Refereed article

Cultural impacts on electronic publishing: experience in Serbia

Duško Vitas and Cvetana Krstev

The authors

Duško Vitas is Assistant Professor at the Computer Science Department of the Faculty of Mathematics, and **Cvetana Krstev** is Assistant Professor at the Library and Information Science Department of the Philological Faculty, both at the University of Belgrade, Yugoslavia.

Keywords

Electronic publishing, Language, National cultures, Serbia

Abstract

Discusses the linguistic influences on an electronic publishing infrastructure in an environment with unstable linguistic standardization from the computational point of view. Essentially, in Serbia in the last half of the century (at least) publishing is based on the following facts: two alphabetic systems are regularly in use with the possibility to mix both alphabets in the same document; the various dialects are accepted as a part of a linguistic norm; orthography is unstable – presently, several linguistic attitudes that have different views of the orthographic norm are under discussion; and, in Serbia, many minority languages are in use, which makes it difficult to provide efficient contact between different communities through electronic publishing. In this context, a systematic solution that responds to this complex situation has not been developed in the frame of traditional Serbian linguistics and lexicography in a way that enables the adequate incorporation of the new publishing technologies. Owing to these constraints, the direct application of electronic publishing tools frequently causes the degradation of the linguistic message. In such an environment, the promotion of electronic publishing therefore needs specific solutions. The paper discusses the general frame based on the specifically encoded system of electronic dictionaries that makes electronic texts independent of some of the mentioned constraints. The objective of such a frame is to enable the linguistic normalization of texts at the level of their internal representation, and to establish bridges for communicating with other language societies. Some aspects of electronic text representation that ensures its correct interpretation in different graphical systems and in different dialects are described. This also allows text indexing and retrieval using the same techniques that are available for languages not burdened with these problems.

New Library World
Volume 100 · Number 1149 · 1999 · pp. 171–178
 · ISSN 0307-4803

Introduction

For the last two decades in Serbia, as well as in former Yugoslavia, all necessary devices for electronic publishing on the technological level have been present. This technology was imported to fulfil the real needs of publishing houses, but the equipment itself proved not to be sufficient to develop the wider environment in which it can produce its best effects. The frequently encountered examples, such as retyping of the same text several times during its production, destruction of electronic texts on a publisher's sites, the lack of a national standard group corresponding to ISO/IEC JTC1 SC18, and many other factors show that the import of technology is not sufficient to prevent technological underdevelopment. Because of these deficiencies, the paradoxical situation arises that, although the technological base for electronic publishing is well developed, it is often not used to improve an efficient information flow.

Although in the recently adopted strategy for development of information technology in Yugoslavia (July 1997) shows that the importance of electronic publishing is recognized, unfortunately, the projects that are proposed continue the former practice. For instance, this document recommends that the products of electronic publishing are put on the Internet, which is in conformity with global trends, but nothing is said about the necessary infrastructural prerequisites to achieve this goal.

Inspired by these problems, the research group for text processing at the Faculty of Mathematics investigated tools that would, at least at the linguistic level, enable efficient information processing and communication.

Text as a natural language object

At least one part of most documents is comprised of text in some natural language. This part of a document, either written by hand or in electronic form, is rather the representation of information than the information itself (Birnbaum, 1995). Text in electronic form is represented by a sequence of bytes that can be interpreted in some way. During the last decade, great effort has been made to formally describe the structure of text, namely its logical and graphical layout, as well as to develop comprehensive character

New Library World
Volume 100 · Number 1149 · 1999 · 171–178

codes. The description of these formal aspects of text structure contributes to its better understanding by a human reader although on the level of its internal representation text itself does not contain the information that enables this understanding. (While looking at the visual representation of text one has the impression that text really contains the information that can be read from it.) The understanding of text stems from understanding the language in which it is written (Schwartze, 1985). The portion of a document which comprises natural language text is organized primarily by natural language and interpreted by its linguistic features and not by the graphical or logical layout or some other non-linguistic characteristics of the text.

