Serbo-Croatian Hyphenation: a TEX Point of View

Cvetana Krstev

On Serbo-Croatian

Serbo-Croatian is one of the South-Slavic languages. It is characterized, as other Slavic languages, by a rich morphology. A particular feature of the language is its almost fully phonological orthography, i.e. on a word level, one letter corresponds to each phoneme and vice versa. As a result, the written text practically represents a phonemic transcription of speech. Still, the Serbo-Croatian literary language has two main pronunciations, ekavian and jekavian, which reflect the different development of the pronunciation of the old Slavic sound ѣ. Sound ѣ is usually replaced by vowel e in ekavian dialect (for instance, dete, mleko, večan, čovek) while in jekavian dialect it is usually replaced either by two-syllable group ije (dijete, mlijeko) or by one-syllable group je (vječan, čovjek). Those differences in pronunciation are recorded in the written text. Accent has a distinctive role in Serbo-Croatian and as it is not marked in written texts there is a number of homographs.

Two alphabets are in use: Latin and Cyrillic. The Serbo-Croatian Latin alphabet is different from the English alphabet. Both letters with diacritics — č, ć, ž, š, đ — and digraphs — dž, lj, nj — are in use and they all have a separate place in the alphabet. The order of the Serbo-Croatian Latin alphabet is therefore as follows: a, b, c, č, ć, d, dž, đ, e and so on. As the letters q, w, x and y don't exist in the Serbo-Croatian alphabet, the total number of letters is 30. Transcription of foreign words and names is compulsory in Serbo-Croatian of ekavian pronunciation while jekavian pronunciation allows the orthography of the source language.

While all the letters with diacritics are assigned separate keys on the standardized national keyboard as well as the positions in the national version of 7-bit code [1, 2, 3], neither keys nor codes are provided for digraphs so they are input by striking two keys, i.e. by entering two codes. Besides that, although the standard provides a separate key for the letter đ, the keyboards of old typewriters often did not have it. As a result, this letter was — and sometimes still is — recorded as the digraph dj, in spite of orthographic rules.

Serbo-Croatian Cyrillic has the equivalent 30 letters but with neither diacritics nor digraphs. The order of the letters in the Serbo-Croatian Cyrillic alphabet is completely different from the order in the Latin alphabet. The Serbo-Croatian Cyrillic alphabet is also different from the Russian alphabet as there are letters which do not exist in Russian Cyrillic: ђ, ј, љ, њ, ћ, џ, and vice versa, which is important as the Russian Cyrillic was the basis for the development of appropriate international coding standards.

The digraphs of the Serbo-Croatian Latin alphabet can cause problems when using formatting and typesetting programs, particularly for hyphenation and automatic transcription from the Latin to the Cyrillic alphabet. These problems can be caused by each combination — lj, nj, dž and dj — which in the text may represent both digraphs and consonant clusters. A digraph is always transcribed into one Cyrillic letter and is never hyphenated. For instance, nadžak-baba is transcribed into наџак-баба and in both cases is hyphenated as na-džak-ba-ba. On the other hand, a consonant cluster is always transcribed into two Cyrillic letters and can, in principle, be hyphenated. For instance, nadživeti is transcribed into надживети and is hyphenated as nad-ži-ve-ti.

Serbo-Croatian Hyphenation Rules

Several sets of hyphenation rules for Serbo-Croatian were proposed on different occasions [4, 5], but to none of them did linguists gave unqualified support. Thus, the *Serbo-Croatian Orthography Book* [6] only gives the recommendations on how to hyphenate words, avoiding formulating precise rules. These recommendations can briefly be described as follows [7]:

(a) Two adjacent vowels should be divided, but it is not wrong if they are not. For instance, ža-oka or žao-ka.

(b) It is not allowed to carry over to the next line two or more final consonants without a vowel. For instance, not mlado-st but mla-dost.

(c) If there is only one consonant between two vowels, the consonant belongs to the second vowel and it is carried over with it to the next line. For instance, not bor-ac but bo-rac.

(d) If there are two or more consonants between two vowels, only those consonants that can be easily pronounced with the vowel that follows can be carried over to the next line (for instance, ze-mlja but also zem-lja). On the other hand, it is not recommended to carry over to the next line

a consonant cluster which is difficult to pronounce (for instance, not bu-mbar but bum-bar).

(e) If the constituent parts of a compound word can be distinguished, the break point is between those parts and each part is further hyphenated as if it were a separate word (for instance, raz-vući and raz-oružati). If those parts can't be distinguished, the word is hyphenated as if it were not compound (for instance, not raz-um but ra-zum). In both the examples the words are formed of prefix (raz-) and stem, but in the first case this prefix can be distinguished while in the latter case it can't be recognized any more, and the word is hyphenated according to the previous rules ((a)–(d)).

