Extending Serbian E-dictionary by the Use of the Lexical Transducers

Cvetana Krstev (1), Duško Vitas (2)

Faculty of Philology, University of Belgrade
Studentski trg 3, CS – 11000 Belgrade
e-mail: cvetana@matf.bg.ac.yu
(2) Faculty of Mathematics, University of Belgrade
Studentski trg 16, CS – 11000 Belgrade
e-mail: vitas@matf.bg.ac.yu

Abstract: In this paper we present one approach in reducing the number of unrecognized words that is based on the use of the transducers with lexical constraints incorporated in Intex 4.3x. We describe the derivational processes in Serbian that were modeled using this kind of transducers, and how the problems of phonological alternations and inflection of derived forms were solved. We present the analysis of the obtained results on three test subcorpora and we outline the possible solutions for the spotted problems. Finally, we present some ideas for future improvements.

1 Introduction

The basic resource used for processing Serbian texts is a system of morphological electronic dictionaries of Serbian, in DELA format (B. Curtois, M. Silberztein, 1990), which corresponds in size to the one volume dictionary of Serbian. It consists of the dictionaries of simple words—DELAS (approximately 77,000 lemmas at present), simple word forms—DELAF (more then 1,000,000 word forms), and a
dictionary of compounds (multiword expressions)—DELAC (its derivation is still in the initial phase). Besides these basic dictionaries of general lexica, the dictionaries of proper names are being developed in the same format, and they now consist of approximately 16,000 entries to which more than 127,000 word forms correspond. Presently, dictionaries of toponyms and personal names are being developed (G. Pavlović-Lažetić et al, 2004; Vitas et al, 2005).

All the entries in the DELAS dictionary have the form \texttt{lemma.Knn [+SinSem]}* where \texttt{lemma} is the simple word, in the revised form compared to the one used in the traditional dictionaries (Vitas et al., 2001), \texttt{K} is the part of speech mark, \texttt{nnn} is the number denoting the class of lemmas that all share the same inflectional properties described by the appropriate transducer \texttt{Knn}, and \texttt{+SinSem} is the freely attached marker that describes the syntactic, semantic, derivational, or other properties of a lemma. For instance, the marker \texttt{+p2} applied to a preposition means that the noun phrase following it has to be in the genitive case, the marker \texttt{+Hum} applied to a noun determines it as a human being, the marker \texttt{+PP} applied to an adjective determines that it has been derived from the passive past participle, the marker \texttt{+Ek} applied to any lemma means that it is specific to the Ekavian pronunciation, etc.. These markers do not form a closed set, so they are being modified and expended in the course of the dictionary development. More specifically, the semantic markers are being added in a more consistent manner using the information from the Serbian Wordnet (Krstev et al, 2004).

All the entries in the DELAF dictionary have the form \texttt{form,lemma:[categories]}* where \texttt{form} is a simple word form of a \texttt{lemma} that is represented by its DELAS entry form, and \texttt{:categories} are the possible grammatical categories of the word \texttt{form}, each category represented by the single character code. The set of grammatical categories is specific to Serbian, though it has been defined along the same line as for French, as described in (Vitas, 2004).

It has been shown that the e-dictionary in DELA format can be transformed in the standard MULTTEXT format, thought not in a quite straightforward way (C. Krstev et al., 2004a).
2 Motivation

Although the size of the e-dictionaries provides for high percentage of successful text tagging and lemmatization, a significant number of words still remain unrecognized. In order to determine the source and scope of this problem, three subcorpora of contemporary Serbian have been compiled on which some experiments have been performed: first consists of daily newspaper texts, second of literary texts, and third of popular economy journal texts. The results of these experiments are represented in Table 1. Although dictionaries of proper names are still far from being complete it can be seen that the recognition rate is improved when they are included in the recognition process, especially for the daily newspaper and economy journal subcorpora.

