PRACE FILOLOGICZNE

Warszawa 2012 TOM LXIII

DUŠKO VITAS, CVETANA KRSTEV
University of Belgrade

PROCESSING OF CORPORA OF SERBIAN USING ELECTRONIC DICTIONARIES

SŁOWA KLUCZOWE: korpus języka serbskiego, słowniki elektroniczne, przetwarzanie danych, korpusy równoległe, język serbski

KEYWORDS: corpus of Serbian language, electronic dictionaries, data processing, parallel corpora, Serbian language

1. Serbian language resources

Among language resources we distinguish, on the one hand, corpora and, on the other hand, dictionaries and grammars. The construction of dictionaries and grammars is slow, because the largest part of the process of their construction is done manually (Laporte 2009). Their interaction with corpora, however, enables more sophisticated processing that cannot be easily achieved without their support.

The existing Serbian language resources can also be analysed in this light (Vitas et al. 2003b). The paper will provide a description of the manually constructed resources for Serbian, namely, the system of morphological electronic dictionaries and semantic networks followed by a review of some of the Serbian language corpora. The aim is to demonstrate how corpora can be successfully exploited using the high-recall tagging that differs significantly from the mainstream approach that is based on the one-to-one tagging prior to any processing.

2. The system of electronic dictionaries

An electronic dictionary is a dictionary intended for text processing, meaning that it contains the information which makes it able to solve the problems related to the segmentation and morphological processing of text. The electronic dictionary model that proved to be useful for Serbian has been developed while using the dictionary concept formulated as part of the RELEX network. From the standpoint of the formal theory of language, this model is based on the finite state automata theory (*Electronic Dictionaries*, 1989).

The construction of such a dictionary requires precise and thorough classification of the inflectional properties of words, where every string of alphabetical characters delimited by separators (punctuation signs or control characters) is considered to be a word. A word, defined this way, is usually referred to as *simple word*, as compared to a *compound word* that consists of a contingent string of simple words. The notions of simple and compound words are introduced with the aim of providing a precise description of objects appearing in the text during the automatic analysis and they do not have true linguistic equivalents. In terms of these definitions, in addition to the usual ones, the character sequence *ti* in *10-ti* (10th) or *Hong* and *Kong*, for example, qualify as simple words, while the corresponding compounds are *10-ti* and *Hong Kong*. Because of the need to differentiate between simple and compound words, the electronic dictionary is organized into two separate subsystems: a dictionary of simple words, named DELAS and a dictionary of compound words DELAC.

The structure of DELAS resembles that of grammatical dictionaries. Every entry is associated with a code describing its part-of-speech and possibly inflectional, syntactic and semantic properties. Inflectional properties are represented by an inflectional class code that describes inflectional endings and the grammatical meaning assigned to them (Vitas 1997). For example, the class N1 in Serbian consists of a set of endings of the unmarked declension of nouns of the first type, labelled as *inanimate* and its description is given in the form of a regular expression:

<E>/:ms1q:ms4q + a/:ms2q:mp2q:mw2q:mw4q +u/:ms3q:ms7q + e/:ms5q:mp4q + om/:ms6q + i/:mp1q:mp5q + ima/:mp3q:mp6q:mp7q

where the symbol "/" is preceded by the case ending (<*E*> for the zero ending) and after the "/" value of the morphological categories of a word, which is arrived at by adding the ending to the noun belonging to the class N1. The meaning of the above-mentioned codes corresponds to gender (*m* for the masculine gender), number (*s* - singular, *p* - plural, *w* - paucal), case (*1* for the nominative, etc.), while *q* refers to a noun marked as *inanimate* respectively. The following list provides an example of entry processing in a dictionary of the DELAS type:

desetorica ('ten men'), N623+NumN+MG+Pl
eksportovati ('export'), V18+Imperf+Perf+Tr+Iref+DerOvIr
ispod ('below'), PREP+p2
muslinski ('muslin'), A2+PosQ+Mat
otpozdraviti ('return a greeting'), V651+Perf+It+Iref

ponešto ('something'), PRO13+Indef+ProN
sporije ('slower'), ADV+Adj+Comp
prozor ('window'), N1
zanovijet ('broom'), N688+Bot+Ijk

Apart from providing a description of inflectional properties, class codes also indicate parts of speech: *N* is the code for nouns, *V* for verbs, *PREP* for preposition, etc. The information following the part-of-speech code describes syntactic and semantic properties of a lemma. Some of these parameters refer to the following features:

MG – natural masculine gender
Pl – natural plural
DerOvIr – dual forms of the infinitive in -ovati/-irati
Hum – human
Bot – botanical
Ijk – jekavian dialect, etc.

