pres.3p.sg.
KNJȈGA
Nc nom.sg.f.
KNJÎGA
Nc gen.pl.f.
ADJ nom.sg.f.["00] / Nc nom.sg.f.[\00] 12
ADJ nom.sg.f.["00] / Vtr/ref pres.3p.sg.[0\0] 6
ADJ nom.pl.n.["00] / Vtr/ref pres.3p.sg.[0\0] 6
ADJ acc.pl.n.["00] / Vtr/ref pres.3p.sg.[0\0] 5
ADJ nom.pl.n.["00] / Nc nom.sg.f.[\00] 4
ADJ acc.pl.n.["00] / Nc nom.sg.f.[\00] 3
ADJ nom.pl.n.["00] / Nc gen.sg.f.["00] 2
ADJ nom.sg.f.["00] / Nc gen.sg.f.["00] 2
ADJ acc.pl.n.["00] / Nc gen.sg.f.["00] 1
ADJ voc.sg.f.["00] / Vtr/ref pres.3p.sg.[0\0] 0
ADJ voc.pl.n.["00] / Vtr/ref pres.3p.sg.[0\0] 0
ADJ voc.sg.f.["00] / Nc nom.sg.f.[\00] -2
to be discarded
ADJ voc.pl.n.["00] / Nc nom.sg.f.[\00] -2
ADJ voc.sg.f.["00] / Nc gen.sg.f.["00] -4
ADJ voc.pl.n.["00] / Nc gen.sg.f.["00] -4

Figure 1. An example of a step in the disambiguation algorithm for the sentence “*Velika gomila knjiga stoji na stolu*” (“*A large | heap | of books | stands | on | the table*”). The diagram shows the situation after all the hypotheses of length two are considered, and three of them with lowest scores are to be discarded (in this example stack size limit is *L* = 12).

$$H_1 = \{(t_{11}), (t_{12}), \ldots (t_{1N_1})\}. \qquad (2)$$

In every following step of the algorithm, each variant of the next word is combined with each of the existing partial hypotheses. A set of all hypotheses of length two is thus:

$$H_2 = \{(t_{1m}, t_{2n}) \mid m = 1, 2 \ldots N_1, n = 1, 2 \ldots N_2\}. \quad (3)$$

Each time a new word is appended in such a way, the score of each partial hypothesis is recalculated, based on the likelihood that a word with such a tag can follow. If the number of all hypotheses exceeds a previously set limit *L*, only *L* hypotheses with highest scores are retained, and all the others are discarded. The procedure continues until all words are included and the hypothesis with the highest score is selected as the estimate of actual tag sequence $T = t_1 t_2 \ldots t_N$. Fig. 1 shows an example of such analysis. The algorithm described here performs in time proportional to the length of the sentence, and one of its interesting features is that it produces partial results very quickly. The first word in the sentence is assigned its tag long before the analysis is over, which is consistent with the notion that, when reading a sentence, humans are usually able to start pronouncing it far before they reach its end, and that they organize the sentence into simple prosodic units which can be obtained from local analysis (Dutoit, 1997). Furthermore, this feature of the algorithm is especially useful from the point of view of speech synthesis, because synthesis of the speech signal can start as soon as the first partial results are obtained, which minimizes the delay introduced by context analysis.

The criteria for the actual scoring of the hypotheses are based on rules defined according to the statistics of different parts-of-speech in Serbian language, as found in (Jovičić, 1999), as well as most regular short-range dependencies among them, as found in (Stanojčić, Popović, Micić, 1989). For instance, since it is known that adjectives modifying a noun have to agree in gender, number and case with the noun in question, hypotheses where such pairs adjective-noun occur are considered more likely than hypotheses containing mismatching pairs. Further rules based on dependencies among specific words have also been defined in case it has been proven that such rules could eliminate a significant number of errors still present after the application of an algorithm based on rules of general type only. Some of the templates for rules of general type are as follows:

Award *n* points to a partial hypothesis $h = (w_1, w_2, \ldots w_l)$:

§ If w_l is tagged t_i

§ If w_l is tagged t_i and w_{l-1} is tagged t_j

§ If w_k is tagged t_i, w_{l-1} is tagged t_j and w_{l-2} is tagged t_k

§ If w_l is tagged t_i and w_{l-1} is tagged t_j and the value of a morphologic category *c* contained in the tag t_i is the same (is not the same) as the value of the corresponding morphologic category contained in the tag t_j

§ If w_l is tagged t_i and w_{l-1} is tagged t_j and all of the values of morphologic categories c_1, c_2,... c_k contained in the tag t_i are the same (are not the same) as the values of corresponding morphologic categories contained in the tag t_j

After the (presumably) correct tag sequence has been discovered, the next step consists of modifying accent patterns to account for some words changing their accent type and/or location in a specific context, as described in previous section. If the algorithm is used within a text-to-speech synthesis system, this accent pattern will be used for obtaining a rich prosody structure, defining phoneme durations and variations of fundamental frequency and energy in time.