<table>
<thead>
<tr>
<th></th>
<th>total number of simple Forms</th>
<th>different Forms</th>
<th>recognized simple words before</th>
<th>recognized simple words after</th>
<th>simple words not recognized</th>
</tr>
</thead>
<tbody>
<tr>
<td>newspaper</td>
<td>582K W</td>
<td>689 (9.8:1)</td>
<td>55581 (80.64 %)</td>
<td>59103 (85.75 %)</td>
<td>9825 (14.25 %)</td>
</tr>
<tr>
<td>literary</td>
<td>245K W</td>
<td>417 (5.9:1)</td>
<td>36943 (88.43 %)</td>
<td>37246 (89.16 %)</td>
<td>4528 (10.84 %)</td>
</tr>
<tr>
<td>economy</td>
<td>404K W</td>
<td>407 (9.9:1)</td>
<td>36257 (88.97 %)</td>
<td>37124 (91 %)</td>
<td>3641 (8.93 %)</td>
</tr>
</tbody>
</table>

Table 1: Recognition rate before and after application of the dictionaries of proper names.
In the next step of the experiment the sets of unrecognized words were analyzed in order to establish why they could not have been recognized. This analysis showed:

1. All texts contain quite a number of typographical errors, such as administrator instead of administratör ‘administratieur’, agenciyu instead of agenciju, accusative form of ‘agency’.

2. All dictionaries suffer from insufficiency, both general dictionaries (no entries for algoritam, ‘algorithm’, and aminovao and aminuju that are forms of the verb aminovati, ‘give consent for something’), and dictionaries of proper names (no entries for Aleksandrovac, town in Serbia, Alžirac, the inhabitant of Alger, Abrašević, Serbian surname).

3. Newspaper and journal texts contain a number of foreign words, written according to the orthography of the language of origin (for instance, agency, Alexander, Apple form English, tour de force from French, staccato from Italian, and tempora mutantur from Latin).

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

5. 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 ambasador are entries in traditional dictionaries and in Serbian e-dictionary.

The problems listed in points 1 and 2 can simply be solved by correcting the text and by improving existing dictionaries, problems listed in points 3 and 4 can be solved by producing new dictionaries, such as dictionaries of common Latin words, and foreign words in general, or otherwise can remain unsolved, while problems listed in point 5 ask for a different solution that will be presented in this paper.
3 Derivational processes

The analysis of the unrecognized words belonging to the fifth listed class shows that many of them follow some known, and usually productive, derivational pattern. Such a pattern can be well known and applicable in many languages, as, for instance, prefixation, while the others are specific to Serbian language (and Slavic languages in general) such as derivation of possessive adjectives from nouns. The aim of this analysis was to determine what are the most frequent and/or the most regular derivational processes in Serbian.

The characteristic of the prefixation of the considered type is that it does not change the part-of-speech of the lemma to which the prefix is applied; moreover, the derived lemma shares the inflectional class with the basic lemma. Also, it can be applied to various PoS, nouns (e.g. antikorupcija ‘anti-corruption’), adjectives (e.g. antivirusni ‘that which is anti-virus’), verbs (e.g. odskakati ‘to bounce’), and adverbs (samozađovoljno ‘self-contently’). Particularly productive is prefix ne- that produces negation from nouns (neregularnost ‘irregularity’), adjectives (nepotpun ‘incomplete’) and adverbs (e.g. nezvanično ‘unofficially’). Negation from verbs in Serbian is not produced with the prefix ne- but with the particle ne that is written separately from the verb. The case of prefix naj- is specific, as it is used for the production of the superlatives forms of adjectives and adverbs. Contrary to the other prefixes, lemma of the superlative form of an adjective is the adjective itself, for instance, lemma of najbogatiji ‘the richest’ is bogat ‘rich’, while in the case of all the other prefixes it is the derived adjective, for example lemma for nepotpun is nepotpun, not potpun. However this is not the case for adverbs, since they are regarded as uninflected words by traditional Serbian grammars, and superlative forms are separate lemmas (e.g. najradije ‘most willingly’, is lemma itself, although it can be also regarded as the superlative form of the adverb rado ‘willingly’).

Derivation of possessive adjectives is also highly productive. Possessive adjective is usually produced from the animate nouns, for instance, premijerov ‘belonging to the prime minister’ and sovin ‘belonging to the owl’. Particularly frequent are possessive adjectives
derived from the first names (e.g. Milanov ‘belonging to Milan’) and surnames (e.g. Đinđićev ‘belonging to Đinđić’). Moreover, feminine gender forms of the possessive adjectives derived from surnames are frequently used to denote a feminine person using that surname (e.g. Arsićeva, woman whose surname is Arsić). Possessive adjectives are less frequently produced from inanimate nouns, however the cases exist, for instance bancin (‘belonging to the bank’, meaning to the particular bank), and Zastavin (derived from zastava ‘flag’, meaning belonging to “Zastava”, the Serbian motor car factory).

