Local Grammars in Word Counting  
Duško Vitas, Cvetana Krstev

1. Introduction

The results of word counting in text depend on the level of its linguistic annotation. If a text can is regarded as a sequence of alphabetic character strings, without any information on their possible linguistic interpretation we are talking about the rough text. Some quantitative characteristics of texts can be obtained by the application of formal operations on a rough text, but these results can differ significantly from the results obtained on the annotated text. Moreover, the results of formal operation on an annotated text change with the change in the level and precision of text annotation.

In this paper we consider the conditions under which the rough text can be supplied with the morphosyntactic information that will transform it into a linguistically relevant object. By achieving this goal the examination of text structure, including its quantitative characteristics, becomes possible.

The approach can be illustrated with the following example. If in the rough Serbian text string mora occurs, then it can represent the realization of one of the several lexicon elements: mora.N ‘dread’, more.N ‘sea’, and morati.V ‘to have to’. The results of formal operations will depend on how the equality among the text elements is defined. Namely, on the level of rough text the string mora is the unique text element, on the simplest level of grammatical annotation mora.N and mora.V are two different elements, while on the next level mora,mora.N, mora,more.N and mora,morati.V become three different elements. If we proceed further in this manner we can obtain as much as ten different objects, since mora,mora.N can be the form of nominative singular and genitive plural, mora,more.N can be the form of genitive singular and nominative, genitive, accusative and vocative plural, and mora,morati.V can be the second person singular aorist tense and the third person singular present and aorist tense.

The goal of this paper is to investigate to what extent it is possible to transform a text into a precisely annotated linguistic object by the use of formal methods only. The formal methods used for the processing of Serbian texts are based on the theory of finite state automata (FSA) and transducers (FST). All lexical data, as well as text that is analyzed are represented in this form. We will show how FSTs can be used to obtain the linguistically annotated text that is both rich with linguistic information and precise. The limitations of such approach are not known in advance although the natural languages are essentially finite and can be formally reduced to regular expressions. (see Kornáí 1999).

In the section 2 of this paper we will briefly describe the model of electronic dictionary and its application to Serbian. In the section 3 we will describe some problems that are encountered in attempt to obtain the precisely annotated text and we will outline the suggested solutions. Finally, in section 4 we will give some directions for future work

2 Morphological electronic dictionaries

A morphological electronic dictionary represents the list of simple word forms, that is, the strings of alphabetic characters between two separators that can be realized in a text, accompanied by the corresponding normalized form (usually called lemma) and the list of
the values of the grammatical categories that point the possible relation between lemma and the simple word form (Grzybek 2006). For instance, one line in such a dictionary for Serbian is

\[ \text{zrikavcima, zrikavac.N42+Zool:mp3v:mp6v:mp7v} \]

which states that the lemma zrikavac ‘cricket’ and the list of codes N42+Zool:mp3v:mp6v:mp7v correspond to the simple form zrikavcima. The list of codes N42+Zool:mp3v:mp6v:mp7v describe zrikavcima as a noun (N) of masculine gender (m) marked as animate (v) in dative, instrumental or locative case (3, 6, 7) in plural (p). The number 42 uniquely determines the inflective class to which the lemma belongs (Vitas 1993). The Serbian tagset used in e-dictionaries is explained in (Krstev 2006). The marks following the plus sign additionally specify its syntactic and semantic features (Zool stands for zoological in the given example).

This format of morphological dictionary is known as LADL format (Courtois 1990) and it is compliant to the ISO draft standard of the framework for the morphological annotation (Clément et al. 2005). It should be noted that the morphological dictionaries in this format exist or are being developed for many languages (French, English, Greek, Portuguese, Thai, Korean, Italian, Spanish, Norwegian, Arabic, German), including a number of Slavic languages: Russian, Polish, Bulgarian, Serbian and, recently, Croatian.

The present size of the morphological electronic dictionary of Serbian is given in the Table 1. It can be seen that on the average 13.40 different simple words are associated to each lemma of the general lexica, not taking into consideration possible different grammatical categories of one simple form.