Even on the level of international standards from the field of information technology (e.g. ISO/IEC, group JTC1/SC18), electronic text is not seen as an object organized by the rules of some natural language. The lack of this kind of linguistic information can lead, on one side, to the degradation and corruption of text to the point of inability to reconstruct the encoded information it has to convey and, on another side, it disables every automatic transformation of text based on this linguistic knowledge.

In the case of languages for which linguistic standardization is achieved these facts can be hidden to some extent. However, in case of a language system such as Serbo-Croatian, the lack of this information can cause serious problems in every step of its processing.

Serbo-Croatian

We use the term Serbo-Croatian to cover a linguistic system in a sense described in Popović (1996): Serbo-Croatian is used as an accepted name for one linguistic base from which several different language standards were derived: Serbian, Croatian, Bosnian. We will confine ourselves in this article to the use of the language in the territory of Serbia as primarily relevant to our work.

The source of this diffuse situation can be found in an orthographic reform dating from the middle of the nineteenth century that introduced a phonetically based orthography. However, cultural and historical conditions did not enable the support of this reform by appropriate language standardization. The consequences were twofold: on a cultural level, this reform produced a rough separation from the former cultural heritage (Selimović, 1987) and, on the linguistic level, it enabled the reproduction of many pronunciations in a written message. The latter phenomenon created a situation in which variations in contemporary text resemble the problems encountered today by researchers of old Italian, old French, etc. The same phenomena are present in other contemporary languages with stable standardization but in these cases they are a result of occasional graphical variations (Gross, 1989a) rather than a systemic feature of the orthographic system.

As a result of close connections with different cultures in recent history, two alphabets are in use in Serbia: Latin and Cyrillic. Although Cyrillic is recommended as the official alphabet, in a large number of documents the Latin alphabet is used for various reasons, sometimes political but also practical (such as a lack of appropriate Cyrillic fonts etc.). Consequently, recent attempts, for instance in the frame of ISO TC46, to define only the part of Serbo-Croatian that uses the Cyrillic alphabet as Serbian were not justifiable. The corpus of daily newspapers published in Serbia shows that in some of them the Cyrillic alphabet prevails while in the others the Latin alphabet prevails. An illustrative example is the regular bulletin of the Yugoslav Standards Organization in which the Latin and Cyrillic alphabet are mixed even in its title: *ЈUSинформације*. It means that, for the purpose of automatic processing, dialects and subdialects used in particular text as well as the alphabet in which they are written are its essential attributes that should be explicitly encoded.

The above two phenomena – reproduction of pronunciation in a written text and the use of two alphabets – have to be taken into consideration before processing a text. The lack of appropriate support that would address these requirements diminishes the possibilities of electronic publishing. For instance, some of the newspapers and journals published in Serbia often use for their Internet presentations the reduced Latin alphabet which consists of 22, instead of 30, letters: that is the English alphabet without w, x, y, z. Diacritics are omitted and some graphemes are substituted by digraphs[1].

New Library World
Volume 100 · Number 1149 · 1999 · 171–178

Diacritics are, however, distinctive in Serbian as is shown by the case of the following word forms:

reci, dative singular of *reka* (Eng. river) or imperative second person singular of *reći* (Engl. to say);
reći, infinitive (Eng. to say);
reči, nominative plural of *reč* (Eng. word).

All of these forms are reduced to only one, *reci*, if this reduced Latin alphabet is used.

The problems occurring on this lowest level of text representation are nevertheless complex enough that they cannot be solved by simple string matching methods: even the transliteration between the Cyrillic and Latin alphabet is not unique, and pronunciation variants are standardized neither on the orthographic nor on the morphological level, etc. For instance, if text is transformed from Latin to Cyrillic alphabet only by changing the font and actually preserving all the codes, *saopštenje* (Eng. communication) becomes *саопштење* instead of *саопштење*, and *саопштење* becomes *saopštenje* instead of *saopštenje* if transformation is done the other way round. But, both *саопштење* and *saopštenje* are misspelled in Serbo-Croatian. Moreover, digraphs can be ambiguous as is shown by the case of the noun *konjugacija* (Eng. conjunction) where the group *nj* remains in Cyrillic: *конјугација*.