4.2. Testing the algorithm

Accuracy of POS tagging and assignment of morphological categories and accentuation patterns are of great importance for efficient development of language resources as well as for high quality and naturalness of synthesized speech. If a wrong accentuation pattern were assigned to the sentence, or if there were errors in identification of syntactic units, the resulting f_0 curve would carry misleading prosodic information and a human listener would have trouble recognizing what had been said. The remarkable importance of lexical accent was shown in (Sečujski et al., 2002), where an experiment is described in detail, showing the improvement in human speech recognition caused by introducing accentuation-based prosody into synthesized speech in Serbian language. Twelve listeners were given synthesized sentences corrupted with noise and were asked to recognize what had been said. As many as 83% of the sentences with prosody based on accentuation only were correctly identified at once, compared to 52% of the sentences without any f_0 variations and 31% of the sentences with misleading accentuation. F_0 contours of the sentences were constructed by concatenating and postprocessing initial word f_0 contours based on accent type, location and position relative to a punctuation mark. Speech signal was synthesized by concatenation of prerecorded speech segments selected from a large speech database at runtime, according to (Beutnagel, Mohri, Riley, 1999) using TD-PSOLA model described in (Dutoit, 1997).

It is clearly of interest to establish the accuracy of such an algorithm on a large text corpus. The algorithm described in this paper was tested on a small text corpus containing 3064 sentences (36692 words). The corpus consists of three parts. Part 1 contains 1144 relatively short sentences of general type content (7054 words). Part 2 contains 915 medium length sentences from children's stories (11239 words). Part 3 contains 1008 relatively long sentences from encyclopedic articles (18399 words). Results presented in Table 1 show that 4.03% words in the entire corpus were assigned incorrect POS, 12.77% were assigned incorrect POS or values of one or more morphological categories, and 2.78% were assigned incorrect accent type and/or location. A certain dependence of accuracy on text type was also observed. Results of tests on Part 2 of the corpus were significantly inferior to the others. The subsequent analysis showed that main reasons for this included relatively free word order and somewhat archaic language with frequent use of aorist tense. Aorist forms of a number of verbs are heterophonous homographs to corresponding forms of present tense, and thus errors in morphologic category annotation lead to errors in accentuation. There is a similar relationship between present tense and imperative verb forms, and the number of errors that occur as a consequence depends on the type of text, since imperative forms are rare in encyclopedic articles, but quite often in children's stories.

	Part 1	Part2	Part3	Total
Words	7054	11239	18399	36692
POS	4.58%	4.09%	3.78%	4.03%
POS, morph.	12.94%	13.58%	12.21%	12.77%
Accentuation	2.55%	3.29%	2.55%	2.78%

Table 1: Percentage of error in POS, morphology and accentuation assignment.

5. Conclusion

In this paper a software tool for semi-automatic part-of-speech tagging and sentence accentuation in Serbian language was presented. The rule-based algorithm for initial automatic POS tagging is based on statistics of different parts-of-speech as well as regular short-range dependencies between them as found in available literature. Results of testing this algorithm on a text corpus containing 36692 words show that there is still some room for improvement as regards annotation of part-of-speech tags and morphology-related information. Reducing this error should lead to further reduction of the number of errors related to accentuation, which is of special importance for application of this algorithm within text-to-speech systems. One of the main sources of errors is the lack of more sophisticated knowledge of lexical short-range dependencies. The second one is the inherent inability of such an algorithm to capture long-range dependencies between words, which is one of the main shortcomings of N-gram based algorithms in general. The third and the most difficult one is lack of the "knowledge of the world", which still remains largely unsolved.

6. Acknowledgment

This work was supported in part by the Ministry of Science and Environment Protection of the Republic of Serbia within the Project “Interdisciplinary Research into Speech and Language Resources of Serbian Language”.

7. References

Beutnagel, M., Mohri, M., Riley, M., 1999. Rapid Unit Selection from a Large Speech Corpus for Concatenative Speech Synthesis. In Proc. of EUROSPEECH’99, Budapest, 607-610.

Dutoit, T., 1997. *An Introduction to Text-to-Speech Synthesis*. Dordrecht: Kluwer academic publishers, 149-152.

Grimes, B. F., 1996. *Ethnologue – languages of the world*. SIL International, SIL International.

Hladká, B., 2000. *Czech language tagging*. Ph.D. thesis, Institute of Formal and Applied Linguistics, Charles Univ., Prague.

Jovičić, S., 1999. *Speech communication: physiology, psychoacoustics and perception*. Belgrade: Nauka, 98-103.

Kostić, Đ., 2001. *Quantitative description of Serbian language structure: Corpus of Serbian language*, Institute for Experimental Phonetics and Speech Pathology, Faculty of Philosophy, Belgrade.

Krstev, C., Vitas, D., Erjavec, T., 2004. MULTEXT-East resources for Serbian. In Proc. of IS-LTC, Ljubljana, 108-114.

Manning, C., and Schütze, H., 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, ch. 3.

Sečujski, M., Obradović, R., Pekar, D., Jovanov, Lj., and Delić, V., 2002. AlfaNum System for Speech Synthesis in Serbian Language. In Proc. of 5th Conf. Text, Speech and Dialogue, Brno, 8-16.

Stanojčić, Z., Popović, Lj., and Micić, S., 1989. *Contemporary Serbo-Croatian language and the culture of expression*. Belgrade: Zavod za udžbenike i nastavna sredstva.