New nouns are produced from nouns by diminution and augmentation. However, the derived noun does not share the same inflectional class with the basic noun. In Serbian, diminution is a productive process (e.g. medvedić, diminutive from medved ‘bear’, and milijardica, diminutive from milijarda ‘billion’), while augmentation is much less productive. In the economy sub-corpus no augmentative nouns were found, while only two occurred in the newspaper sub-corpus, namely glasina (augmentative of glas ‘voice’), and zidina (augmentative of zid ‘wall’), that have both lexicalized in the specific meaning as ‘rumor’ and ‘fortress walls’, respectively. In the literary sub-corpus four augmentatives were found, three of which have lexicalized in a specific meaning: already mentioned glasina and zidina, and gradina (augmentative of grad ‘city’) with two specific meanings ‘garden’ and ‘fence’, and only one true augmentative komadina (augmentative of komad ‘piece’). In Serbian, diminutives can be produced both from diminutives and augmentatives, for instance, diminutives and augmentatives of kesa ‘bag’ are respectively kesića and kesićina, and their diminutives are kesićica and kesićina. However, such cases do not occur in any of the analyzed sub-corpora (D. Vitas 2004a).

Special case of derivation is the formation of compounds from simple words. This kind of process is highly productive in Serbian, but only some compounds are recorded in traditional dictionaries. The most frequent are compounds obtained by connecting numbers with
nouns (e.g. *pedestro*[-]*godišnjak* ‘52 years old man’) and adjectives (e.g. *jedno*[-]*i*[-]*po*[-]*satni* ‘lasting one hour and a half’), and by connecting adjectives with other adjectives (*sivkasto*[-]*žut* ‘of color between gray and yellow’), nouns (*plitko*[-]*umnost* ‘narrow-mindedness’), and adverbs (*pravo*[-]*linijski* ‘straight-forwardly’). In the later case, the adjective as a first part of a compound is usually in a neuter gender, singular number, and nominative case form. For this kind of compounds, both part-of-speech and inflectional class are same as for the final constitutional element of the compound. Occasionally, the second part of a compound is an adjective that is used for production of many compounds, for instance with numbers, but is never used as a simple adjective. Such is the case of the adjective *spratni* that is used in many compounds (e.g. *peto*[-]*spratni* ‘having five floors’, *mnogo*[-]*spratni* ‘having many floors’) but never as a simple adjective.

In some languages, as in French, the gender motion is considered the part of the inflective paradigm. This is not the case in Serbian, where it is considered as separate lemma that is sometimes recorded in a traditional dictionary but in a rather unsystematic way (C. Krstev and D. Vitas, 2004b). Although both the basic and the derived lemma belong to the same PoS, they always belong to different inflectional class. For instance, *novinar* ‘journalist’ belongs to the class of masculine gender animate nouns with unmarked declension (denoted in DELAS as N2), while *novinarka* ‘woman journalist’ belongs to the class of feminine gender animate nouns marked by the use of the inflectional ending –*i* in genitive plural (denoted in DELAS as N661).

The abstract nouns denoting the attribute or the quality of an entity are frequently derived from the adjectives, and they often belong to the same inflectional class of feminine gender inanimate nouns ending in consonant and marked by the use of two possible endings –*i* and –*u* in instrumental singular (denoted in DELAS as N704). Many of those derived nouns are recorded in the traditional dictionaries and are

---

1 In this and the following examples the hyphen is used to denote the point were simple words meet to form the compound; however, it is not used in the writing.
characterized by the prototype definition of the form: “quality of that who is X”. For instance, the noun bezosećajnost ‘insensitivity’ is characterized in the dictionary by the definition osobina onoga ko je bezosećajan ‘quality of that who is insensitive’. However, the analysis of unrecognized words in three test sub-corpora showed that many nouns derived in this manner are still missing from the dictionary.

Active past participle, passive past participle, and present gerund active are traditionally regarded as a part of one verb paradigm. For instance, for the verb zlostavljati ‘to abuse’, active past participle for masculine gender singular is zlostavljao (used in several active compound tenses), passive past participle in masculine gender singular is zlostavljan (used in several passive compound tenses), and present gerund active is zlostavljajući ‘while abusing’. These three verb forms are sometimes used as adjectives as well, which means that grammatical cases are added to their paradigm. Passive past participle is most frequently used as an adjective (e.g. automatizovan rad ‘automated work’) and thus it is sometimes, especially when it acquires some specific meaning, recorded as such in traditional dictionaries. Due to its frequent use as an adjective, it was, however, systematically added to the e-dictionary. On the other hand, active past participle (e.g. požutele školske čitanke ‘school textbooks that become yellow’) and present gerund active (e.g. zlostavljajuće proganjanje ‘the abused prosecution’) are also sometimes used as adjectives, but that is practically never recorded in traditional dictionaries, nor were additional entries made for them in e-dictionary.