Although it follows from this rule that the recognized prefix can be further hyphenated as a separate word, it is considered a good typographic practice not to hyphenate a polysyllabic prefix. For instance, the word novootvoren with prefix novo- should be hyphenated novo-o-tvo-ren rather than no-vo-o-tvo-ren.

The recognition of vowels is fundamental for hyphenation in Serbo-Croatian, as hyphen positions often coincide with syllable boundaries and syllables are formed around the central phonemes which are usually vowels. The Serbo-Croatian alphabet, both Latin and Cyrillic, has five vowels: a, e, i, o, u. However, the phoneme r can take on the role of a vowel and be a central phoneme of a syllable in the following circumstances:

- in interconsonantal position (svr-stavanje);
- when preceding a consonant, at the beginning of a word (r-vanje);
- when following a vowel, in compounds (po-r-vati se).

Besides that, the sonants r, l, m and n when preceded by a consonant, at the end of a word, also behave as vowels (ma-sa-kr and bi-ci-kl). Marginal phonemes are consonants. Two types of syllables can be distinguished: open syllables, with the structure **-V-** or **-CV-**, where **V** is any vowel in the sense described above and **C** is any consonant, and closed syllables, with any other structure.

The application of recommendations (a)–(c) is almost self-evident and beyond questioning. Recommendation (d) refers to the identification of closed syllables which means that the consonant cluster has to be divided. Still, the *Serbo-Croatian Orthography Book* gives no strict rules for the division of consonant clusters, but only introduces the intuitive notions of consonant clusters that are easy or difficult to pronounce and illustrates them by few examples.

A semantic criterion, whose automatic implementation can be difficult to achieve, is introduced into hyphenation by recommendation (e). For example, ob- is a prefix in the word ob-istiniti but not in the word o-bi-čaj. The particular problem here is to decide whether the constituent parts of a compound word can be distinguished. The *Serbo-Croatian Orthography Book* gives no hints on how to make such a decision (for instance, whether the parts in the word preduzeti can be recognized). An additional problem to the application of this recommendation is posed by homographs. For instance, in the word podići, podiđem the prefix is pod- while in the word podići, podignem the prefix is po-.

Definition of Hyphenation Rules

Research into some aspects of consonant clusters in Serbo-Croatian has been undertaken before [8], but their occurrences in contiguous text have not been investigated. It was thus necessary to establish which consonant clusters do occur in Serbo-Croatian in order to state precisely the notions of easy- and difficult-to-pronounce consonant clusters. For that purpose, the analysis of consonant cluster occurrences has been undertaken that was based on the corpus of modern Serbo-Croatian texts of ekavian dialect [9]. In addition to frequency dictionaries that take into account the position of a consonant cluster in a word — initial, final or medial — results were obtained pointing out the occurrences of consonant clusters dj, dž, nj and lj as well as of those occurring only at the junction of the prefix and the word stem.

This analysis has made it possible to formulate the following rules for consonant cluster division in Serbo-Croatian:

I Binary consonant clusters.

(a) Two consonants are carried over to the next line ($\text{-}C_1C_2V$) only if they are usual at the beginning of a word in Serbo-Croatian, that is, if C_1C_2 belongs to one of the following sets:

1. $C_1 \in \{$s, z, š, ž$\}$ (C_1 is a fricative) and C_2 is any consonant; or
2. $C_1 =$ m and $C_2 \in \{$n, l, r$\}$; or
3. $C_1 =$ v and $C_2 \in \{$l, r$\}$; or
4. $C_1 \notin \{$r, l, lj, v, j, m, n, nj$\}$ and $C_2 \in \{$r, l, lj, v, j, m, n, nj$\}$ (C_2 is a sonant).

(b) In all other cases, one consonant must be left in the current line ($C_1\text{-}C_2V$).

II Three-member consonant clusters.
(a) Three consonants are carried over to the next line (-$C_1C_2C_3V$) only if they are usual at the beginning of a word in Serbo-Croatian (for instance, izdajni-štvo), that is if

$C_1 \in$ {s, z, š, ž} and
C_2 is any consonant, and
$C_3 \in$ {r, l, lj, v, j} (C_3 is a sonant);

(b) Two consonants are carried over to the next line (C_1-C_2C_3V) only if two consonants C_2C_3 are usual at the beginning of a word in Serbo-Croatian (for instance, rot-kva — see Ia).

(c) One consonant is carried over to the next line (C_1C_2-C_3V) only if the two consonants that remain in the current line (C_1C_2) are usual (or possible) at the end of a word in Serbo-Croatian (for instance, funk-cija), which means that C_1C_2 belongs to one of the following sets:

1. $C_1 \in$ {r, l, lj, v, j, m, n, nj} (C_1 is a sonant) and
   C_2 is any consonant; or
2. $C_1 \in$ {s, z, š, ž, f, h} (C_1 is a fricative) and
   $C_2 \notin$ {r, l, lj, v, j, m, n, nj} (C_2 is not a sonant); or
3. $C_1C_2 \in$ {ps, bz, ks, gz, pt, bd, kt, gd}.