This structure of the DELAS dictionary allows precise and automatic generation of all forms of entries, resulting in another type of dictionary named DELAF. The following lines are an excerpt from this kind of dictionary of forms of entries:

bacili, baciti ('threw, throw'), V951+Perf+Tr+Iref:Gpm
devetih, deveti ('ninth'), A2+Ord:aemp2g:aefp2g:aenp2g
gdekoja, gdekoji ('some'), PRO22+ProA+Indef+Ek:fs1g:..
glavama, glava ('head'), N600:fp3q:fp6q...
očev, očev ('father's'), A1:akms1g:aems4q
plavo, plav ('blue'), A17+Col:aens1g:aens4g:aens5g
prozora, prozor ('window'), N:ms2q:mp2q:mw2q:mw4q
pripoveci, pripovetka ('short story'), N620:fs3q...
tvrđih, tvrd ('harder, hard'), A11:bemp2g:befp2g:...

For example, the meaning of the line *prozora,prozor.N1:ms2q:mp2q:mw2q:mw4q* in a dictionary of the DELAF type is as follows: the form *prozora* is either the genitive case (singular or plural) or paucal of the entry *prozor* ('window') which is a masculine noun, marked as *inanimate*. Other lines are interpreted in a similar manner.

On the basis of the content of the DELAF dictionary, it is possible to perform automatic segmentation of text into words and a morphological analysis by applying the method of lexical recognition (Silberztein 1993). This procedure consists of comparing simple words in the analysed text with the content of the dictionary. If the simple word is found in the dictionary, it will be assigned the information about the possible lemmas and the corresponding grammatical categories.

Let's look at the result of the application of the electronic dictionary method to the following text fragment:[1]

E moj Tirke, ti ode onako gospodski, na crvenom tepihu - i Tadić ti se umusio ispred kapele, i Matija, i Tijanić, i Kapor, i Manjo, tvoj Koštunica se zarozao i ridao kod kuće, jeste, jeste: ti serbez ode, a nas ostavi ovde. Sa njima.

The matching up of this text fragment with DELAF generates a text dictionary, a selection of lines from which is shown here:

crvenom, crven ('red') A+Col:aefs6g:adms3g:adms7g:adns3g:adns7g

Kod, ('at') N:ms1q:ms4q
kod, ('at') PREP+p2
Koštunica, N+NProp+Hum+Last+SR:ms1v
koštunica, N+Bot:fs1q:fp2q
kuće,kuća ('home') N:fs2q:fw2q:fw4q:fp1q:fp4q:fp5q
kuće,kućiti ('home, build a home, set up household') V+Imperf+Tr+Ref:Pzp
Manjo, N+NProp+Unk

ode, oda ('odes, ode') N:fs2q:fw2q:fw4q:fp1q:fp4q:fp5q
ode, otići ('went, go') V+Perf+It+Iref:Pzs:Ays:Azs
onako ('like that') ADV

se ('himself') PAR
se, sebe ('oneself') PRO+PrsJB+Ref:4i
Tadić N+NProp+Hum+Last+SR:ms1v

tvoj ('your') PRO+ProA+Pos:ms1g:ms4q
umusio,umusiti ('dejected, become dejected')V+Perf+Tr+Ref+Der:Gsm
zarozao,zarozati ('blubbering, blubber') V+Perf+Tr+Ref:Gsm

Let us remark first that if a word in the text has the same form as that of the entry, the entry is not cited, (e.g. *koštunica*,.N:fs1q:fp2q). Some simple words in the text can have several entries. For example, the form *ode* can be the genitive case singular of the noun *oda* or the present tense/aorist of the verb *otići*, while the simple word *kuće* can, if context is not considered, be reduced to the lemma *kuća* or the lemma *kućiti*. In the terminology of text processing, this phenomenon is usually referred to as *ambiguity*[2] in lexical recognition. There are 44 simple

[1] Vladimir Jokić: *Poruka iz boce*. "Danas Vikend", 24–25.01.2009, p. XX

[2] This is the phenomenon of the homography of forms which is present in Serbian to a significant degree.

words, 34 of which are different and matching with the dictionary yields 52 different interpretations.