4 Phonological alternations

When during the derivation process certain phonemes occur one beside other, they influence each other that can lead to the phonological alternations, that is, some phonemes when found in certain context are replaced by some other phonemes. In Serbian, these phonological alternations are often recorded in written text. The
phonological alternations occurring in derivational processes described in section 3 are:

- **Palatalization** is the process of consonant mutation in which velar consonants k, g, h when found in certain context are replaced by the palatal consonants č, ž, š, respectively. The same process changes dental consonants c and z to palatal consonants č and ž. This alternation occurs, for instance, when producing diminutives of the feminine gender nouns, e.g. *ruka* ‘hand’ → *ručica* → *ružica*, diminutives of the masculine gender nouns, e.g. *vrtilog* ‘whirl’ → *vrtožić* → *vrtožić*, augmentatives of the masculine gender nouns, e.g. *dak* ‘pupil’ → *dak-ina* → *dačina*, and possessive adjectives, e.g. *kraljica* ‘queen’ → *kraljičin*.

- **Sibilarization** is the process of consonant mutation in which velar consonants k, g, h when found in certain context are replaced by the hissing dental consonants (sibilants) c, z, s, respectively. This alternation occurs in gender motion, e.g. *književnik* ‘writer’ → *književniča*.

- **Assimilation of voiced and voiceless consonants** occurs when two consonants that differ in voicing make the contact. In that case the first one is adapting to the second one in voicing. For instance, if the first consonant is voiced (b, g, d, ž, z, dž) and the second consonant is voiceless (p, k, t, č, š, s, č, f, h, c), than the first one is replaced by its voiceless counterpart (p, k, t, č, š, s, č, respectively). This alternation occurs in deriving the possessive adjectives in conjunction with the fleeting –a, e.g. *predak* ‘ancestor’ → *predak-ov* → *predk-ov* → *pretkov*.

- **Fleeting –a’** is a vocal ‘a’ that occurs in some word forms and disappears in others. In a word it is the last vocal, followed by one consonant or the consonant cluster. This alternation is not specific to the derivational process itself; however, if it occurs in inflection it may occur in derivation as well. Such is the case of the noun *predak* given in the previous item: two alternations, fleeting ‘a’ and alternation of voiced d to the voiceless t, govern the production of the genitive singular *pretk* and the possessive adjective *pretkov*. The same process occurs when producing the abstract nouns from the
adjectives, as, for instance, in moralan ‘moral’ → moralan-ost → moralnost, and diminutive forms from the masculine gender nouns, e.g. tigar ‘tiger’ → tigar-ić → tigrić.

- Alternation of the sonant ‘l’ into vocal ‘o’ takes place when the sonant ‘l’ occurs at the end of the syllable in which case it is replaced by the vocal ‘o’. As with the fleeting ‘a’, this alternation occurs in derivation from the basic lemma only if it occurs in the inflection of the same lemma. Such is the case of the noun posetilac ‘visitor’: two alternations, fleeting ‘a’ and alternation of sonant l to the vocal o, govern the production of the genitive singular posetilac → posetilaca → posetića → posetioca, and possessive adjective posetilac → posetilac-ev → posetić-ev → posetić-ev, when at the end palatalization applies as well.

Basic and derived lemma can differ, not only due to the phonological alternations, but also due to the stripping of a suffix: for instance, suffix -ac is stripped from the noun muškarac ‘man’ in order to obtain its diminutive form: muškarac → muškarac-ćić → muškarčić.

5 Transducers with lexical constraints

The lexical transducers incorporated in Intex 4.3 allow the expression of the morphological rules that govern word formation (Silberztein, 2004). The input of a lexical transducers is used to recognize word forms while the output is used to compute the corresponding lemma and other grammatical information. They can perform the tokenization of the word forms into linguistic units, which are established on the basis of imposed constraints expressed in terms of recognition by e-dictionaries. Furthermore, during the recognition process the values of the recognized linguistic units can be stored into variables, which can later be used for the computation of lemmas and grammatical categories. The computed lemmas and associated grammatical categories enter the dictionary of the text and can be used
in text processing in the same way as the information retrieved from the e-dictionaries themselves.