III Four-member consonant clusters.
(a) Three consonants are carried over to the next line (C_1-$C_2C_3C_4V$) only if these three consonants are usual at the beginning of a word in Serbo-Croatian (see IIa). For instance, demon-stracije.

(b) Two consonants are carried over to the next line (C_1C_2-C_3C_4V), only if the two consonants which are left in the current line are usual at the end of a word in Serbo-Croatian (see IIc) and if the two consonants which are carried over to the next line are usual at the beginning of a word in Serbo-Croatian (see Ia). For instance, student-ski.

Moreover, it should be stressed that no four-member consonant cluster was identified on the analyzed corpus that would not have the structure (a) or (b).

In order to enable the implementation of recommendation (e) given by the *Codex* it was necessary to analyze the use of prefixes in Serbo-Croatian. For that purpose a list of 79 common Serbo-Croatian prefixes of 2 or more letters in length was composed on the basis of traditional Serbo-Croatian grammar textbooks, consisting of 52 basic prefixes and 27 phonologically altered prefixes that occur as a result of phonological changes at the junction between the prefix and the word stem. For instance, the basic prefix iz- has three altered prefixes: is- (in front of p, t, k, f, c, h), iž- (in front of đ and dž) and iš- (in front of č and ć).

The occurrences of all these prefixes were analyzed on the corpus of modern Serbo-Croatian texts of ekavian dialect [10]. For each prefix string, it was thus possible to establish its productivity and its ability to be combined with other prefixes. Also, such prefix strings were identified that are prefixes in some, but not all, cases (for instance, ob- is a prefix in obuhvatiti but not in običaj). Finally, instances were identified where the prefix is not the longer but the shorter prefix string (for instance, naj- is the prefix in najelegantnija while in najednom the prefix is na-).

The Serbo-Croatian prefixes were, exclusively for the purpose of hyphenation, classified on the basis of this analysis in such a way that the recognition of prefixes of the same group requires the same conditions. The prefix types are as follows:

1. **The break point is always after a prefix string of type 1.** A prefix was assigned type 1 either because all instances of the prefix string in the corpus were prefixes or because it never occurred. For all the prefix strings that never occurred in the corpus the additional check was done in the Serbo-Croatian dictionary that supported the decision to categorize them as type 1. For instance, the prefixes of type 1 are anti- (antihrist) — always a prefix — and iž- (iždžikljati) — never occurred.

2. **The break point is after a prefix string of type 2, except when a difficult-to-pronounce consonant cluster follows it.** A prefix falls into this group if it is monosyllabic and ends with a vowel. The prefix za- is of type 2 (za-svoditi, but zač-koljica because čk is a difficult-to-pronounce consonant cluster, that is, condition Ia is not satisfied).

3. **The break point is after a prefix string of type 3, except if it is followed by a vowel, when additional information is needed.** All prefixes of type 3 end with a consonant. If a prefix string of type 3 is followed by a consonant, it cannot be wrong to break the word after it, because the prefix string without that last consonant can't be a prefix for various reasons. For instance, eks- is a type 3 prefix because ek- is not a prefix at all. On the other hand, naj- is a type 3 prefix although na- can be a prefix, but as no initial consonant cluster beginning with consonant j exists, na- can't be a prefix when the prefix string naj- is followed by a consonant. On the other hand, eks- and naj-

are not prefixes in the words ekser and najaviti, respectively.

4. **Additional information is always necessary to decide whether a prefix string of type 4 is a prefix.** Prefix strings of type 4 are polysyllabic. If we consider the break point after the prefix to have a higher priority than other possible break points and if, on the other hand, we do not want to miss any break point, we can never tell beforehand if the first break point in a word is after a prefix string of type 4. For instance, polu- is a prefix in the word polu-pismen while in the word polupati the prefix is po-. In the former case the break point after po- is possible but should be avoided due to typographical conventions, while in the latter case the break points after po- and polu- have equal value.

5. **Additional information is always necessary to decide whether a prefix string of type 5 is a prefix, except when its final consonant and consonants that follow form a difficult-to-pronounce consonant cluster.** Prefixes of type 5 end with a consonant and for all of them the prefix string without the final consonant can also be a prefix. We can be sure that a word can be hyphenated after a longer prefix string only if the final consonant and consonants that follow form a difficult-to-pronounce consonant cluster, no matter whether that prefix string is really a prefix. For instance, the word ob-zidati must be hyphenated after ob- because bz doesn't satisfy the condition Ia (and ob- is actually a prefix). On the other hand, ob- is a prefix in ob-lizati while in o-bližnji the prefix is o-.