The words that are not found in the dictionary are *unknown words*. The phenomenon of unknown words presents a sensitive problem from the point of view of constructing electronic dictionaries. Namely, an unknown word opens up the issue of the choice of entries when constructing the dictionary. One of the strategies for overcoming the occurrence of unknown words is processing every such new word and entering it in the dictionary. However, this approach leads to excessive and unsystematic expansion of the dictionary, as unknown words are most often possessive or relational adjectives, diminutives, etc. Thus, a more suitable approach would be approximating unknown words, based on the content of the dictionary. Some of these approximations are simple, like the recognition of the superlative:

najraznovrsnijim,raznovrstan ('the most diverse, diverse'). A:cems6g:cemp3g: ...

that is formed by prefixing the *naj*-to the comparative, while inheriting the grammatical codes of the comparative; similarly, recognition of negated adjectives:

neobične,neobičan ('unusual'). A:aemp4g:aefs2g:aefw2g

which will inherit the codes of the adjective *običan*. More complex cases of recognition of the unknown word status are described in Vitas (2007a). The filtering of unknown words can be done using heuristics. One such heuristic method allows that every unknown word consisting of alphabetical characters with a capitalized initial be considered proper noun. The above excerpt from the dictionary yields the following line:

Manjo,.N+NProp+Unk

where the string *Manjo* is determined to be a noun (N) and a proper name (NProp) and that it, unlike other proper names (compare *Koštunica, Tadić*), belongs to the class of unknown words. The application of these methods leaves only one unknown word in the text: *serbez* ('carefreely'), that is, no grammatical information has been assigned to it.

As shown in the example of the text analysed, a dictionary of the DELAF type generates all interpretations contained in the dictionary and consequently a significant number of ambiguities in the sense specified above. Eliminating ambiguity involves including the information about the context. The DELACF dictionary, containing inflectional forms of compound words offers one of the ways of introducing additional information in processing. Building such a dictionary is somewhat similar to building a dictionary of the DELAF type, but it poses problems that are not present in DELAF construction. These include the choice of compound words, the description of their inflection, the analysis of effects in terms of eliminating ambiguity, etc (Krstev et al. 2010). An analysis of the problem of identification of compound words in the corpus and their tagging for classes of compound nouns and adverbs is given in Laporte et al. (2008a), and Laporte et al. (2008b).

In the text we analysed, two compound words, the noun clause *crveni tepih* ('red carpet') and the adverbial clause *kod kuće* ('at home') are recognized on the basis of the content of DELACF. The ambiguity that is eliminated by the introduction of DELACF can be comprehended by analysing the results of text tagging at the level of DELAF only. The results of processing at the DELAF level for the string *crvenom tepihu* can be seen in the following excerpt:

crvenom,crven ('red'), A+Col:aefs6g:adms3g:adms7g:adns3g:adns7g
tepihu,tepih ('carpet'), N:ms3q:ms7q

where the simple word *crvenom* can have five different grammatical meanings (dative or locative case, masculine or neuter gender singular or the instrumental case singular, feminine gender). By stating that *crveni tepih* is a compound word in a dictionary of the DELAC type, just two out of five possible interpretations are retained for the reasons of congruence. Therefore, the corresponding string from the DELACF dictionary is as follows:

crvenom tepihu,crveni tepih ('red carpet'). N+AXN+Comp+Conc:ms3q:ms7q

The process of defining inflectional properties of compounds using DELAF has been described in (Krstev et al. 2006). This formalism is based on the unification principle and uses a system of finite state transducers. As a rule, the information necessary for the construction of a DELAC dictionary is not featured in traditional dictionaries.

In addition to these two types of dictionaries, whose content resembles that of their traditional counterparts, local grammars are used in the system of electronic dictionaries for describing certain phenomena. They make it possible to describe the structure and assign grammatical meaning to strings of simple words that cannot be explicitly cited in a dictionary. Some examples of local grammars are those used for recognition of numerals written as words, (Krstev et al. 2007), dates in texts (Krstev et al. 2008a) or grammars for compound tense recognition (Vitas et al. 2003a).