In order to effectively produce the transducers with lexical constraints that would recognize in text the derived forms of lemmas already in the e-dictionaries certain decisions had to be made. As already pointed in section 3, the derived lemma do not necessarily belong to the same PoS as the basic lemma, and even if it does, the basic and the derived lemma often belong to the different inflective classes. Although the inflective class of the derived lemma is predetermined, the inflectional forms and grammatical categories are in Serbian numerous, especially for the adjectives, which makes it cumbersome to deal with them in lexical transducers themselves, as suggested for Portuguese in (Mota, 2003). For that reason the small dictionary of dummy entries was produced, neither of which is actually a lemma in Serbian (see Table 2). Their role is to enable the recognition of the inflective forms of the derived entries.

| wyyoe,A41+Dummy | for the adjectives derived from the active past participle |
| wqost,N704+Dummy | for the abstract nouns derived from the adjectives |
| wyyinxa,N601+Dummy | for nouns of feminine gender derived from the nouns of masculine gender |
| wqyka,N661+Dummy | for nouns of feminine gender derived from the nouns of masculine gender |

Table 2: A few dummy entries from the auxiliary dictionary.
This dictionary is used with the lowest priority.

The usage of the dictionary of dummy entries is illustrated by the transducer from Figure 1 that recognizes the abstract nouns derived from adjectives. For instance, *aljkavost* ‘sloppiness’ is derived from
the adjective *alkav* ‘sloppy’. Its inflected forms and possible grammatical categories are:

\[
\begin{align*}
alvkavost, & \text{alkkavost.N704:fs1q:fs4q} \\
alvkavostxcu, & \text{alkkavost.N704:fs6q} \\
alvkavostima, & \text{alkkavost.N704:fp1q:fp2q:fp4q:fp5q} \\
alvkavosti, & \text{alkkavost.N704:fs2q:fs3q:fs5q:fs6q:fs7q:fp1q:fp2q:fp4q:fp5q} \\
\end{align*}
\]

However, these inflected forms need not be in the e-dictionary as long as the adjective *alkav* is. The dummy entry *wqost* has been assigned the same inflectional class as *alkkavost* N704 (see the second row in Table 2), so its inflected forms are:

\[
\begin{align*}
wqost, & \text{wqost.N704:fs1q:fs4q} \\
wqosxcu, & \text{wqost.N704:fs6q} \\
wqositma, & \text{wqost.N704:fp6q:fp7q} \\
wqostii, & \text{wqost.N704:fs2q:fs3q:fs5q:fs6q:fs7q:fp1q:fp2q:fp4q:fp5q} \\
\end{align*}
\]

2 Serbian e-dictionary is encoded in pure ASCII in order to neutralize the regular usage of two alphabets in Serbian written texts: Latin and Cyrillic. The specific Serbian characters are represented as follows: š → sx, đ → dx, č → cy, ć → cx, ž → zx, nj → nx, lj → lx, dž → dy.

---

Figure 1: The transducer that tokenizes attributive nouns derived from adjectives.
For instance, recognition of \textit{alxkavosxcxu}, the inflected form of \textit{aljkavost}, by the transducer from Figure 1 proceeds as follows: since there is no fleeting ‘a’, the first constraint is $\langle \text{BSNounE.A:am1s} \rangle$, it is rewritten as $\langle \text{alxkav.A:am1s} \rangle$ and is validated by a dictionary lookup as \textit{alxkav} is the adjective in positive form (a), masculine gender (m), nominative case (1), singular number (s). The second constraint is $\langle \text{wq$L.N+Dummy} \rangle$, it is rewritten as $\langle \text{wqosxcxu.N+Dummy} \rangle$ and is validated by a dictionary lookup as \textit{wqosxcxu} is the form of a dummy noun. Since both lexical constraints are validated, entry is computed in the DELAF form: lemma is obtained by concatenating the adjective stored in the variable $\text{SB}$ with the derivational suffix $\text{–ost}$ (variable $\text{NounE}$ contains empty string), any possible syntactic and semantic features are inherited from the adjective (lemma met by the first constraint), while the inflectional features are inherited from the dummy entry (lemma met by the second constraint).

As pointed in section 4, derivational processes are characterized by various phonological alternations. However, in order to impose lexical constraints on derived forms it is necessary to reverse (undo) the effect of one or more phonological alternations that could have taken place. Since many of these alternations occur in more than one derivational process they are implemented as subgraphs that are called by transducers that recognize certain derivational process (see Figures 2 and 3).
Figure 2: The subgraph palatalizacija.grf that reverses the palatalization. The value of the variable $P$ is the recognized consonant. The output of the transducer is the original consonant that could have palatalized or not and it is concatenated to the previously produced output. Different consonants can produce same output since palatalization is not obligatory.