<table>
<thead>
<tr>
<th></th>
<th>lemmas</th>
<th>Simple forms</th>
<th>Simple forms/lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>General lexica</td>
<td>77,347</td>
<td>1,036,408</td>
<td>13.40</td>
</tr>
<tr>
<td>Geographic names</td>
<td>3,337</td>
<td>34,877</td>
<td>10.45</td>
</tr>
<tr>
<td>Personal names</td>
<td>23,767</td>
<td>145,697</td>
<td>6.13</td>
</tr>
<tr>
<td>TOTAL</td>
<td>104,451</td>
<td>1,216,982</td>
<td>11.64</td>
</tr>
</tbody>
</table>

Table 1. The size of the morphological e-dictionary of Serbian in December, 2005

The comparison of morphological e-dictionary of Serbian and corpus of contemporary Serbian (Krstev 2005) has shown that these e-dictionaries cover approximately 98% of running simple word forms (tokens) in the corpus and approximately 62% of different simple word forms (types). However, the simple word forms whose frequency in the corpus is greater than 100 are all covered by the morphological e-dictionaries. The same analysis has shown that approximately 26.5% of all entries in the morphological e-dictionaries are represented in corpus.

As the illustration of the results of the morphological analyzer we will use one example from the novel Ljudi govore by Serbian writer Rastko Petrović. This text has 18,232 tokens (word occurrences), 5,071 of which are different simple word forms. The analyzer recognizes 4,896 of these simple words by applying morphological dictionaries, leaving 175 simple word forms unrecognized. The analyzer associates 5,714 lemmas to the recognized simple forms. The number of recognized simple forms and the number of associated lemmas do not agree because to some simple forms more than one lemma is
associated: for instance, the simple form *morali* can be a realization of two different lemmas, the noun *moral* ‘moral’ (in the nominative plural) and the verb *morati* ‘to have to’ (in the masculine plural form of the active past participle).

For the fragment *Prozori okolnih kuća su sasvim mračni* ‘The windows of the surrounding houses are completely dark’, the analyzer generates the following result:

\{Prozori,prozor.N1:mp1q:mp5q\}
\{okolnih,okoli.A2:aemp2g:aefp2g:aenp2g\}
\{kuća,.N600:fs1q:fp2q\}
\{su,jesam.V575+Imperf+It+Iref+Aux:Pzpi\}
\{sasvim,.ADV\}
\{mračni,mračan.A18:adms1g:aemp1g:aemp5g:adms4q:aems5g\}

This result is sometimes called a *tagged text*, which in this case has not yet been disambiguated. Namely, in the above example, only the word forms *su* and *sasvim* are unambiguous, since to all the other simple word forms more than one set of grammatical categories is associated. The same result can be illustrated by the transducer from Figure 1.

Figure 1. The finite state transducer of the fragment *Prozori okolnih kuća su sasvim mračni*. Each node in the graph corresponds to a simple form and its associated lemma, while the transducer’s output represents simple form’s possible grammatical categories (the codes below the nodes).

The example sentence is obtained as a result of the following query: “Find all sentence segments with the structure: the noun phrase of the form $N_A\, A_{gen}$ followed by an auxiliary verb in a congruent form followed by an adjective phrase of the form $ADV\, A_{nom}$”. This result is obtained by the relatively simple procedure that compares the codes of the grammatical categories between the query and the tagged text. The retrieved segment is the only segment in the analyzed text comprising of 2190 sentences that satisfies the posed query although this structure may be regarded as common.

3 The problems of lexical recognition

3.1 The unrecognized words

The occurrence of unrecognized words represents the first problem in the morphological analysis that is based on lexical recognition. In the text *Ljudi govore* there are 3% of unrecognized words, while in the corpus of contemporary Serbian, as mentioned in section 2, this number is greater, almost 38%. The problem of so called *unknown words* has various sources, besides the obvious ones such as typographical errors, and we may classify them as:

1. All developed dictionaries suffer from insufficiency, both general dictionaries and dictionaries of proper names.
2. Newspaper and journal texts contain quite a number of foreign words, written according to the orthography of the language of origin (for instance, *agency*, *Alexander*, *Apple* from English, *tour de force* from French, *staccato* from Italian, and *tempora mutantur* from Latin).

