Corpus and Lexicon - Mutual Incompleteness

Dr Cvetana Krstev
Faculty of Philology
University of Belgrade
cvetana@matf.bg.c.yu

Dr Duško Vitas
Faculty of Mathematics
University of Belgrade
vitas@matf.bg.c.yu

1 Introduction

The natural language processing group (NLP group) at the Faculty of Mathematics, University of Belgrade is engaged for many years now in a task of producing various language resources, both corpora and lexicons (Vitas et al. 2003). However, in the past our main goal was to produce as many resources as possible in order to try to keep the pace with the so called "big" languages. After producing resources of considerable size we focused our attention to the evaluation of their quality. In order to support this process we performed an experiment by applying the Serbian morphological dictionary to the corpus in order to establish:

a) The extent and content of the corpus lexica that is not covered by e-dictionary. Here we are trying to see what kind of tools have to be developed for the recognition and tagging of unrecognized words such as derivatives, proper names, acronyms, foreign words, etc.
b) The part of e-dictionary not covered by the lexica found in the corpus. We are looking for uncovered lemmas (for instance, to what extent corpus covers the names of zoological species), and uncovered forms (for instance, is imperfect tense really vanishing from contemporary Serbian), etc.

In section 2 we will discuss the structure of Serbian monolingual corpus, its size and accessibility of its part that is presented on web, in the section 3 we will present our Serbian morphological e-dictionary. In section 4 we will present the results of the analysis of the coverage of the corpus by the e-dictionary, while in section 5 we will analyse the coverage of e-dictionary in corpus. Finally, in section 6 we will give some concluding remarks, mainly concerning our future work on the further development of both the corpus and the e-dictionary on the basis of the results presented in this paper.

2 The Corpus of Contemporary Serbian

Many projects have been initiated recently in order to develop reference corpora for the less well-resourced languages, particularly for the languages spoken in former Yugoslavia. Some of these projects started as national projects (Tadić 2002), some were sponsored by the consortium of research and comertial teams (Erjavec 1998) (Krek 2002), and some were initiated as a part of international projects (Santos 1998). Although the activities in corpora constructing for Serbian have been vivid for quite a long period, they have not been officially and financially supported until recently when their importance has been recognized and they have been given some modest support in the scope of some larger national projects. Despite these difficulties the NLP group at the Faculty of Mathematics has achieved some significant results in the construction and usage of both monolingual and multilingual corpora. In this paper we will present only the part of the monolingual corpus that is accessible on the web for on-line searching. This part, known as SrpKor, consists of rough texts, that is, texts in which the logical structure is not marked, and they are not morphosyntactically tagged. The HTML tags have, however, been stripped from the texts downloaded from the web.

Before constructing the corpus of contemporary Serbian certain problems have to solved, for which the experiences and solutions for other languages are of little use. They include, but are not restricted, to the following: (1) regular usage of two alphabets, Cyrillic and Latin; (2) the usage of different encoding schemas for e-texts: ISO 646 IRV, ISO 8859-2 and 8859-5, Windows CP 1250 and 1251, Unicode, to mention only the most frequently used; (3) usage of two pronunciations, Ekavian and Iekavian; (4) the problems to define the scope of Serbian language in the larger community that was once known as Serbo-Croatian.

Latin	č	ć	ž	š	đ	lj	nj	dž
Cyrillic	ч	ћ	ж	ш	ђ	љ	њ	џ
Corpus encoding	cy	cx	zx	sx	dx	lx	nx	dy

Table 1 Internal encoding used for Serbian language resources

In order to neutralize the use of two alphabets, as well as various encoding schemas, the whole corpus is encoded in plain 7-bit ASCII, by encoding the Serbian specific letters by digraphs (Table 1).