On the other side, the efforts of traditional lexicography are mainly concentrated on the production of a Serbo-Croatian dictionary of literary language and vernacular through a project of the Serbian Academy of Science and Art which is due to be finished and published in paper form by the year 2050. The solution for linguistic problems that arise in the field of electronic publishing is therefore often found in *ad hoc* orthographic guidelines. As a consequence of the separation of the Serbian and Croatian languages, the key interest of most orthographers during the last few years in Serbia was, and still is, to stress the differences between these two languages. In spite of many differences between them, most orthographers agree on two points:

(1) The Serbian language consists of two pronunciations: *ekavian* and *jekavian*. (*Jekavian* pronunciation is also the base for Croatian). Both pronunciations are reproduced in written text. For instance, both the *ekavian* form *reč* and *jekavian* form *rijereč* (Eng. word) occur in written form.
(2) The Serbian language uses both the Cyrillic and Latin alphabet.

However, there is not full agreement about other language phenomena. As a result of this unstable situation, one can find, even in the documents of the recently established official Council for Language, the following records:

preds(j)ednik (Eng. president), encompassing both *predsednik* (*ek.*) and *predsjednik*, (*jek.*)
r(ij)eč (Eng. word), encompassing both *reč*, (*ek.*) and *riječ*, (*jek.*)

From the formal point of view, this solution introduces parenthesis in the alphabet and substantially complicates the recognition of formal words as a sequence of characters between separators.

The factor that remains neglected in linguistic discussions is the multilingual situation in Serbia. Namely, there are several minority groups in Serbia, most of them located in the north, in the region of Vojvodina. There are approximately 18 minority languages, of which Hungarian is the most important. In some of these languages there is a rather vivid publishing activity in Serbia. Automatic linguistic support for information exchange between these different language systems is practically non-existent. Also, the influence of other languages, such as English, French, etc., as a result of the state of lexicographic and linguistic theory, and particularly due to unstable terminology, generates additional difficulties in the transfer of information and knowledge.

Electronic dictionary

One possible solution to the mentioned problems is found in the theoretical frame of the lexicon-grammar (Gross, 1975) which gives one answer to the problem of distribution of information between grammar and dictionary[2]. One of the consequences of this approach is the precise and systemic encoding of the grammatical information for every lexical entry of the lexicon (Gross, 1989). The specific form of electronic dictionary (which is intended for text processing only and not for human use) developed in the framework of this theoretical model extensively describes the features of lexical units. It is therefore possible to assign to the sequences of characters from the text the lexical unit with the corresponding grammatical information. The basic unit of

New Library World
Volume 100 · Number 1149 · 1999 · 171–178

this model is a simple word defined as a sequence of characters between two separators. The components of the system are:

- a dictionary of simple lexical units (DELAS) with an accompanying dictionary of the corresponding inflected forms (DELAF);
- a dictionary of compounds (DELAC) with an accompanying dictionary of corresponding inflected forms (DELACF);
- a system of local grammars that describes the wider text fragments in the form of finite transducers.

The dictionaries of compounds and the local grammars are used as a device for the elimination of ambiguity (Roche, 1997). These components, or some part of them, are developed for several European languages, namely for French, English, German, Italian, Spanish, Portuguese, Polish, Greek, Bulgarian, and Serbo-Croatian.

For instance, the DELAF dictionary of simple words for English has the following form:

- abbreviating,abbreviate.V4:ing
- abbrevated,abbreviate.V4:Pp:Pret
- abbreviates,abbreviate.V4:Pr3s
- abbreviation,.N1:Ns
- abbreviations,abbreviation.N1:Np
- abbreviator,.N1:Ns
- abbreviators,abbreviator.N1:Np

where, for instance, abbreviating represents the textual word, abbreviate represents the lexical word and V4:ing represents the corresponding grammatical code.