Dictionaries of proper names are a special subsystem in the system of electronic dictionaries, namely:

- a dictionary of given names and surnames initially excerpted from the 1993 list of residents of Belgrade;
- a dictionary of transcribed English names;

- a dictionary consisting of geographical entries, excerpted based on the content of a school atlas and the register of settlements in the former Yugoslavia.

These dictionaries are in the same format as DELAF, the difference being that in them the properties describing the semantics of given names are added after the part-of-speech code. For example, in the line:

Amsterdama,Amsterdam (*Amsterdam*).N+NProp+Top+Gr:ms2q

the attribute *NProp* indicates that the item is a proper name, *Top* - toponym, *Gr* – grad (town), and so on. A subsystem that automatically derives and incorporates derivatives into the dictionary has been described and developed for compound proper names. For example, the relational adjective *novosadski* and the noun *Novosađanin*, referring to the resident of the town are incorporated into the dictionary, along with the toponym *Novi Sad*. This subsystem is described in Utvić (2008). The role of these additional dictionaries is especially important for reducing the number of unknown words during the analysis.

The semantic markers, described earlier are partially integrated by transmitting information from a semantic network of the WordNet type for Serbian (Krstev et al. 2004).

The system of electronic dictionaries of the DELAS type currently includes around 127,000 simple word entries (87,000 as part of the general dictionary and 30,000 proper names) and around 6,000 compound word entries.

Apart from the system of electronic dictionaries, two multilingual semantic networks have been developed for Serbian. The first, whose structure partially corresponds to that of Prinston's WordNet, has been developed as part of the BalkaNet project. This network connects lexicalized concepts represented by synsets in a complex graph whose nodes are linked via semantic relations. The Serbian WordNet contains around 17,000 nodes and for the most part conforms to the corresponding parts of the Princeton WordNet. A small part of this network represents Balkan-specific concepts that have no lexical counterparts in the English WordNet (Krstev et al. 2008b).

The other semantic network describes semantic relations between proper names (Vitas et al. 2007c). A network describing conceptual and linguistic properties of proper names in several European languages has been developed as part of the Prolex[3] project. It features around 2,000 proper names (the names of the contemporary countries and their respective capitals).[4]

[3] http://www.cnrtl.fr/lexiques/prolex/

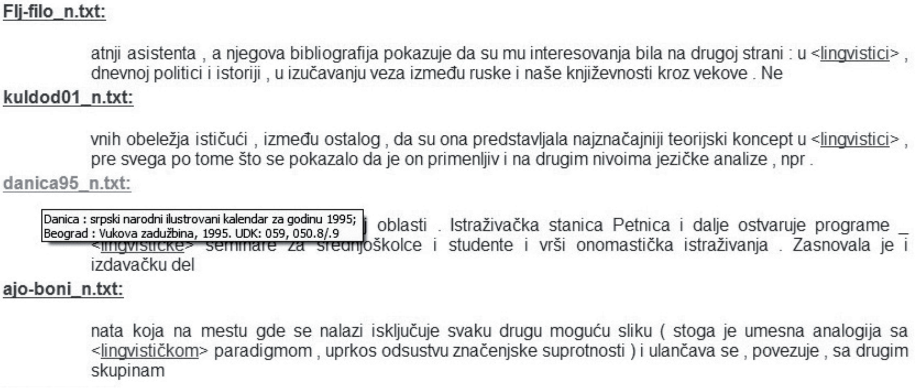
[4] The synonyms, if any, were also quoted (*Pariz ~ grad svetlosti*) ('Paris – the city of light'*)*, the relational adjective (*pariski*), residents, etc.

3. The corpus of contemporary Serbian

The corpus of contemporary Serbian was established in 2002, at the initiative of Professor Ljubomir Popović (Popović et al. 2003), with the aim of enabling researchers to consult the chosen collection of texts in Serbian via the Internet.[5] The part of the corpus available on the web includes around 25 million words,[6] while the entire material prepared for the corpus is four times that size. The main problem in expanding the corpus concerns the issue of finding ways of maintaining balance between different functional styles, especially between newspaper articles and other kinds of text. In the version available on the web, newspaper articles make up 50 % of the material, literary works (including translated ones) about 15 % and monographic publications 17 %.