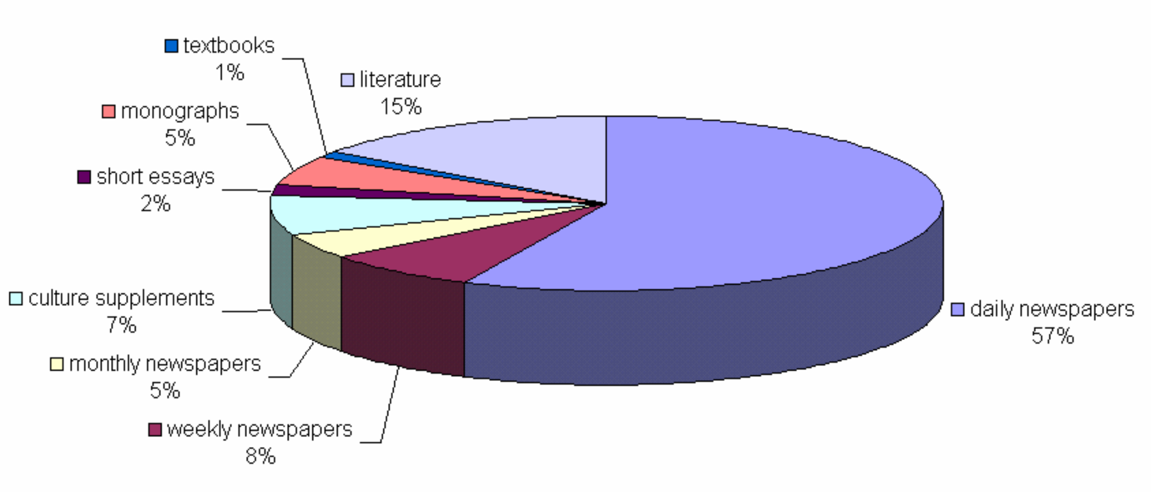

Figure 1 The structure of the corpus of contemporary Serbian accessible on web (http://www.korpus.matf.bg.ac.yu/, authorization required)

The structure of SerKor regarding the text types is given in Figure 1. Newspaper texts date from the 1993, textbooks and monographs date from 1980, while literature part dates from 1920 and it consists of both original and translated works. It consists mainly from the texts published in Belgrade, and therefore the Ekavian pronunciation prevails.

The software IMS Corpus Workbench (CWB) produced at University of Stuttgart is used as a corpus manager (Christ 1994). The web interface was produced at the Faculty of mathematics, and it enables the retrieval using the restricted regular expressions. The concordances obtained as a result of the retrieval are represented in the standard Serbian Latin alphabet, not in the encoding schema used for the internal corpus representation. For instance the regular expression

h?(l|lj|lx)eb(a|u|om|e|ov(i|e|a|ima))

applied on the untagged corpus produces the concordances given in Table 2. This regular expression corresponds to the inflectional paradigm of the noun *hleb* (Engl. bread) together with its pronunciation and dialectic variants. The example shows that although the corpus is in Ekavian pronunciation, occurrences of the Iekavian pronunciation can also be retrieved.

```
pare za to da se vozači kamiona sa hlebom spreče da do prodavnice idu preko tra
 i kukuruza. Među svima pobrojanim hlebovima najukusniji je hleb od pšenična br
lima brane svoju fabriku, svoj hljeb. Nije patetično... Na Terazijama m
e, vare mlijeko, spravljaju pite i hljebove. Slivljanima je najteže. Pod oružje
 ko kad dete umre. Po pet-šes kila leba jedu naše mečke na dan, jedu i kukuruz.
eskonačno da se deli, k'o onih pet lebova u Jevanđelju: komandovao bi bitkom dan
Oblajavaš ljude po novinama, pa se ljebom raniš! "Jesi li u tom "oblajavanju”
```

Table 2 The concordance lines produced by the given regular expression. In bold are Iekavian occurrences.

Though the existence of this corpus has only recently been widely known, it has already been used by many researches, most of them foreign Slavists, for the variety of applications mostly due to its free of charge use for the research purposes and its user friendly web interface.

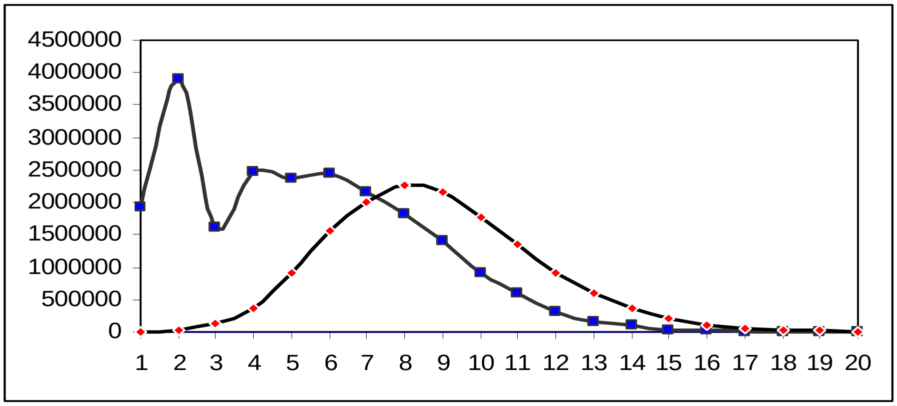

Figure 2 The distribution of frequencies of lengths of types and tokens measured in number of characters in SrpKor (data for types was scaled by factor 300)