3. Word forms that are usually covered in no dictionary, such as parts of Internet addresses (www, yu, uk), names of companies (*Adrijaervejz*, *Aeroflot*), and products (*AntiSpam, DinaCard* and *dinakard*, Serbian debit card), etc.

4. Words that usually cannot be found in a dictionary, but that belong either to general lexica or to the sublanguage of some specific domain and are derived from common dictionary entries. Such are, for instance, *autoodgovornost* ‘autoresponsability’, derived from *odgovornost* ‘responsability’ by prefixation, and *ambasadorka* ‘woman ambassador’, derived from *ambasador* ‘ambassador’. Both *odgovornost* and *ambasadorka* are entries in traditional dictionaries and in Serbian e-dictionary as well.

The problems listed in point 1 can simply be solved by improving existing dictionaries, problems listed in points 2 and 3 can be solved by producing new dictionaries, such as dictionaries of common Latin words, and foreign words in general, or otherwise they remain unsolved, while problems listed in point 4 ask for a different solution. The results of derivational processes can be described by a special type of finite state automata, known as morphological grammars (Vitas 2005). These grammars extrapolate the existing morphological e-dictionaries can be applied to the words that remained unrecognized during the morphological analysis (see Figure 2). For instance, *majčica*, the diminutive of *majka* ‘mother’ that is not in the e-dictionary is recognized according to the following principle: an arbitrary string of alphabetic characters is recognized inside the parenthesis **NounB** (in this case *maj*), which is followed by a letter that may have palatalized (in this case it is č obtained from k), and to which the final a is added. If a concatenated string (in this case *majka*) exists as a lemma of a feminine gender noun in a morphological dictionary, and the following characters (recognized in the parantheses **Suf** and **S**) represent, respectively, the diminutive suffix (*ic* in this case) and the inflectional endings for the feminine gender diminutives (*a* in this case), than *majčica* is the diminutive of the noun *majka*.

The morphological grammars significantly enhance the dictionary and enable the recognition of various derivational structures that are difficult to include in the dictionary, such as *pedestdvogodišnjak* ‘52 years old man’, *petospratni* ‘having five floors’, *sedmominutni* ‘lasting 7 minutes’, etc.
Figure 2. The morphological graph that recognizes the inflected forms of the diminutives of the feminine gender nouns of the third declination type.

In some cases the derived forms are ambiguous themselves, for instance, the simple word form Arijetina can be correctly associated with two derived lemmas: one is the augmentative of the noun arija ‘tune’ and the other is the form of the possessive adjective Arijetin of the feminine personal name Arijeta. In other cases some inflected derived forms are ambiguous with some other inflected forms of the lemma they were derived from, e.g. nastavnici is the dative or locative case in singular of the derived form nastavnica ‘woman teacher’ and the nominative and vocative case in plural of nastavnik ‘man teacher’. In both cases, the ambiguity is inherent to the language.

Sometimes the ambiguities from the e-dictionaries are reproduced in the derived forms, offering some erroneous solutions. The word form glasnogovornici is the nominative plural form of the noun glasnogovornik ‘spokesman’ that is derived from the adjective glasan ‘loud’ and the noun govornik ‘speaker’. As govornici is besides being the nominative plural form of the masculine animate noun govornik also the dative and locative singular form of the feminine non-animate noun govornica ‘the speaking platform’, glasnogovornici has been associated two different lemmas, glasnogovornik and glasnogovornica, with different sets of grammatical categories, of which only the first one is correct. The best way to avoid these false ambiguities is to include the derived lemmas in the dictionary – in this case glasnogovornik.

3.2 The compounds

Compounds, as defined in Silberztein (1993), are similar to simple words except that they can contain separators, that is characters that are usually regarded as non-alphabetic. The constituent parts of the compounds are usually simple word forms of the language, as in leptir-mašna ‘bow-tie’ or s vremena na vreme ‘from time to time’, but they need not be, as, for instance, in lož-ulje ‘heating fuel’ lož is not a simple word of the language. In either case, they ask for a special treatment. In the later case, with the recognition of compounds some unrecognized words are avoided, while in the former case the recognition of compounds allows more precise morphological, syntactic and
semantic text tagging. For instance, the meaning of the constituent leptir ‘butterfly’ is changed in the compound leptir-mašna. On the other hand, the lexical recognition of the adverbial phrase s vremena na vreme would yield the following result:

\{s,s.PREP+p2\}+\{s,s.PREP+p6\}
\{vremena,vreme.N324+Ek:ns2q:np1q:np2q:np4q:np5\}
\{na,na.INT\}+(na,na.PREP+p4)+(na,na.PREP+p7)
\{vreme,vreme.N324+Ek:ns1q:ns4q:ns5q\}

After manual or automatic disambiguation (see section 3.3), the same phrase would be tagged as follows:

\{s,s.PREP+p2\}
\{vremena,vreme.N324+Ek:ns2q\}
\{na,na.PREP+p4\}
\{vreme,vreme.N324+Ek:ns4q\}

However, in such tagging, though correct, the adverbial usage of the frozen expression s vremena na vreme is not apparent.

Although the limit between, for instance, noun compounds and free nominal groups is not always easy to establish, in many cases it is clear that the noun compound: (a) has its own rules of inflection; (b) does not allow inserts; and (c) represents a concept that differs from the concepts of its constituents. The compound Ujedinjene nacije ‘The United Nations’ illustrates these points. Namely, only if used in plural and with constituents in given order can it represent the institution (compare to … Još jednom smo pokazali da smo ujedinjena nacija, bez obzira gde se nalazimo ‘Once again we have shown that we are an united nation, no matter where we are’ and … Velika Britanija i Amerika, kao i ostale slobodne nacije ujedinjene su u bolu…”...The Great Britain and America, as well as other free nations are united in their pain…”). Also, when another adjective is inserted in the sequence Ujedinjene nacije between the adjective ujedinjen ‘united’ and the noun nacija ‘nation’ the free nominal group is realized in the text (e.g. Ujedinjene mrežne nacije: Čiji je Internet i ko određuje pravila ponašanja na globalnoj mreži? ‘United network nations: To whom belongs Internet and who makes the rules of the behavior on the global net?’). Finally, the concept Ujedinjene nacije differs from the concept nacija: the first one is according to the English Wordnet (Fellbaum 1998) a social group, more precisely an organization, while the other represents a group of human beings collectively.

When considering the inflection of the compounds two things have to be taken into consideration: (1) each constituent in the compound is inflected according to its specific inflectional rules; and (2) the compound as a whole has its own inflectional rules. This is illustrated by the entry for Ujedinjene nacije in the dictionary of compound lemmas DELAC:


In this example, the inflectional code NC_AXN3 corresponds to the compounds consisting of a noun preceded by an adjective that agrees with the noun in gender, case and number. For the compounds of this class, the noun constituent does not inflect in number, that is, the number of the compound does not change and it always equals to the number of the noun constituent of the lemma entry (number is always plural for the example Ujedinjene nacije). The inflection of the entry itself is performed by the finite
state transducer represented in Figure 3 (Savary 2005). This graph shows that the gender, number and animatness of the compound are inherited from the noun (indicated by the double equal sign in the graph), while the adjective agrees with the noun in gender, number, and case (indicated by one equal sign and usage of same variables $g$, $n$, $c$). The graph also shows that the agreement in case for the adjective depends on the animatness of the noun. Besides that, the graph takes into consideration that only the so called longer, or definite, forms of the adjectives are used in the compounds (indicated by the usage of the fixed values e and d for Det feature). This is not illustrated in the example of *Ujedinjene nacije* since there is no difference in writing between shorter and longer feminine forms. On the other side, the constituent forms themselves are inflected according to the inflectional rules for simple words. The output of this transducer shows that the compound animatness, number, and gender are inherited from the noun constituent.