6. **The additional information is always necessary to decide whether a prefix string of type 6 is a prefix, except when its final consonant and consonants that follow form the difficult-to-pronounce consonant cluster.** Prefixes of type 6 end with a consonant and are followed by a consonant, as they are all phonologically altered prefixes. Also, for all of them a prefix string without the final consonant can also be a prefix. As for the prefixes of type 5, we can be sure that the longer prefix string is a prefix in a word only if the final consonant and consonants that follow form a difficult-to-pronounce consonant cluster. For instance, the word is-psovati must be hyphenated after is- as sps isn't an initial consonant cluster in Serbo-Croatian. On the other hand, is- is a prefix in is-kititi while in i-skakati the prefix is i-.

Production of Hyphenation Patterns

For Serbo-Croatian, as for some other languages — French and Polish, for example — dictionaries that indicate the hyphenation break points for all the entries do not exist [12, 13]. Thus, in order to apply the PATGEN program [14] for the automatic generation of a pattern dictionary to be used with the typesetting system TEX [15], it would be necessary to include that information in some machine-readable dictionary of Serbo-Croatian. But that would not be enough as this dictionary would have to contain, in addition to the usual dictionary entries, all the derived forms. As Serbo-Croatian is a language with a very rich morphology, this extended dictionary would be at least ten times larger than the original one, and it would be time-consuming and erroneous to prepare it for PATGEN.

For this reason, the decision was made to produce the pattern dictionary "by hand", founding the generation on the precise hyphenation rules described in previous sections and checking the obtained results on the sample words extracted from the existing Serbo-Croatian paper dictionaries. In this section will be described the procedure which generates the pattern dictionary to be used for Serbo-Croatian texts of the ekavian dialect that use digraphs dj, lj, nj and dž encoded as two separate codes [11].

(a) The break point is between two vowels (recommendation (a) of the *Orthography Book*). From this rule, 25 patterns were generated of the form

$$V1V,$$

where **V** belongs to the set {a, e, i, o, u} of "real" vowels.

(b) The break point is before a consonant surrounded by two vowels (recommendation (c) of the *Orthography Book*). From this rule, 105 patterns were generated of the form

$$1CV,$$

where **C** belongs to the set $\mathcal{A} \setminus \{$a, e, i, o, u, dj, nj, lj, dž$\}$, where $\mathcal{A}$ denotes Serbo-Croatian alphabet.

(c) As the analysis of consonant cluster occurrences showed, the strings dj, nj, lj, dž are much more frequently digraphs than consonant clusters. Therefore, the following four patterns were produced which disable the division of digraphs:

1d2j 1n2j 1l2j 1d2ž

The same analysis also showed that the consonant clusters dj, nj, lj and dž occur only at the junction of a prefix with a stem, so these cases will be covered in items (j)–(o). At this point, only 4 patterns were

added for the identification of prefixes in- and kon- which had not been included in the prefix analysis as they are not usual in Serbo-Croatian:

```
.in3jekc        .in3junkt
.ko2n3jug       .ko2n3junk
```

(d) The break point is not between two consonants of a binary consonant cluster if this consonant cluster is usual at the beginning of Serbo-Croatian words (rule Ia for consonant cluster division). From this rule, 205 patterns were generated of the form

$$1C_1C_2$$

where C_1 and C_2 are such consonants that C_1C_2 satisfies rule Ia.

Having in mind that consonant clusters mnj and vlj, whose existence is confirmed on the corpus, do not satisfy rule Ia, in contrast to mn and vl, two more patterns were introduced:

```
2m3nj           2v3lj
```

The patterns of form (c) function as desired if a digraph is followed by a vowel. On the other hand, if the digraph is followed by a consonant, the break point before the digraph has to be disabled in all the cases where the resultant consonant cluster does not satisfy rule Ia. So, 11 more patterns were added:

```
2ljn    2ljk    2ljs    2ljš
2ljc    2ljd    2ljb    2njs
2njc    2djs    2dž3b
```

The generation of these patterns was based on the results of the analysis of consonant cluster occurrences.

(e) The phoneme r between two consonants behaves as a vowel, which means that the break point can, in principle, be after r. Analysis of the occurrences of the phoneme r in interconsonantal position, as part of the analysis of consonant cluster occurrences, showed that only 106, of 576 possible strings of the form $C_1\mathtt{r}C_2$, were actualized. Besides that, the actualized strings had such a form that the pattern $1C_1\mathtt{2r}$ was generated in step (d) while pattern $\mathtt{1r}2C_2$ was not generated at all. When pattern $1C_1\mathtt{2r}$ is applied to string $C_1\mathtt{r}C_2$, the obtained result is

$$1C_1\mathtt{2r}C_2$$

and that is precisely the pattern necessary to identify the phoneme r in interconsonantal position — no new patterns need to be generated.