The size of the SerKor is 23,532,367 tokens and 495,043 types. Various statistical analyses were performed on corpus, one of them related to the word length. The obtained results correspond to the previous results obtained on smaller data (Vitas 2005). The peculiar form of the curve for the distribution of the frequencies of the lengths of tokens is found in other languages (Przepiórkowski 2005).

The analysis shows that the four most frequent types, conjunctions *i* and *da*, preposition *u*, and a form of the copula verb *je* cover more than 10% of all tokens[1]. With additional two more words, preposition *na* and reflexive particle *se*, 15% of corpus size is covered. The 80% of corpus size is covered by 19,462 types, that is, by less then 4% of all types. It can be seen from Figure 4 that 1000 most frequent types cover more then half of all corpus.

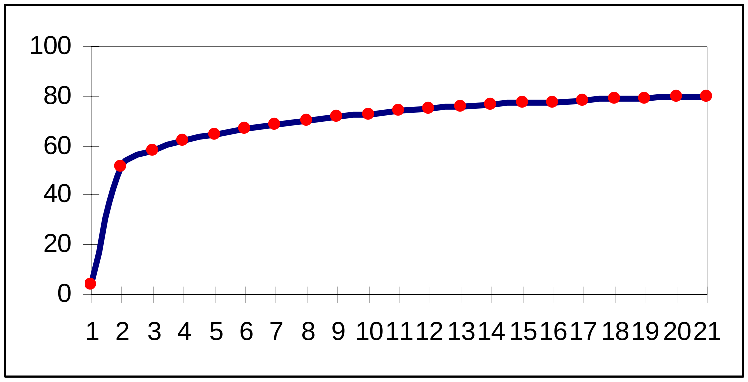

Figure 3 Coverage of corpus by the most frequent types.

3 Serbian Morphological E-dictionary

We have developed the morphological electronic dictionary for Serbian based on the model adopted for the construction of this kind of dictionary in the scope of the network RELEX (Laport 2003). Our system of morphological dictionaries consists from dictionaries of simple words (a sequence of alphabet characters), dictionaries of compounds, and the set of lexical transducers that approximates the unknown words, that is, the words that are not in the dictionaries of the system. For the analysis that we are presenting in this paper we have used the dictionaries of simple words and derivational lexical transducers. A simple words module consists of three parts: (a) dictionary of lemmas DELAS; (b) set of transducers describing the properties of inflectional paradigms; (c) dictionary of inflected forms DELAF. For instance, an entry that corresponds to the lemma *kralj* (Engl. king) in the Serbian dictionary of simple words DELAS is:

```
kralj,N84+Hum
```

[1] It should be noted that the word forms *da* and *je* are ambiguous: *da* can also be the inflected form of verb *dati* (Engl. to give), and *je* can be clitic form of the pronoun *ona* (Engl. she). However, these realizations are much less frequent.

The marker +Hum assigned to this entry describes it as *human*. In Serbian DELAS dictionary there are 22 entries whose inflection is described by the transducer N84. From this particular entry, and using the transducer N84, all the inflectional forms are computed that belong to the dictionary DELAF:

```
kralj,kralj.N84+Hum:ms1v
kralja,kralj.N84+Hum:ms2v:ms4v
kraljem,kralj.N84+Hum:ms6v
kraljeva,kralj.N84+Hum:mp2v
kraljeve,kralj.N84+Hum:mp4v
kraljevi,kralj.N84+Hum:mp1v:mp5v
kraljevima,kralj.N84+Hum:mp3v:mp6v:mp7v
kralju,kralj.N84+Hum:ms3v:ms5v:ms7v
```

Each entry in DELAF dictionary is a word form to which its lemma, that is, entry in DELAS dictionary, is associated together with the set of possible grammatical categories, each category represented by the single character code. For instance *kralja* has two such sets associated to it, denoting that it can be the genitive (**2**) or accusative case (**4**)**,** singular (**s**) of the masculine gender (**m**) of the animate (**v**) lemma *kralj*.