The software used for searching the corpus is IMS CQP Workbach[7] (Christ 1994), while the online interface was developed at the Faculty of Mathematics of the University of Belgrade. The corpus is used by 220 Slavists all over the world.

The corpus can be searched by using the extended syntax of regular expressions. A snapshot of a screen listing concordances for the search query *lingvist[a-z]+* (i.e. all occurrences of the words beginning in *lingvist* ('linguist') regardless of the string of alphabetic characters that follows it) is given in Figure 1.



```
Flj-filo_n.txt:

            atnji asistenta , a njegova bibliografija pokazuje da su mu interesovanja bila na drugoj strani : u <lingvistici> ,
            dnevnoj politici i istoriji , u izučavanju veza između ruske i naše književnosti kroz vekove . Ne
kuldod01_n.txt:

            vnih obeležja ističući , između ostalog , da su ona predstavljala najznačajniji teorijski koncept u <lingvistici> ,
            pre svega po tome što se pokazalo da je on primenljiv i na drugim nivoima jezičke analize , npr .
danica95_n.txt:

   Danica : srpski narodni ilustrovani kalendar za godinu 1995;
   Beograd : Vukova zadužbina, 1995. UDK: 059, 050.8/.9
                        j oblasti . Istraživačka stanica Petnica i dalje ostvaruje programe _
            <lingvističke> seminare za srednjoškolce i studente i vrši onomastička istraživanja . Zasnovala je i
            izdavačku del
ajo-boni_n.txt:

            nata koja na mestu gde se nalazi isključuje svaku drugu moguću sliku ( stoga je umesna analogija sa
            <lingvističkom> paradigmom , uprkos odsustvu značenjske suprotnosti ) i ulančava se , povezuje , sa drugim
            skupinam
```

Figure 1. *An example of concordances (in square brackets – the search query; in the left column, the abbreviation of the name of sources; in the square: full reference to the source)*

[5] The corpus is available at: http://www.korpus.matf.bg.ac.rs/korpus (freely available for scientific and educational purposes; username and password required for all users can be obtained on demand)

[6] The new version will be available in August 2011; it will have 113 million words.

[7] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

The distribution of frequencies in the corpus indicates that high frequency layers (simple words with a frequency higher than 50 - around 20,000 such words) make up 82% of all occurrences of simple words in the corpus, but just 5% of the total number of different simple word types. On the other hand, the layers with a low frequency (simple words with a frequency lower than 5 - around 390,000 of them) constitute half of the number of different words in the corpus or just 1.6% of the corpus. An analysis of the relation between an electronic dictionary and a corpus is given in (Krstev et al. 2005).

This corpus is not morphologically annotated. The reasons are twofold. The systems based on the methods of machine learning that assign morphological tags with a certain degree of accuracy are usually used for annotation purposes. Different systems solving the problem of the automatic part-of-speech tagging are analysed in (Popović 2010). The TnT system (Brants 2000) yielded the best results when applied to different text samples (the success rate was approximately 95%). However, the accuracy of all systems analysed drops to just 50-60% when they are applied to the words that were not included in the samples used for program training (unknown words). Since the majority of corpus users are sensitive to errors in tag assignment, this kind of tagging has been omitted at this stage of corpus development.

On the other hand, the same corpus can be processed in a different way, using the Unitex[8] system, available under the GPL licence, which efficiently connects corpus processing to the system of electronic dictionaries and local grammars. From the point of view of current users of the corpus, the inability of this system to be accessed on the web represents a shortcoming. Unlike with a program like TnT, the corpus tagged with Unitex is, in fact, tagged with all the information given in the system of electronic dictionaries. This makes complex queries in the form of either regular expressions or recursive transition networks possible. As an illustration of this approach, let us take a look at several examples of the processing of Andrić's novel *Travnička hronika* (*Bosnian Chronicle*). In the text that contains around 394,000 simple words, Unitex identifies around 26.000 different words, 250 of which belong to the class of unknown words. Let us also examine a number of simple search queries that can be applied to this text.