(f) The break point is before a three-member consonant cluster that is usual at the beginning of Serbo-Croatian words (rule IIa). It may seem that the implementation of this rule requires the generation of 460 patterns of the form

$$1C_12C_22C_3$$

However, analysis of consonant cluster occurrences showed that C_2 is actualized only as a plosive consonant (p, b, k, g, t, d), the fricative f, the affricate c or č, or the sonant m or v. That means that patterns of the form $1C_12C_2$ and $1C_22C_3$ have already been generated in step (d) as $C_1 \in \{\mathtt{s}, \mathtt{š}, \mathtt{z}, \mathtt{ž}\}$, $C_2 \notin \{\mathtt{r}, \mathtt{l}, \mathtt{lj}, \mathtt{j}, \mathtt{n}, \mathtt{nj}\}$ and C_3 is a sonant. When these two patterns are applied, the necessary pattern $1C_12C_22C_3$ is obtained. This means that for implementation of this rule no new patterns need to be generated.

(g) The break point in a three-member consonant cluster that is not usual at the beginning of Serbo-Croatian words is between the first and the second consonant only if the two last consonants of the cluster are usual at the beginning of Serbo-Croatian words (rules IIb and IIc). The three-member consonant cluster $C_1C_2C_3$ that can match no pattern of form (f), while C_2C_3 cannot match any pattern of form (d) either, will be divided as C_1C_2-C_3, because C_3V matches the pattern of form (b). This means that for the implementation of these rules no new patterns need to be generated either.

However, the analysis of consonant cluster occurrences, as well as the interactive checking of the correctness of the generated patterns, showed that there are some cases when it is better to divide a three-member consonant cluster as C_1C_2-C_3, although C_2C_3 is usual at the beginning of Serbo-Croatian words, which means that pattern $1C_22C_3$ exists. Those cases occur mainly in the words of foreign origin and very often in derivative forms. For instance, it is better to hyphenate konkursni as konkurs-ni than as konkur-sni, or tekstil as teks-til than as tek-stil. In order to cover these cases too, 6 more patterns were generated:

```
ur2s3n      k2s3t       k2t3n
l2t3n       n2t3n       or2f3n
```

(h) The majority of four-member consonant clusters in Serbo-Croatian appear at the junction of a word stem with suffixes -sk(i) and -stv(o). Namely, the analysis of consonant cluster occurrences showed that only 3 of the 22 four-member consonant clusters that were identified on corpus did not have the form **CCsk** or **Cstv**. The breaks between the stem and the suffix will thus be provided by the patterns of form (d) and (f) respectively. The remaining identified consonant clusters are also correctly handled, as they match the patterns of type (f) or the appropriate patterns for prefix recognition: nstr in demon-stracija is controlled by the pattern of type (f) while kspl

in eks-plozija and jstr in naj-strašniji are controlled by the patterns of form (l) for prefixes eks- and naj- respectively.

(i) The break point cannot be before the final consonants (recommendation (b) of the *Orthography Book*). The patterns of form (b) provide that at least one vowel is carried over to the next line. However, the patterns of form (d) and (f) that match the consonant clusters satisfying the conditions Ia and IIa respectively would allow the break points before these consonant clusters at final positions as well. The analysis of consonant cluster occurrences showed that only 4 of the 33 final consonant clusters that were identified on corpus match the patterns of form (d). In order to overcome this problem it would thus be necessary to add 4 more patterns:

```
2st.    2sl.    2št.    2kl.
```

But as the hyphenation routine of TEX hyphenates the words in a way that enables at least three letters to be carried over to the next line, these patterns were not included in the pattern dictionary for Serbo-Croatian.

Aside from these final clusters, the special case of the phonemes r, m, n and l at the end of a word behind a consonant should be mentioned. An example is the word masakr, in which the final r behaves as a vowel. But the final consonant cluster kr matches the pattern of form (d), which means that this word would be correctly hyphenated: ma-sa-kr.

Before we continue to describe the generation of patterns that control the hyphenation at the junction of prefixes and word stems, it should be noted that the pattern dictionary produced thus far has 362 patterns.

(j) Prefix strings of type 1 are always prefixes. In other words, the break point is always after a prefix string of type 1. For the 27 prefixes of type 1, 36 patterns were generated which make break points after the prefix strings possible. For instance, for the prefix anti- only one pattern was generated, .an2ti, while for the prefix beš-, which is a variant of the prefix bez- that is realized in front of č and ć, two patterns were generated: .be2š3č and .be2š3ć. On the other hand, for the prefixes nat-, op-, ot-, pot- and pret-, which are also variants of the basic prefixes, no pattern was generated as the hyphenation after these prefixes is controlled by the patterns for consonant cluster division (steps (a)–(i)).

(k) The question whether a prefix string of type 2 is a prefix in a word or not, does not influence word hyphenation. So, for the 19 prefixes of type 2 no pattern was generated.