	DELAS	DELAF	DELAF/DELAS
General lexica	77,136	1,037,263	13.45
Geographic names	3,293	34,531	10.49
Personal names	22,130	138,414	6.25
TOTAL	**102,559**	**1,210,208**	**11.80**

Table 3 . The present size of Serbian morphological e-dictionary

The system of DELAS / DELAF dictionaries consists of three main parts. The largest is the dictionary of general lexica that corresponds in size to the Serbian one-volume dictionary. The dictionary of geographic names is still under development. The dictionary of personal names consists itself of several parts: the largest part consists of Serbian personal names, while the development of dictionaries of English personal names transcribed to Serbian orthography and the dictionary of celebrities are still in initial phase. The largest part of these dictionaries was constructed using as a source various traditional dictionaries, grammars, gazetteers, and word lists, but they have also been enriched by lexica found in many processed text. However, texts used for dictionary development, with the exception of Orwell's *1984*, are not included in SerKor, the part of corpus used in the analysis. The dictionary is encoded using the same encoding schema used for the corpus. The main tool for the exploitation of e-dictionaries is the system Intex v. 4.33 (Silberztein, 2004).

Due to the recent political changes in the region, Serbian language is going through the phase of redefining its position in the scope of what was known as Serbo-Croatian (Popović 2003). Since we no not what to predict future solutions and thus restrict our dictionaries to a particular pronunciation or variant, we have encompassed in our dictionaries both major pronunciations, Ekavian and Iekavian, as well as several other variants. All lemmas specific to a certain pronunciation or variant are marked by an appropriate marker, e.g. +Ek and +Ijk for Ekavian and Ijekavian pronunciations, respectively, and +Sr and +Cr for Serbian and Croatian form. Also, a number of lemmas can be produced using different suffixes, some of which are more specific to Serbian and other to Croatian. We have included all those lemmas as well, and marked them with special markers, such as +DerRaSa and +DerSaRa (Table 3).

lemma		**Marker**	**English**
deca	Ekavian	+Ek	children
djeca	Iekavian	+Ijk	
delirijum	Serbian	+Sr	delirium
delirij	Croatian	+Cr	
deformisati		+DerSaRa	to deform
deformirati		+DerRaSa	
istorija		+Der0H	history
historija		+DerH0	

Table 4 The illustration of pronunciation and derivational variant forms

The Serbian language is characterized by the rich morphological system, which is reflected not only on the inflective but also on the derivational level. Particularly productive are derivational processes that produce new lemmas with predictable meaning. We call this *regular derivation*. Since derived lemmas are regularly produced, we have chosen, as a rule, not to include them in the dictionaries, but rather to recognize them with lexical transducers incorporated in Intex v.4.33. As an illustration, the possessive adjective *kraljev* (Engl. belonging to the king) and all its inflected forms, *kraljevog*, *kraljevom*, *kraljevim*, etc., although not in DELAF dictionary, are recognized on the basis of several facts checked by the appropriate lexical transducer: *kralj* is in a dictionary, *-ev* is a possessive adjective suffix, and *–og*, -o*m*, *-im*, etc. are its inflectional endings. Numerous lexical transducers have been produced that make use of the similar derivational patterns to recognize possessive adjectives, diminutives, augmentatives, gender motion, prefixation, etc. (Vitas 2005a).

4 Coverage of the Corpus by E-dictionary

In order to estimate the incompleteness of the dictionaries, we have performed the experiment on one large subset of all corpus types. We have chosen the subset of all the word forms beginning with letters “D”, “Đ”, and “Dž”, the letter “D” being very frequent in the initial position, the other two much less. There are in corpus 24,909 (5.03%) types beginning with these letters, and 1,598,521 (6.79%) tokens. We have applied the lexical resources described in Section 3 in several steps:

a) Application of the dictionary of general lexica;
b) Application of the dictionary of personal names;
c) Application of the dictionary of geographic names;
d) Application of the lexical transducers for regular derivation;
e) Application of the English morphological dictionary.

We found out that after applying the dictionary of general lexica only less then 50% of all types are recognized (46.32%). At the same time, however, almost 93% of all tokens are already recognized (92.99%). When all lexical resources are applied more then 62% of all types are recognized and almost 97% of all tokens (Figure 4)..