A query that takes the form <*N+Hum+MG-FG:f*> extracts from the text all nouns referring to persons (+*Hum*), whose natural gender is masculine (+*MG*) and cannot be feminine (-*FG*), which are declined like nouns whose grammatical gender is feminine. As an output, Unitex produces concordances for 143 simple words that

[8] http://igm.univ-mlv.fr/~unitex/. Starting from the version 1.2, the distributions of this system feature a module for Serbian (Latin and Cyrillic) consisting of the text of Voltaire's Candide and the accompanying dictionaries.

satisfy the criteria of this query: *aga* ('aga'), *amidža* ('uncle'), *binjedžija* ('skilful horse rider'), *braća* ('brothers'), etc. For the query <*N+NProp+Top+Gr*>, Unitex lists 606 occurrences of nouns that are toponyms (*Top*) and refer to towns (*Gr*) like *Amijen* ('Amiens'), *Beč* ('Vienna'), *Beograd* ('Belgrade'), *Carigrad* ('Constantinople, Istanbul'), *Jena* ('Jena'), *Jerusalim* ('Jerusalem'). The query <*N+Bot*> extracts the terms related to plants, such as *badem* ('almond'), *bob* ('broad bean'), *bor* ('pine tree'),while the query <*A+Col*> yields adjectives referring to colours like, *beo* ('white'), *bledocrven* ('pale red'), *bledoplav* ('pale blue'), *bledorumen* ('pale rosy'), *crn* ('black'), etc. The pattern <*V:W*> retrieves the concordances of all infinitives used in the novel - 1236 from *alakati* ('shout "Allah"') to *žrtvovati* ('sacrifice') and *žvakati* ('chew'). Each of the above-mentioned patterns consists of tags included in the system of electronic dictionaries. These patterns can be incorporated into more complex ones, which opens up possibilities for complex linguistic queries, for example, if we are looking for strings of pronominal enclitics followed by verbal enclitics (or the other way round), the query make take the form of the graph shown in Figure 2, taking the general model offered in Popović (Popović 1997) as a starting point. Here, the parameters *i* and *h*, refer to enclitic forms of pronouns and verbs.

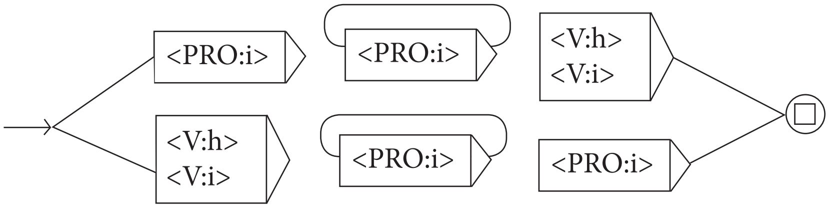


Figure 2. *The formulation of a query that extracts strings of enclitics*

The result of the application of this graph to Andrić's text yields just eight instances shown in Figure 3.

(...) ne brini kako	Će ti se	kmeti pobuniti.
da	joj se nije	što uz put desilo?
- Ah, da	Mi se je	sresti s njime...
	Nije ti ga	majka rodila -
Žalio	Sam mu se	kako je strm (...)
(...) glasovi	su im se	ukrštavali, napregnuti i zatalasani.
Tako	su mi se	u slučajnom preseku ukazale
Oči	su mu se	same sklapale

Figure 3. *The concordances extracted by using the graph shown in Figure 2.*

The nodes of graphs like the one shown in Figure 2 can contain references to other graphs, rather than references to the content of the dictionary, which makes formulation of exceptionally complex corpus queries possible that cannot be expressed in the usual query languages. An example of such complex processing is given in Krstev et al. (2007).

4. Parallel corpora and texts

The term *parallel corpus* refers to two semantically equivalent texts, as a rule, but not necessarily, in different languages, between which a link shared by certain elements of logical layout has been established. Parallel corpora are usually aligned at the paragraph or sentence level. An example of a parallel text[9] in the simplified TMX-format is shown in Figure 4.

```
<tu>
<tuv>Obudziłem się na głos pustelnika, który zdawał się niesłychanie
cieszyć, widząc mnie zdrowego i wesołego.</tuv>
<tuv>Probudio me je glas isposnika koji se izgleda vrlo obradovao što
me vidi zdravog i veselog.</tuv>
<tuv>Probudi me isposnik, veoma zadovoljan što me vidi živa i
zdrava.</tuv>
<tuv>Je fus réveillé par l'ermite, qui parut très content de me voir sain
et sauf.</tuv>
</tu>
```