(l) The break point is after a prefix string of type 3, except when it is followed by a vowel when additional information is needed. The generation of patterns for prefixes of this type will be illustrated on the example of the prefix naj-, which has a high frequency in Serbo-Croatian as it is used to express the superlative of adjectives. First of all, the prefix .na2j3 is generated. Then, the cases when naj- in front of a vowel is not a prefix have to be covered, because naj- can, in principle, be a prefix for all Serbo-Croatian adjectives that begin with a vowel. So, for instance, the patterns .na3j4av (for words as na-ja-vi-ti or na-jav-lji-va-či-ca) and .na3j4el (for words as na-je-la or na-je-lo) are generated. Finally, as the adjectives avetinjski and elementaran, which have superlatives exist, the patterns .na4j5avet and .na4j5elem had to be generated. In that way, 19 patterns were generated for the prefix naj- and a total of 108 patterns for the 9 prefix strings of type 3.

(m) Additional information is always necessary in order to decide whether a prefix string of type 4 is a prefix in a word or not. It should be remembered that all prefixes of type 4 are polysyllabic and, despite the recommendation given by the *Orthography Book*, it is usually considered better not to hyphenate the prefix itself. Patterns should, therefore, prevent the hyphenation of a prefix string in the cases when it is a prefix. The generation of patterns for prefixes of type 4 will be illustrated on the example of the prefix preko-, that occurs, for example, in words prekomerno or prekosutra. However, the analysis of occurrences of the prefix string preko- in the corpus, as well as a lookup in Serbo-Croatian dictionaries, showed that the instances where pre-, rather then preko-, is a prefix, are more frequent. Some examples are pre-koračiti, pre-kositi, and pre-komandovati. Because of that, only the patterns that provide the correct hyphenation in the cases when preko- is prefix were entered in the pattern dictionary. For instance, for the examples given above two patterns were generated: .pre2kome and .pre2kosu. Similarly, 9 patterns were generated for the prefix preko-, and a total of 59 patterns for the 14 prefix strings of type 4.

(n) The break point is compulsory after a prefix string of type 5 only if the last consonant of the prefix string and the initial consonants of the word stem form a difficult-to-pronounce consonant cluster. In all other cases, additional information is needed. The generation of patterns for prefixes

of type 5 will be illustrated on the example of the prefix od-, that occurs, for instance, in words odraditi or odsvirati. In view of the fact that the analysis of occurrences of this prefix string in the corpus, as well as a lookup in Serbo-Croatian dictionaries, showed that it is a very frequent prefix, pattern .od3 was entered in the pattern dictionary. However, in front of the vowel i the prefix string od- very often is not a prefix (for instance, in words odignuti or odista), so the pattern .od4i is generated. At the end, in the word odigrati the prefix string od- is a prefix, so the pattern .od5igr is added. In that way, 35 patterns were generated for the prefix od-, and a total of 151 patterns for the 6 prefixes of type 5.

(o) Prefix strings of type 6 are similar to the prefix strings of type 5, except that, being variants of the basic prefixes, they can be prefixes only if they are followed by a consonant of a certain kind. The generation of patterns for the prefixes of type 6 will be illustrated on the example of the prefix ras-, which occurs, for instance, in words rastumačiti or rascvetati. The prefix ras- is a variant of the prefix raz- that emerges from the substitution of the voiced consonant z by its unvoiced counterpart s in front of the unvoiced consonants p, k, t, f, c or h. Thus, for the prefix ras- we introduce the patterns

.ra2s3p	.ra2s3k	.ra2s3t
.ra2s3f	.ra2s3c	.ra2s3h

However, ras- is not always a prefix in front of these consonants. Moreover, in words rastaviti and rastezati, for instance, the prefix is ra- and therefore the patterns .ra3s4ta and .ra3s4te are added. At the end the pattern .ra4s5tanj is generated because the word rastanjiti has the prefix ras-. Similarly, for the prefix ras- 20 patterns were generated, and a total of 82 patterns for the 4 prefixes of type 6.

(p) Combinations of two, rarely three, prefixes can be identified in Serbo-Croatian. Examples of the latter case are, for instance, is-po-raz-boljevati or po-iz-o-stavljati. For the solution of the hyphenation problem, only those combinations are interesting for which additional information is needed to identify the second prefix. The analysis of occurrences of a combination of prefixes in the corpus, as well as a lookup in Serbo-Croatian dictionaries, showed that besides the very frequent prefixes naj-, which expresses the superlative of adjectives, and ne-, which expresses the negative form of nouns, adjectives and adverbs, prefixes that combine with other prefixes are o-, po-, pro-, za- and novo-. The generation of patterns needed to identify the combination of prefixes will be illustrated on the example of the combination o-bez-, which occurs in the words obezvrediti and obezglaviti, while in the word obezubiti the combination o-be- appears. Bearing in mind that patterns .ob3 and .ob4e were generated in step (n), two patterns, .obe2z3v and .obe2z3g, were added for the examples given above.