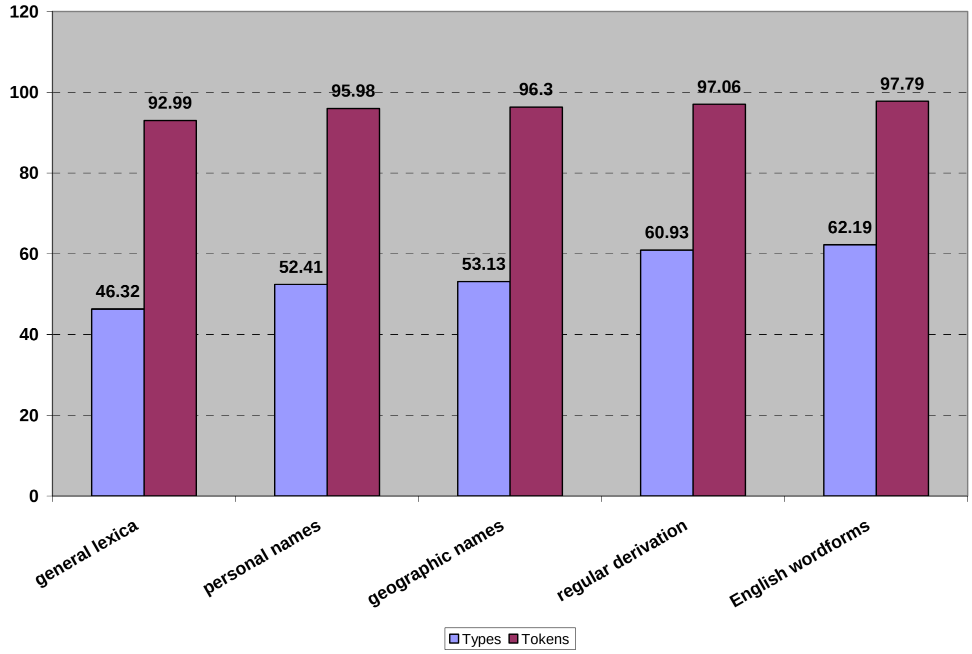

Figure 4 The increase of the coverage of the corpus after applying different lexical resources

After applying the dictionary for general lexica only the most frequent unrecognized word form is *DOS* (7239), the acronym of a political association, it is followed by the surname *Đinđić* (2540), and the first name *Dušan* (2298). Among the ten most frequent unrecognized words are only acronyms and personal names. After applying the dictionary of personal names, among the ten most frequent unrecognized words are 7 acronyms, mostly for political parties and currencies, and two inflected forms of *Dunav* (Engl. Danube). The last word form among this top-ten is *Džon*, which is the Serbian transcription of the English name *John*. Our dictionary of English personal names transcribed according to the Serbian orthography is still under development, and we have systematically processed only the English names with initial A, B, and C. None of those names, however, has in Serbian orthography the initial D, Đ, or Dž. It should be, therefore, expected that our results would improve once we finish this dictionary.

After adding the dictionary of geographic names, among the ten most frequent word forms are only acronyms and the English name *Džon*. Among the next ten most frequent word forms are, besides the acronyms, the English name *Džejms* (Engl. *James*), and *departmenta*, the inflected form of *department*, which is one part of a compound *Stejt department* (a transcription of the State Department). It is interesting to note that on the top of the list is one Serbian surname *Drulić* that is not in the list of all the citizens of Belgrade that was used as one of the sources for the development of our dictionary of personal names. Apparently, it is the surname of a football player popular at one moment.

The word forms recognized by regular derivation lexical transducers are not among those most frequent, which is illustrated in Figure 5: the number of types recognized increased by 7.8%, while at the same time the number of the recognized tokens increased by only

0.76%. The most frequent word form recognized by some lexical transducer is *dvodnevnog* (110), the inflected form of the adjective *dvodnevni* (Engl. lasting two days), which is derived from the numeral *dva* (Engl. two) and the adjective *dnevni* (Engl. daily) (Table 4).