Figure 4. *Equivalent sentences from Jan Potocki's novel* The Manuscript Found in Saragossa

Serbian is featured in two important parallel corpora. One is the Intera-corpus[10] containing one million words per language, aligned at the sentence level in the TMX format. The registers included in this corpus are education, economy, law and health care. Serbian texts in this corpus are lemmatized at the level of simple words with part-of-speech tagging. The form of lemmatized Serbian texts is illustrated by the following example:[11]

{Kao,kao.ADV} {lideri,lider.N2+Hum} {imamo,imati.V4+Imperf+Tr+It+Iref} {dužnost,.N704} {prema,.PREP+p7} {svim,sav.PRO19+ProA} {ljudima,ljudi.N+Hum} {sveta,svet.A17}

[9] The example has been extracted from the parallel version of Jan Potocki's novel *The Manuscript Found in Saragossa*, translated into Serbian from French (Prosveta, Belgrade, 1964) and from Polish (SKZ, Belgrade, 1988).

[10] http://www.clarin.eu/intera-corpus

[11] An excerpt from the United Nations Millennium Declaration, 2000.

Another important corpus is the French/Serbian literary corpus consisting primarily of a selection of French literary works translated into Serbian (Vitas et al. 2006). This corpus spans the period from the 18th century to the present, as far as the material in French is concerned, while the earliest translation into Serbian dates back to 1926. It features around 1,300,000 French words and 1,100,000 Serbian ones and is aligned at the sentence level. For the majority of texts in the corpus, unambiguous alignment of the source text and its translation was ensured thanks to manual control.[12]

In addition to parallel corpora, a number of individual parallel texts have been built such as Orwell's *1984* – in 12 languages[13] or Verne's *Around the World in Eighty Days* – in 15 languages (Vitas et al. 2008). Among individual parallel texts, there are a number of literary translations into Serbian and Croatian.

Starting from the version 2.0, Unitex also includes processing of parallel texts by using the lexical recognition method.

Parallel corpora are applied not only in contrastive research, but also in the experiments in the field of automatic translation into Serbian (Tufis et al. 2008).

5. The prospects of Serbian language processing

In the near future users of the corpus of contemporary Serbian on the web will be offered the access to 100 million words corpus. This new experimental version of corpus will incorporate tags generated by TnT into user interface. At the same time, all Serbian lexical resources will be further developed, especially dictionaries of multi-word units and proper names. To that end, the development of the workstation LeXimir (old name WS4LR) for manipulation and exploitation of various resources and its incorporation in other applications will continue.

References

Brants T., 2000, *TnT – A Statistical Part-of-Speech Tagger*, [in:] *Proceedings of the 6th Conf. on Applied Natural Language Processing*, ACL, Seattle, Washington, pp. 224–231.

Christ O, 1994, *A modular and flexible architecture for an integrated corpus query system. COMPLEX'94*, Budapest.

[12] This means that the parts of the original that have not been translated and the differences as compared to the original text, if any have been identified.

[13] http://nl.ijs.si/ME/

Electronic Dictionaries and Automata in Computational Linguistics, [in:] *Lecture Notes in Computer Science*, 377, eds. M. Gross, D. Perrin, Berlin–New York 1989.

Krstev C., Pavlović-Lažetić G., Vitas D., Obradović I., 2004, *Using Textual and Lexical Resources in Developing Serbian Wordnet*, "Romanian Journal of Information Science and Technology", vol. 7, No. 1–2, pp. 147–161.

Krstev C., Vitas D., 2005, *Corpus and Lexicon - Mutual Incompletness*, [in:] *Proceedings of the Corpus Linguistics Conference*, eds. P. Danielsson, M. Wagenmakers, Birmingham, http://www.corpus.bham.ac.uk/PCLC/.

Krstev C., Vitas D., Savary A., 2006, *Prerequisites for a Comprehensive Dictionary of Serbian*, "Lecture Notes in Artificial Intelligence", 4139, Berlin–Heidelberg, pp. 552–564.

Krstev C., Vitas D., 2007, *Treatment of Numerals in Text Processing*, in: *Proceedings of 3rd Language & Technology Conference*, ed. Z. Vetulani, Poznań, pp. 418–422.