To the prefixes naj- and ne- a partial solution was applied. For prefix naj- only the patterns that correspond to the combinations of the prefix naj- with prefixes that participate in adjectival derivation were generated. Similarly, for the prefix ne- only patterns that correspond to the combinations of the prefix ne- with prefixes that participate in adjectival, adverbial and nominal derivation were generated. At the same time, for the second prefix in each combination, only the patterns that "reflect the rule" were taken into consideration. For example, for the combination naj-bez- (for instance, in words najbezgrešniji and najbezvoljniji) only one pattern was generated .najbe2z3 while the other patterns generated for the recognition of the prefix bez- were *not* taken into account. In this manner, for the recognition of the combination of prefixes 90 patterns were generated.

(r) The pattern dictionary was amended with an exception dictionary which contained only seven words. The words were added to the exception dictionary for one of these two reasons:

- The pattern dictionary would have to be expanded with a pattern that matches only one word. Such is the case of the undeclined word po-dne which due to the pattern .po2d3 generated in step (o) wouldn't otherwise have any break point.
- The word is a homograph and its break points depend on the meaning. For instance, the word form uzori can be the nominative plural of the noun uzor, in which case the break points are u-zo-ri, or the second person singular of the imperative of the verb uzorati, which has a prefix uz-, and in that case the word should be hyphenated as uz-ori. Such a word forms are added to the exception dictionary to suspend all the break points.

The complete pattern dictionary has 888 patterns. The highest coefficient that appears in some pattern is 5. The analysis of occurrences of prefixes of a particular type in the corpus showed that the 362 patterns generated in steps (a)–(i) and which reflect the hyphenation rules can be expected to

provide the breaks for approximately 97% of the words in some document.

Conclusion

TEX has been in steady use at the Faculty of Mathematics at the University of Belgrade for several years now. The Latin alphabet is predominant, but Cyrillic has been used too. However, the Serbo-Croatian version of TEX has not been produced yet, in the sense that there are no font tables, neither for the Latin nor for the Cyrillic alphabet, that would reflect the national standard 7-bit codes. This means that for the Latin alphabet the commands `\v` and `\'` are used to set the diacritics. It is well known that words containing such diacritics can't be hyphenated by TEX. For this reason, the generated pattern dictionary has still not been included in any TEX implementation.

The validity of the generated pattern dictionary was, thus, tested with the command `\showhyphens`. As a Serbo-Croatian dictionary in machine-readable form was not available, chosen words found in corpus and traditional dictionaries were used as the arguments of this command. All the words were coded using only the letters of the English alphabet. The words were chosen in a way to reflect most of the problems of digraphs, consonant clusters and prefix recognition. On the basis of this test it can be said that the generated pattern dictionary provides word hyphenation according to the formulated hyphenation rules. However, the undertaken test can, by no means, be considered exhaustive enough and the true validity of the produced pattern dictionary will be confirmed only through regular use.

There are some aspects of Serbo-Croatian hyphenation that have not been covered by the performed analysis. First of all, the problem of compound words is still unsolved. As in Serbo-Croatian many compound words are formed by inserting the vowel `o-/-e-` between the constituent parts (for example, `glavo-bolja` or `gluvo-nem`), all those words will be correctly hyphenated by the existing rules. As for the compound words that are formed simply by connecting the constituent parts, many of them will be correctly hyphenated, as `krompir-čorba` or `star-mali`.

In the end, it should be stressed that this pattern dictionary was generated for use with the Latin alphabet which uses the four digraphs `dj`, `lj`, `nj` and `dž`. If the letter `đ` were used instead of the digraph `dj`, the dictionary would have to be extended. First of all, patterns of form (b) and (d) should be added: for example, patterns `1đa`, etc. and `1đr`, etc. Also, in all patterns for prefix recognition every occurrence of the string `dj` in which `d` and `j` are not separated by a digit should be replaced by `đ`. For instance, the pattern `.na4d5redj` should be replaced by `.na4d5ređ`. It should be noted that during the generation of patterns for prefix recognition, the digraphs were always treated as separate letters. For instance, two patterns were generated — `.na4d5red` and `.na4d5redj` — although one pattern, `.na4d5red`, would have been enough. This strategy would facilitate the replacement of a digraph with a separate letter. Of course, all the patterns that refer to the digraph `dj` could be deleted, but that is not necessary at all. Namely, for the hyphenation routine of TEX, the letters `đ` and `dj` could be considered as two different letters that can occur in the same document, which sometimes even happens.

If the Cyrillic alphabet were used instead of the Latin alphabet, then the same procedure that was suggested for the replacement of the digraph `dj` by the letter `đ` should be applied to the letters љ, њ and џ. Of course, the patterns that refer to digraphs should be deleted as they would have no meaning any more.

It should be stressed once again that the obtained results apply only to Serbo-Croatian of ekavian dialect as some crucial results are based on the analysis of corpus of texts of ekavian dialect. However, besides some small lexical differences, the differences between the two dialects are mostly the result of the different pronunciation of the sound ђ. As a result, in jekavian dialect some new consonant clusters may be introduced: for instance, `cvjetova` (jekavian) *vs.* `cvetova` (ekavian) or `snjegova` (jekavian) *vs.* `snegova` (ekavian). According to hyphenation rule Ia, most of those consonant clusters wouldn't be hyphenated which is in this case appropriate. Nevertheless, the produced pattern dictionary should be applied to the texts of jekavian dialect only with due precaution.