The computation of DELAF entry can be rather complex, as seen in Figure 3. The recognized form of the augmentative is obtained by concatenating five variables, while its lemma is obtained by concatenating four variables and the final ending $-a$. Its syntactic features are inherited from the basic lemma (the first constraint). For instance, lemma *magarac* ‘donkey’ is marked by $+\text{Zool}$ as animal and this marker would be transferred to the augmentative form *magarčina* as well. Besides these inherited features, the new ones are added. First of all, derivational transducer with lexical constraints adds the marker $+D$ to all the derived forms they recognize. Next, the marker is added that characterizes the type of the derivational process, for instance $+\text{Aug}$ for augmentatives. Finally, additional markers can be added as well, as are $+\text{MG}$ and $+\text{FG}$ for the augmentatives recognized by the transducer in Figure 3. Namely, although the basic noun is masculine, the derived augmentative is feminine; however, it is used for both the
masculine and feminine sex beings, and this information is important for modeling the agreement constraints in syntactic processing.

Figure 3: Transducer that recognizes augmentatives derived from the masculine gender nouns with the suffix –čina. It uses the subgraph palatalizacija.grf that restores the palatalized consonant and suf-AC.grf that restores suffix –ac.

The syntactic and semantic markers that are assigned to the lemmas in DELAS dictionary are used in many transducers with lexical constraints in order to better qualify which lemmas can actually participate in the derivational process, and thus reducing the possibilities of false recognitions. The usage of these markers in lexical constraints is illustrated in Figure 4. For instance, nouns denoting animals (-Zool) and plants (-Bot) usually cannot be prefixed by numbers and prefixes. Also, prefix ne- is used only for those adverbs that were derived from adjectives (+Adj).
Figure 4: Transducer that recognizes various forms of prefixation of nouns, adjectives and adverbs.

All developed derivational graphs are assembled in one graph (see Figure 5) which is given the lowest priority. This is of the utmost importance, as their usage with the regular priority would create a numerous, often false, ambiguities.

All the occurrences of derived nouns, adjectives and adverbs recognized by the developed transducers can be retrieved from the text by using the pattern `<A+D>+<N+D>+<ADV+D>`. An excerpt from the concordances produced using this pattern on one of the subcorpora is:

SAD na prostoru Srbije i SRJ, sxe fica Stejt departmenta Madlen Olbrajt, pristal ticaja na lxudstvo ove nasxe planetice. Oni koji tvrde suprotno nisu u pravu. o pozivu, krojacyki, sad vecx kalifex, bio pred direktorom i urednikom "Politik" sko zaduzenxe da agitujem za neizlazak za savezne izbore, pa time i rusenxe otovo je za ne verovati da je pedesetvodisxnxak, gospodin Eli Malka, reditel Kakva su stremlxenxa u neuroracyunarstvu? Na cyemu se zasnivaju neuralne anxa u Zxenevi), zamislio je paukovu mrezxu, nicyim ograniencyenu, preko koje bi gde dobodosxli) - kad su na zemlixinom tlu - u oklopnova vozila i smestiti u ka
Trbojevic docyarava Cyngalovicxevo bdenxe nad svakom ulogom, svakim li Srb, Romi, Turci i drugi nealbanski nxivalx, stiglo se do brojke od milion onstataciju da su Hrvati, kao najmalobrojnih narod u BiH u podredxenom polož zxe je recy. Sve je visxe medxunarodnih ukrsxtanxa brakom, nxihov broj je ove xko povredxeno, a poginuo je tridesetosmogodisxeni Nenad Nikolicx o kome je pr e nemaju precya posla, nasxu namrsxtenosti i ozbilxnosti bez radosti. Pa do stinxaka. Nerado se vracxa u opustelu Belu kucxu - supruxa Hilari je u Nxujork nepotkuplxiva, opominxucxa, razdirucxih sxumova i tonova, sa minimalnim otkl RJ. Ispunxenxe te odredbe neopravdana je do sada blokirala medxunarodna mi.

Figure 5: Transducer Derivacija+.grf that calls transducers for particular derivational processes as subgraphs.