Part of speech	**Type of derivation**	**N. of types**	**example**	
Numerals	collective	54	*dvadesetdvoje*	twenty two people
Nouns	diminutives	35	*dabarcyicx*	small beaver
	augmentatives	5	*dimcyina*	big smoke
	gender motion	52	*desnicyarka*	right-wing female person
	action	5	*demonizacija*	the action of demonizing something or somebody
	attribute	81	*decentralizovanost*	the attribute of not being centralized
	compounds & prefixes	391	*deprofesionalizacija*	reversal of the action of profesionalization
			dvojezicyar	bilingual person
Adjectives	possessive	686	*detektivov*	belonging to a detective
	compounds & prefixes	724	*dvadesetoslovni*	having twenty letters

Table 5 The frequency of the various types of derivations and some examples of recognized forms

The English morphological dictionary is incorporated in the Intex programming system that we use for dictionary development and maintenance (Curtois 2004). We have made use of these dictionaries to try to recognize among the remaining unrecognized words the English words, and by this we mean the words written using the English spelling. As in a previous step, there are no such words among the most frequent unrecognized words. However, there were quite a number of them: 466 different word forms that are possible English word forms. The most frequent English word in corpus is *daily* (47), being the part of many newspaper and news agency names: *Daily Telegraph*, *Daily News*, *Daily Mirror*, etc.

It should be stated that the last two steps are only approximations, that is, they do occasionally erroneously recognize a word forms as a regularly derived form or as an English word. For instance, *datotetku* is recognized as a compound produced form the adjective *dat* (Engl. given) and the noun *tetka* (Engl. aunt), although it is actually the misspelling of *datoteka* (Engl. data file). Also *Damask* is recognized as the English noun *damask*, type of a fabric, when it is actually the name of the capital of Syria that is still missing from the dictionary of geographic names. These cases are not very frequent and cannot blur the overall results.

After applying all lexical resources 9417 word forms still remain unrecognized. Although we cannot give the precise figure, the first results show that among them there are a lot of typographic errors. For instance, among unrecognized words there are 186 word forms with doubled letter, by far the most of them being errors, foreign words, or acronyms since doubled letters rarely occur in Serbian texts: for instance *danaas* instead of *danas*

(Engl. today), *Decca* (English company Decca Music Group), and *debugging* (obviously missing from the English morphological dictionary supplied by Intex). There are also 108 word forms without vocals, most of them being very short, up to four characters, and it seems that the most of them are acronyms. Finally, there are 70 word forms that use letters X, Y, Q, and W that are not part of the Serbian Latin alphabet, almost all of them being either acronyms or foreign words like *Dreamweaver*, *diplomatique* (from Le Monde diplomatique), and *DaimlerChrysler*.

Apart from typos and acronyms, in the remaining 9053 word forms there are number of word forms actually missing from some of the dictionaries, most of them either geographic names or English personal names, since these two dictionaries are still under development. The dictionary of personal names known in history or from literature still has to be developed as it is now in its initial stage, thus failing to recognize, for instance, *Daladje* (Daladier) and *Dalamber* (D'Alambert). The analysis of this list also shows that small dictionary of prefixes used to recognize prefixed nouns and adjectives has to be enhanced, for instance by *dez-*, *dis-*, and *dermo-*. Also, some compounds with numerals remain unrecognized, and this problem we will solve by improving the appropriate lexical transducers. We should note that there are a number of cases that will probably never be covered by an e-dictionary. Consider the corpus example: "*A ne da se, po Interkontinentalima, Dedinjima, Kiprima, Grčkama, Kinama, Havanama, vilama, hotelima, bahanalijama kocka sa celim narodom stavljajući ga kao žeton…*" in which several toponyms (Dedinje – residential part of Belgrade, Cyprus, Greece, China, Havana) are metaphorically used in plural form.

5 Coverage of E-dictionary in the Corpus

	n. of entries	*n. of realized*	*ratio* (%)
general lexica	46542	12804	27.51
personal names	8961	1931	21.55
geographic names	1320	287	21.74
Total	56823	15022	26.44

Table 6 Realization of dictionary word forms in SrpKor

In order to measure the representativness of dictionary forms in corpus we have again chosen a subset of dictionary entries beginning with letters "D", "Đ", and "Dž". We have established that certain dictionary entry is represented in corpus by simply matching DELAF entries with a list of types from the corpus. Since corpus is not tagged, and is therefore ambiguous, the results obtained are only approximations. Namely, some ambiguous word forms are all realized in corpus, such as *daće* − third person singular and plural form of the future tense of the verb *dati* (Engl. to give), and several inflected forms of the noun *daća* (Engl. feast to honour the deceased). In other cases only one of two ambiguous forms is realized, as in the case of *damo* – only the first person plural form of the verb *dati* in present tense is realized, while the vocative singular form of the noun *dama* (Engl. lady) is not. For that reason, data given in third and forth column in Table 5 are not accurate but maximum values. However, if taking into account the level of the ambiguity in Serbian texts and dictionaries the obtained results can be interpolated.