Version 3.0 of TEX introduces some new possibilities. For example, with TEX 3.0 the user can specify the smallest number of letters that may be left in a current line and the smallest number of letters that may be carried over to the next line. For Serbo-Croatian this is an important improvement, as the old restrictions lead to situations where many 6-letter words with 2 legal break points would not be hyphenated (for instance, `u-sko-ro`) and even some 7-letter words, as `u-stre-li`. If some user would want to set the smallest number of letters that may be carried over to the next line to 2, the

only change that has to be made in a pattern dictionary is the addition of the 4 patterns generated in step (i). The further reduction of the limits would, however, require the detailed check of all patterns, especially those for the prefix recognition.

Acknowledgement

I want to express my gratitude to Dr. Janusz Bień, from the Institute of Informatics at Warsaw University, for all the support and information he gave me during my work as well as for the careful reading of the manuscript. I also owe a lot to Prof. Jacques Désarménien from "Laboratoire de typographie informatique" at the University Louis-Pasteur, Strasbourg, whose work in solving the hyphenation problem for French guided me during my research. In addition, I thank Prof. Ljubomir Popović, from the Faculty of Letters at Belgrade University, who helped me define the hyphenation rules for automatic application.

References

[1] International Organization for Standardization, *Information processing—ISO 7-bit coded character set for information interchange*, ISO 646, (1983)

[2] Savezni zavod za standardizaciju, *Skup znakova za razmenu podataka kodiranih sa 7 bitova za srpskohrvatsko latinično pismo*, [7-bit coded character set for Serbo-Croatian Latin alphabet for information interchange], JUS I.B1.002, (1986)

[3] Savezni zavod za standardizaciju, *Jedinice za unos podataka—Tastatura sa 47 tipki za slovenačko i hrvatsko latinično pismo*, [Data entry units—Keyboards 47 keys for Slovenian and Croat Latin alphabet], JUS I.K1.002, (1986)

[4] Belić, A., *Pravopis srpskohrvatskog književnog jezika*, [Standard Serbo-Croatian Orthography Book], Belgrade, pp. 15–18, (1934)

[5] Stevanović, M., *Savremeni srpskohrvatski jezik*, [Contemporary Serbo-Croatian Language], Belgrade, pp. 152–156, (1964)

[6] *Pravopis srpskohrvatskog književnog jezika*, [Standard Serbo-Croatian Orthography Book] Novi Sad - Zagreb, 130–131, (1960)

[7] Vitas, D., *Podela na slogove srpskohrvatskih reči*, [Syllable Division of Serbo-Croatian Words], Informatika, Ljubljana, 3/1981

[8] Tolstoja, S. M., *Načal'nye i konečnye sočetanija soglasnyh v serbsko-horvatskom jazyke*, [Initial and Final Consonant Clusters in Serbo-Croatian], Issledovanija po serbsko-horvatskom jazyku, Moskva, pp. 3–38, (1972)

[9] Krstev, C., *Frekvencijski rečnik konsonantskih grupa u srpskohrvatskom jeziku i problem rastavljanja na slogove*, [A Frequency Dictionary of Consonant Clusters in Serbo-Croatian and the Problem of Syllabification], Proceedings of the 2. Scientific Meeting "Computer Processing of Linguistic Data", Institute "Jožef Stefan", Bled, pp. 389–404, (1982)

[10] Krstev, C., *Rastavljanje reči srpskohrvatskog jezika na kraju retka*, [Hyphenation of Serbo-Croatian words at the end of the line], Proceedings of the 3. Scientific Meeting "Computer Processing of Linguistic Data", Institute "Jožef Stefan", Bled, pp. 289–301, (1985)

[11] Krstev, C., *Programski sistemi za obradu teksta*, [Text Processing Systems], Master Thesis, Faculty of Mathematics, Beograd, (1989)

[12] Désarménien, J., *French hyphenation by computer: application to TEX*, Technology and Science of Informatics, Gauthier-Villars & John Wiley & Sons, Vol. 6, No. 1. (1987)

[13] Kolodziejska, H., *Dzielenie wyrazów polskich w systemie TEX*, [Hyphenation of Polish words by TEX], Report No. 165, Institut of Informatics, Warsaw University, (1987)

[14] Liang, F. M., *Word Hy-phen-a-tion by Comput-er*, Ph. D. Thesis, Department of Computer Science, Stanford University, Report No. STAN-CS-83-977, (1983)

[15] Knuth, D. E.: *The TEXbook*, Addison-Wesley, Reading, Mass., (1984)

◇ Cvetana Krstev
Computer Laboratory
Faculty of Sciences
Studentski trg 16
11000 Belgrade
Yugoslavia
Bitnet: xpmf101@yubgss21