The application of the inflectional graph NC_AXN3 to the compound *Ujedinjene nacije* yields the following seven inflected forms to which appropriate grammatical information is associated (the text following the slash is a comment):

- Ujedinjene nacije, Ujedinjene nacije.NC_AXN1:fp1q/United Nations
- Ujedinjeneh nacija, Ujedinjene nacije.NC_AXN1:fp2q/United Nations
- Ujedinjenim nacijama, Ujedinjene nacije.NC_AXN1:fp3q/United Nations
- Ujedinjene nacije, Ujedinjene nacije.NC_AXN1:fp4q/United Nations
- Ujedinjene nacije, Ujedinjene nacije.NC_AXN1:fp5q/United Nations
- Ujedinjenim nacijama, Ujedinjene nacije.NC_AXN1:fp6q/United Nations
- Ujedinjenim nacijama, Ujedinjene nacije.NC_AXN1:fp7q/United Nations

These inflected forms, as a part of a dictionary of compound inflected forms DELACF, can be used in lexical recognition and tagging in much the same way as dictionary of simple words. Figure 4 represents text FST after application of this dictionary.
3.3 Lexical ambiguities and their elimination by grammars

The resolution of lexical ambiguity is a prerequisite for many automatic procedures on written texts, such as text indexing, context-sensitive spelling checkers, and syntactic parsing. The reduction of lexical ambiguity can considerably speed up the parsing process by filtering out invalid analyses before parsing (Laporte 2001). The problem is even more serious as many cases of ambiguity are actually the cases of the ‘artificial ambiguity’, since some of the tags associated to the simple words or compounds during the lexical analysis are not felt as ambiguous in context by the human reader. The problem is all the more critical that the tagset is more informative and fine-grained and the e-dictionaries more comprehensive.

The objectives of the lexical disambiguation process, as a step that can precede the word counting is, on one side, to increase the precision of tagging of the words by removing as many as possible of incorrect analyses, and, on the other side, to maintain the achieved recall by never discarding a correct analysis. These objectives can best be met by observing and maintaining the distributional data required for the reduction of lexical ambiguity. These data are elaborated by taking into account underlying linguistic structures, by abstracting general rules from observable facts, and by expressing them in readable form. We will present here and illustrate by some Serbian examples, a formalism for expressing the disambiguation rules that can be expressed by finite state automata and used on a tagged text that is itself represented in the same form. This formalism is called Elimination of Lexical Ambiguities by Grammars, shortly ELAG.

ELAG formalizes distributional and grammatical constraints that are specific to a particular word, particular part of speech or some of its subclasses, or the particular grammatical categories. All these constraints are expressed by using the same tagset that is used by the e-dictionaries for the lexical recognition and tagging. An ELAG rule always comprises of one “if” part and one or several “then” parts. This can be illustrated by one example.

If the beginning of the sentence or the phrase is followed by the simple form ti
Then ti cannot be the enclitic form of the pronoun ti.
In ELAG graph, the “if” part is always signaled and delimited by three boxes with <!>, and each “then” part by three boxes with <=>, as illustrated in Figure 5 that represents ELAG grammar for the stated rule. The boxes with <!> and <=> are used to recognize the structure of the rules: the left and right <!> and <=> are used to mark the beginning and the end of the rule, respectively, while the central <!> and <=> are used as a points of synchronization between the “if” and “then” part. All the other boxes contain linguistic elements that are searched in input sentences when rules are applied to text. In our example, {S} denotes the beginning of the sentence, while <ti.PRO+PrsJB:s3i> represents the form of the personal pronoun ti in the enclitic form of the dative singular, which is also ti. The ELAG rules function as follows: everything that is matched by the “if” part of the rule that is not matched by the “then” part is deleted as the option from the tagged text. Consider the following sentence from the Voltair’s Candide:

Ti nisi ubio nikakvog inkvizitora i prema tebi neće imati nepoverenja. (Engl. as you have not killed an Inquisitor, they will have no suspicion of you)

This sentence’s representation in the form of a FST after applying e-dictionaries is given in Figure 6. It shows that the simple word ti is ambiguous: it can be either the form of the demonstrative pronoun taj (in the nominative case of the masculine plural) or the form of the personal pronoun ti (nominative or vocative case, or the enclitic form of the dative case). After applying the ELAG rule from Figure 5 to this sentence, one of these four possible representations, the enclitic form, is deleted (Figure 7). It should be noted that since the “if” part of the ELAG rule does not mention that ti can be the demonstrative pronoun, this option is unaffected by the rule and it remains in the sentence FST, although in this case it is not realized. It could be removed by some other ELAG rule.