6 Results, problems, solutions

The developed transducers have been tested on the test subcorpora and the results presented in Table 3 show that the recognition rate has improved significantly. The number of recognized simple words has increased for 1.64% for literary text, 3% for newspaper texts, and 3.3% for economy texts.
Table 3: Recognition rate before and after application of derivational transducers with lexical constraints

<table>
<thead>
<tr>
<th>Corpus</th>
<th>total number of simple forms</th>
<th>different forms</th>
<th>recognized simple words before</th>
<th>recognized simple words after</th>
<th>s. words not recognized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newspaper</td>
<td>582K W</td>
<td>689/25 (9.8:1)</td>
<td>59103 (85.75%)</td>
<td>61155 (88.73%)</td>
<td>7770 (11.27%)</td>
</tr>
<tr>
<td>Literary</td>
<td>245K W</td>
<td>417/74 (5.9:1)</td>
<td>37246 (89.16%)</td>
<td>37914 (90.8%)</td>
<td>3860 (9.24%)</td>
</tr>
<tr>
<td>Economy</td>
<td>404K W</td>
<td>407/54 (9.9:1)</td>
<td>37124 (91%)</td>
<td>38419 (94.3%)</td>
<td>2335 (5.73%)</td>
</tr>
</tbody>
</table>

The chart in Figure 1 shows to what extent certain derivational processes contribute to the set of unrecognized words, determined by the capability of the derived derivational transducers to retrieve them. It can be seen that the derivational processes that do not change either the PoS of the basic word or its inflectional class, that is negation, adjective superlatives, prefixation, and compound formation, are by far the most frequent in all three subcorpora. Derivation of possessive adjectives is the most frequent among the remaining derivational processes. The usage of the modeled derivational processes is much less frequent in literary texts than in non-fiction. However, it can be seen that in the former texts diminutives contribute more to the unrecognized words than they do in the later.
Although the usage of the developed derivational transducers has proven to be useful, there are still some problems to be solved. These problems basically fall in two categories: either some simple word has been erroneously recognized as a derivative or some derivative has not been recognized although it was derived by some modeled process.

The most frequent problems occur when certain lemmas are missing from the e-dictionary that leads to their erroneous recognition by the derivational graphs. That is the case, for instance with the noun forms demantiju, the dative form of demanti ‘official denial’, and Redyepa, genitive form of the personal name Redžep, which, as missing from the dictionary, were analyzed as:

\[ \text{demantiju}, \{\text{de}, \text{de.PREF}\} \{\text{mantiju}, \text{mantija.N600:fs4q}\} \]
\[ \text{Redyepa}, \{\text{Re}, \text{re.PREF}\} \{\text{dyepa}, \text{dyep.N81:ms2q}\} \]

The straightforward solution of this problem is to add the missing lemmas to the dictionary. With the improvement of the dictionary itself this kind of problem will become less severe, but to some extent it will always exist. The more sophisticated solution would take into account stronger constraints, such as which prefixes can be used with
a certain type of nouns. In the given example, for instance, prefixes *de-* and *re-* cannot be used with concrete nouns as *mantija* ‘cassock’ and *džep* ‘pocket’.

In some cases the derived forms are ambiguous themselves, as in the following example:

\[
\text{Arijetina,}\{\text{Arija,arija.N600:fs1q:fp2q}\}\{\text{Arijetina.N+Aug+D:fs1q}\}
\]

\[
\text{Arijetina,}\{\text{Arijeta,Arijeta.N622+Hum+NPpos+First+Pos+D:fs1v}\}
\]

\[
\{\text{Arijetin.A+Pos:akms2g:akms4v:ae}\}
\]

The simple word form *Arijetina* is correctly associated with two 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*. This kind of occurrences cannot actually be regarded as problems since they express the inherent ambiguity. However, they can be avoided if some transducers that model less frequent derivational processes, as augmentatives in the previous example, are not used in text analysis, or if they are used with the lowest priority, the possibility offered in NOOJ, but not in Intex (M. Silberztein, 2004a).