5.1 The analysis of the realization of word forms

Data in Table 5 suggests that the ratio of realization of dictionary word forms is low, approximately one quarter. Research can be done as to the realization of some particular forms, which we will illustrate by two examples. It is a common believe that the imperfect tense is rarely used in contemporary Serbian. We have used the third person plural form of imperfect tense in order to investigate the usage of this tense in contemporary Serbian since all other imperfect forms are often ambiguous with some aorist tense forms, and found that only 13 forms have realization in corpus. It seems as a small number taking into account that there are 443 imperfective verbs (with initial D, Đ, Dž) having 486 imperfect forms in third person singular (for some verbs there is more then one possible imperfect form). However, from these 443 imperfective verbs at most 130 have realization of some of their forms in corpus (see section 5.2), and in that light the number 13 does not look insignificant, even more so as it is the minimal number of verbs with a realisation of some imperfect form. On the other hand all these retrieved imperfect forms have a very low frequency, the total is 16.

Our inflectional transducers for imperfective verbs generate the form for the past gerund active. These forms are rarely ambiguous with some other forms. Corpus evidence shows that for imperfective verbs there is only one realization of this form: *dolazivsxi* (Engl. being coming) in "*Međutim, ona je vrlo brzo, uopšte ne dolazivši svesti, preminula…*" (However, she very soon, not coming to consciousness at all, died…). However, in this example the past gerund active is used where present gerund active *dolazecxi* should have been used. This shows that our decision to include this form in the inflectional transducers for imperfective verbs should be reconsidered.

5.2 The analysis of the realization of lemmas

	n. of lemmas	*realized*	
		max	*min*
Nouns	3368	2081 (61.8%)	1300
Verbs	620	451 (72.7%)	281
adjectives	1052	655 (62.3%)	409
Other	227	175 (77.1%)	109
Total	**5267**	**3362 (63.8%)**	**2101**

Table 7 The estimation of the numebr of realised lemmas

Matching the DELAF dictionary with the list of types from corpus enables the estimation of the number of realized lemmas in corpus (Table 7). This estimation gives the maximum number of realized lemmas since the corpus is not disambiguated. There are cases where several word forms are ambiguous and all of them are associated to two different lemmas of which only one is realized. For instance, the word forms *duše* and *duši* are both associated to the noun *duša* (Engl. soul) and the verb *dušiti* (Engl. *to disable somebody to breathe*). The analysis of corpus concordances shows that only the first lemma is realized. The research that we have recently performed on one sample text shows that the ambiguity of tokens in respect to the associated lemmas is 1.6 (Krstev 2005). We have used this coefficient to establish the minimal number of realized lemmas. Taking into account the participation of letters D, Đ, and Dž in the dictionary and the corpus we can estimate that the number of dictionary lemmas realized in corpus is between 41800 and 66800.

Iekavian	207 (7.84%)	2432 (92.16%)
Croatian	43 (14.83%)	247 (85.17%)
Derivational variant	83 (3.24%)	2475 (96.76%)

Table 8 Usage of variant forms in corpus

We performed various useful analyses on obtained data about the potential usage of lemmas in the corpus. First of all, we have established that although corpus consists of texts primarily published in Belgrade all the pronunciation and derivational variants that are not specific to Belgrade language are nevertheless present to some degree (Table 8). This confirms the soundness of our decision to incorporate all these variants in our dictionary. All the botanical and zoological species are marked in DELAS/DELAF dictionary by +Bot and +Zool markers. By using these markers we have established that 20 lemmas for botanical species out of 44, or 45.5%, and 21 lemmas for zoological species out of 35, or 60%, are realized in corpus. Since the realization rate does not differ significantly from the rate for other common nouns (Table 7), we can conclude that corpus is representative for this domain.

The interesting results are obtained by analyzing the inflective lemmas that have only one realized form that, however, occurs rather frequently. These occurrences can point two interesting situations. First, often the lemma is not realized at all, since the one occurring form is actually ambiguous with a form of some other more frequent lemma. The example is lemma *dometati* (Engl. to add something to the conversation) having only one realized form *dometa* (aorist in second and third person singular) with frequency 131. The concordance analysis shows that this form is always either the singular or plural genitive case form of the noun *domet* (Engl. range). These forms are good candidates for a kind of a filter dictionary that can be used during the disambiguation process. Second, such cases can pinpoint that some lemmas do not actually inflect. For instance, the collective numerals, such as *devetoro*, *desetoro* (Engl. nine, ten male persons) have also plural forms *devetora*, *desetora*, but they never occur in corpus for the numerals beginning with D.