The similar situation appears with the following sentence:

Ti su ih prizori naveli da raspravljaju još više... (Engl. Such sights gave occasion to frequent dissertations;...)

Figure 5. One disambiguation rule in the form of ELAG grammar.

Figure 6. The beginning of the sentence Ti nisi ubio nikakvog inkvizitora... lexically tagged

The similar situation appears with the following sentence:
After applying the ELAG rule, the same three options remain – in this case the demonstrative pronoun is realized, while the personal pronoun option can be removed with some other ELAG rule. This ELAG has no effect when *ti* is not at the beginning of the sentence or the phrase, as illustrated by the following sentence, where the enclitic form of the personal pronoun is actually realized:

Nije *ti* dosta gospodin inkvizitor? Treba i taj prosjek da deli sa mnom? (Engl. The Inquisitor was not enough for *thee*, but this rascal must come in for a share with me?)

It should be noted that disambiguation grammars that rely on local context cannot resolve all ambiguities. It is not surprising as in some cases the full syntactic analysis cannot either, as illustrated by the following sentence:

Ja se ne razumem u rodoslovke, ali ako je istina sxto *ti* propovednici govore, onda smo svi mi jedan drugom rod po ocu.
(Engl. As for me, I do not understand anything of genealogies; but if what *these* preachers say is true, we are all second cousins.)

If *ti* is interpreted as a demonstrative pronoun, the sentence gets presumably the right interpretation, as implied by the given English translation. In this sentence, however, *ti* can be interpreted as the enclitic form of *tebi*, personal pronoun *ti* in the dative case, and then this fragment would have the meaning “but if what preachers say *thee* is true…”.

**Conclusion**

In this paper we have presented the results of the complex text annotation that enables various equivalence relations between the words to be introduces. The results obtained in counting the lexical text units depend on how this equivalence relation is defined. Quantitative text analyses show that one word of two in the written Serbian text, and similar results are obtained for other languages (Senellart 1999), is a constituent of the complex lexical unit that can be identified bu some type of local grammar.

**Bibliography**

Courtois, Blandine; Silberztein, Max (eds.) 1990. Dictionnaires électroniques du français. 
*Langue française* 87. Paris: Larousse


Word Length Studies and Related Issues, series “Text, Speech and Language 
Technology”, volume 31, eds. N. Ide and J. Veronis, Springer, Berlin, Heidelberg

Kornai, A. (1999): Extended finite state models of language, CUP

Krstev, Cvetana; Vitas, Duško (2005): Corpus and Lexicon - Mutual Incompleteness,
Proceedings from the Corpus Linguistics Conference Series, Vol. 1, no. 1, ISSN 1747-
9398 (to appear)

Krstev, Cvetana; Vitas, Duško; Pavlović-Lažetić, Gordana (2006) Resources and 
methods in the morphosyntactic processing of Serbo-Croatian, in in Current Issues in 
Formal Slavic Linguistics, eds. Gerhild Zybatow, Uwe Junghanns, Grit Mehlhorn, Luka 
Szucsich (to appear)

XXIV:1, Amsterdam-Philadelphie : Benjamins, pp. 67-103.

Savary, Agata (2005) Towards a Formalism for the Computational Morphology of Multi-
Word Units in Proceedings of 2nd Language & Technology Conference, April 21-23, 
2005, Poznan, Poland, ed. Zygmunt Vetulani, pp. 305-309, Wydawnictwo Poznanskie 
Sp. z o.o., Poznan.

Senellart, Jean (1999) Outils de reconnaissance d'exressions linguistiques complexes 
 dans des grands corpus. Thèse de doctorat, Université Paris 7

Vitas, Duško (1993) Matematicki model morfologije srpskohrvatskog jezika (imenska 
fleksija), PhD Thesis, Faculty of Mathematics, University of Belgrade, 
December 1993

Vitas, Duško; Krstev, Cvetan (2005) Derivational Morphology in an E-Dictionary of 
Serbian, in Proceedings of 2nd Language & Technology Conference, April 21-23, 2005, 
Poznań, Poland, ed. Zygmunt Vetulani, pp. 139-143, Wydawnictwo Poznańskie Sp. z 
o.o., Poznań