In other cases the ambiguities from the e-dictionaries are reproduced in the derived forms, offering some erroneous solutions, e.g.

\[
glasnogovornici,\{\text{glasno,glasan.A...}\}\{\text{govornici,govornik.N+Hum:mp1v}\}
\]

\[
glasnogovornici,\{\text{glasno,glasan.A...}\}\{\text{govornici,govornica.N:fs3q:fs7q}\}
\]

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. Similar problems arise in cases when one prefix is the head of the other, as in the following examples:

\[
\text{neopravdan,}\{\text{ne,ne.PREF}\}\{\text{opravdan,opravdan.A17:akms1g...}\}
\]

\[
\text{neopravdan,}\{\text{neo,neo.PREF}\}\{\text{pravdan,pravdan.A1:akms1g...}\}
\]
Here again only the first tokenization is correct. However, both
tokenizations yield same and correct lemma and grammatical
categories. The way to solve this problem is to develop more elaborate
semantic constraints, which for the given examples would not be quite
clear. In order to avoid false ambiguities the straightforward and best
solution seems to be to enhance the dictionary with lemmas causing
this kind of problem.

The most frequent cases when derivational transducers fail to
recognize correctly some derivational forms occur when one
component of the compound is not in the dictionary. For example, the
adjective krpnooki ‘having big eyes’ has been erroneously analyzed as
a compound noun, with two possible lemmas krupnooka and
krupnooko (neither of them existing), derived from the adjective
krupan ‘large’ and nouns oka ‘the old weight unit’ and oko ‘eye’
respectively:

krupnooka,{krupno,krupan.A18:aens1g…}{oka,oka.N600:fs1q:fp2q}
krupnooka,{krupno,krupan.A18:aens1g…}{oka,oko.N308:ns2q}

This situation occurs as adjective oki ‘having eyes’ is not in the
dictionary. This adjective though used only in the compounds
(plavooki ‘having blue eyes’, etc.) could still be added in the special
dictionary of dummy entries. The other solution would be to add all
the possible compounds in the regular dictionary, since there are not
many of them. Different situation occurs with the adjective
cyetvorocylan ‘having four members’ that was tokenized as
cyetvorocylanu,{cyetvoro}{cylanu,cury.N:ms3q:ms7q}
since cylanu is dative and locative form of the noun cylan ‘member’
while adjective cylan ‘member’ is not in the dictionary and is actually used only
in the compounds. In this case the only solution is to add this adjective
in the dictionary of dummy entries since many compounds can be
formed with other adjectives (e.g. mnogocylan ‘having many
members’ vissecylan ‘having several members’, etc.) and practically
any number.

Interesting are the cases when transducers fail to recognize the
derived words due to the different capitalization used in the basic
word and the derived form. In the newspaper subcorpus the noun
forms nealbanci ‘non-Albanians’, nemadxari ‘non-Hungarians’,


neromi ‘non-Roms’ have not been recognized since nationalities are in Serbian orthography written with upper-case initial letter; however, when negated with prefix ne- this rule can no longer apply. This problem could be solved by modifying the transducers to restore the initial capital in the appropriate lexical constraint.

In some cases some inflected forms of the feminine gender nouns are ambiguous with some other inflected forms of the masculine gender nouns they were derived from, e.g. nastavniči is both the form of nastavnica ‘woman teacher’ and nastavnik ‘man teacher’:

\[
\begin{align*}
nastavniči, nastavnica.N651+Hum:fs3v:fs7v 
nastavniči, nastavnik.N10+Hum:mp1v:mp5v
\end{align*}
\]

As a result, the transducers when used with the lowest priority would not recognize the feminine gender forms. Since this problem occurs systematically for large number of lemmas produced by gender motion, the only way to solve it is to use the transducers that model this process with the regular priority.

The developed transducers and the way they are implemented (see Figure 3) do not allow the recognition of the forms produced by the iterative application of some derivational processes, such as naj[-]ne[-]razvijeniji ‘the most undeveloped’ and poštarkin ‘belonging to the post-woman’. At this moment it is not clear how these cases could be handled with the available tools.

7 Conclusions

The results obtained on test subcorpora are promising, but there are still points where the applied approach can be improved. First of all, more derivational process can still be added, to mention only the most frequent ones such as derivation of relational adjectives from nouns and nominalization form verbs (using the suffix –acija). The examples of simple words of this type that were not recognized in the test subcorpora are relational adjectives agresorski ‘in the way of an aggressor’ and agronomski ‘in the way of agronomy’ and the abstract nouns denoting an action militarizacija ‘the act of militarizing
something and *demonizacija* ‘the act of making a demon from somebody’. Next, more refined syntactic and semantic markers should be added to dictionary entries in order to be able to define more versatile lexical constraints in transducers. Finally, transfer from Intex to NOOJ will enable more precise tuning of the priorities of the used transducers which will add to the recognition precision.

**References**


Silberztein Max D. 2004a. NooJ v1.0. [in this volume]