6 Conclusions

We find the obtained results very useful since they give many hints as to what paths should be followed in order to improve the developed resources, particularly the morphological e-dictionaries. To that end we plan to perform the described procedure on the complete data.

At this moment we see several useful usages of the obtained data. First of all, we plan to use it for the development of the useful rule or dictionary based disambiguation procedure. Next, we plan to stratify our dictionaries in the sense (Garrigues 1999) according to the frequency of usage, both on the level of lemmas and forms. We also think that the investigation on corpus data in a broader context, for instance bigrams and trigrams, can give results useful for the development of dictionaries of compounds.

References

Erjavec, T., Gorjanc, V., Stabej, M. (1998). Korpus FIDA. V zborniku konference Jezikovne Tehnologije za slovenski jezik. Ljubljana, oktober 1998 http://nl.ijs.si/isjt98/zbornik/sdjt98-Gorjanc.pdf

Garrigues, M. (1993). Méthode de paramétrage des dictionnaires et grammaires électroniques: Application à des systèmes interactifs en langue naturelle, Thèse de Doctorat, Paris, Université Paris 7.

Krek, S., Stabej, M., Gorjanc, V., Erjavec, T., Romih, M., Holozan, P. FIDA (2002) Korpus slovenskega jezika. URL: http://www.fida.net

Krstev, C. and Vitas, D. On the Ambiguity of Serbian Texts and Methods to Disambiguate it. In the *Proceedings of the 8th Intex/Nooj Workshop*, Besançon 29 May – 1 June 2005, [to appear]

Laporte, E. (2003). *The RELEX Network*. (http://infolingu.univ-mlv.fr/Relex/Relex.htm)

Maurel D. (2004). Les mots inconnus sont-ils des noms propres?, *Septièmes Journées internationales d'Analyse statistique des Données Textuelles* (JADT 2004), Louvain-la-Neuve, Belgique. pp. 776-784 http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_074.pdf

Christ, O. (1994) A modular and flexible architecture for an integrated corpus query system*. COMPLEX'94,* Budapest, http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

Courtois, B. (2004) Dictionnaires lectroniques DELAF anglais et français. In *Lexique, Syntaxe et Lexique-Grammaire / Syntax, Lexis & Lexicon-Grammar*, Leclère, C., Laporte, E., Piot M. and Silberztein, M. (eds.), Papers in honour of Maurice Gross, Lingvisticae Investigationes Supplementa 24, John Benjamins

Popović, Lj. (2003) Od srpskohrvatskog do srpskog i hrvatskog standardnog jezika: srpska i hrvatska verzija. Wien, *Wiener Slawistischer Almanach*, 57, 201-224.

Przepiórkowski, A. (2005) The IPI PAN Corpus in Numbers, in Proceedings of 2nd Language & Technology Conference, April 21-23, 2005, Poznań, Poland, ed. Zygmunt Vetulani, pp. 27-31, Wydawnictwo Poznańskie Sp. z o.o., Poznań

Santos, Diana (1998). Providing access to language resources through the World Wide Web: the Oslo Corpus of Bosnian Texts. *Proceedings of The First International Conference on Language Resources and Evaluation* (Granada, 28-30 May 1998) http://www.tekstlab.uio.no/Bosnian/lrec.rtf

Silberztein, M. (2004) INTEX Manual, v. 4.33. Available on-line from http://intex.univ-fcomte.fr/downloads/Manual.pdf

Tadić, M. (2002). Building the Croatian National Corpus, LREC2002, Las Palmas, 27 May -2 June 2002, ELRA, Paris-Las Palmas 2002, Vol. II, 441-446
http://www.hnk.ffzg.hr/txts/mt4LREC2002.pdf

Vitas, D. et al (2003) An Overview of Resources and Basic Tools for the Processing of Serbian Written Texts", in the *Proceedings of the Workshop on Balkan Language Resources and Tools*, 21 November, Thessalonica, Greece, [to appear]

Vitas, D. and Krstev, C. (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

Vitas, D., Pavlović-Lažetić, G. and Krstev, C. (2005) About Word Length Counting in Serbian, in *Contribution to the Science of Languages – Word Length Studies and Related Issues*, ed. Peter Grzzbek, Kluwer Academic Publisher, [to appear]