Krstev C., Koeva S., Vitas, D., 2008a., *A Dictionary-based Model for Morpho-Syntactic Annotation*, [in:] *Proceedings of the 2nd Linguistic Annotation Workshop, in scope of LREC'08*, Marrakech, Morocco, ELRA.

Krstev C., Obradović, I., Vitas, D., 2008b, *An Approach to the Development of Language Specific Concepts in Wordnets*, "Southern Journal of Linguistics, Special Theme: South Slavic and Balkan Languages", Vol. 29, No. 1/2, pp. 106–118.

Krstev C., Stanković R., Obradović I., Vitas, D., Utvić, M., 2010, *Automatic Construction of a Morphological Dictionary of Multi-Word Units*, [in:] *Proceedings of the 7th International Conference on NLP IceTAL*, , LNCS 6233, Reykjavik, pp. 226–237.

Laporte É., Nakamura T., Voyatzi S., 2008a., *A French Corpus Annotated for Multiword Nouns*, [in:] *Proceedings of the LREC'08, Workshop Towards a Shared Task on Multiword Expressions*, Marrakech, Morocco, pp. 27–30.

Laporte É., Voyatzi S., 2008b, *An Electronic Dictionary of French Multiword Adverbs*, [in:] *Proceedings of the LREC'08, Workshop Towards a Shared Task on Multiword Expressions*, Marrakech, Morocco, pp. 31–34.

Laporte É., 2009., *Léxico e gramática: dos sentidos ŕ construção da significaçăo*, http://hal.archives-ouvertes.fr/docs/00/40/09/86/PDF/artesOuIndustr.pdf.

Popović Lj., 1997, *Red reči u rečenici*. Beograd.

Popović Lj., Vitas, D., 2003. *Konspekt za izgradnju referentnog korpusa standardnog srpskog jezika*, „Zbornik Naučni sastanak slavista u Vukove dane", 31/1, pp. 221–227.

Popović Z., 2010, *Taggers Applied on Texts in Serbian*, "Infotheca", 11(2), pp. 21–38.

Silberztein M.D., 1993, *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Paris,.

Utvić M., 2008, *Konačni automati u regularnoj imenskoj derivaciji*. Magistarski rad. Matematički fakultet, Univerzitet u Beogradu.

Vitas D., 1997, *O elementarnoj morfografemskoj klasi*, "Naučni sastanak slavista u Vukove dane", 26/2, pp. 195–206.

Vitas D., Krstev C., 2003a, *Composite Tense Recognition and Tagging in Serbian*, [in:] *Proceedings of the Workshop on Morphological Processing of Slavic Languages: 10th Conference of the European Chapter, EACL 2003*, Budapest, pp. 54–61.

Vitas D., Krstev C., Obradović I., Popović Lj., Pavlović-Lažetić G., 2003b, *Processing Serbian Written Texts: An Overview of Resources and Basic Tools*, [in:] *Workshop on Balkan Language Resources and Tools*, eds. S. Piperidis, V. Karkaletsis, Thessaloniki, pp. 97–104.

Vitas D., 2007a, *O problemu ne(pre)poznate reči*, "Zbornik Matice srpske za filologiju i lingvistiku", 50, pp. 111–120.

Vitas D., Krstev, C., Maurel D., 2007c, *A note on the Semantic and Morphological Properties of Proper Names in the Prolex Project*, [in:] *Lingvisticae Investigationes, Special issue on Named Entities: Recognition, Classification and Use*, eds. S. Sekine, E. Ranchhod, Vol. 30(XXX), No. 1, Amsterdam, Philadelphia, pp. 115–134.

Przetwarzanie języka w serbskich korpusach i słownikach elektronicznych

Streszczenie

Artykuł dotyczy problemu obróbki elektronicznej materiału językowego w języku serbskim, w tym tworzenia słowników elektronicznych i korpusów. Przedstawione są w nim mechanizmy informatyczne używane do opracowania słowników obejmujących słownictwo ogólne i nazwy własne.

Artykuł opisuje też obecny stan prac nad Korpusem Współczesnego Języka Serbskiego i podkorpusami wchodzącymi w jego skład. Ostatnia część tekstu poświęcona jest korpusom równoległym, w skład których wchodzi język serbski, takich jak korpus serbsko-francuski.