Compounds are defined (Silberztein, 1993) as sequences that include several simple words. However, compounds, that are sometimes called frozen expressions, have to be distinguished from any free sequence of simple words: the fact that makes them different is that the syntactic property of a compound usually cannot be deduced from the syntactic properties of its constituent simple words. Examples of such expressions in English belonging to different parts of speech are: morning glory, make-believe, a piece of cake. More often than not, the compounds, are written without the characteristic separation sign. On the level of compounds ineffective restrictions in the sintagmatic constructions are also precisely and extensively described.

The system of electronic dictionaries, including the local grammars integrated in the system INTEX, enables the transformation of texts that are based on natural language organization, such as automatic lemmatization which associates to every textual word the corresponding lexical word (for instance, lexical word abbreviation would be associated with the textual word abbreviation). This system is rather a resource for different applications than an application itself. For instance, a spelling checker can be obtained as an excerpt from the electronic dictionary.

Taking this methodological base and format as a starting point, a prototype system of morphological electronic dictionaries is developed for simple words and compounds in Serbo-Croatian (Vitas, 1993; Krstev, 1997; Nenadić, 1997). In Table I the short excerpts are given from the constructed dictionaries[3].

Taking into account the state of affairs in traditional lexicography and problems outlined earlier the construction of the system of electronic dictionaries of Serbo-Croatian has to take care of the following:

- It must be independent of the alphabet, that is, for instance the word *saopštenje* (Eng. communication) has to have one entry in the electronic dictionary which is independent of its coding in text.
- The dictionaries have to synthesize different pronunciations and dialect variations by reducing them to some canonical form. This means that, for instance, words *reč*, (*ek.*) and *riječ*, (*jek.*) (Eng. word) have to be connected appropriately.
- The dictionaries have to neutralize the orthographic variations. For instance, due to the phonologically-based orthography, variations *hleb* and *leb* as well as *hljeb* and *ljeb* (Eng. bread) are possible. Also, *dan-i-noć*, *dan i noć* and *daninoć* (Eng. pansy) are orthographically all correct and should be covered by a dictionary of compounds in the latter case.

The research has shown that variations due to different pronunciations or orthographic origin do not influence the morphological behaviour of lexical units. For instance, in the mentioned examples *reč*, ek./*riječ*, jek. (Eng. word) and *hleb*, ek./*leb*, ek./*hljeb*, jek./*ljeb*, jek. (Eng. bread) all the variations of the same lexical unit have, on the morphological level,

New Library World
Volume 100 · Number 1149 · 1999 · 171–178

Table I Short excerpts from the constructed dictionaries

DELAS	**DELAF**
dno,N51.01-*,Pre*	do,.Pre*,.Adv*,.N15.08-*:msn-:msa-
do,N15.08-*,Pre*,Adv*	doba,.N64.01-*:nsn-:nsg-:nsa-:nsv-:npn-:npg-:npa-:npv-
doba,N64.01-*,N90.00-*	dobar,.A14.01*:p#msn*:p#msa-
dobar,A14.01*	dobara,dobro.N51.02-*:npg-
dobaviti,V33.51.3*	dobave,dobaviti.V33.51.3*:P3p
dobijati,V01.00.2*	dobave&cx;i,dobaviti.V33.51.3*:AdvPr
dobiti,V24.50.2*	dobavi,dobaviti.V33.51.3*:P3s:Y2s:A2s:A3s
dobivalac,N17.18+*	dobavih,dobaviti.V33.51.3*:A1s
dobivati,V01.00.2*	dobavila,dobaviti.V33.51.3*:PPsf:PPpn
dobo&sx;,N22.01-*	dobavile,dobaviti.V33.51.3*:PPpf
DELAC	**DELACF**
fiksna.A/ta~ka.N:N	disjunktne/skupove,disjunktni/skupovi.:Nma-
fundamentalan.A/niz.N:N	disjunktni/skupovi,.:Nmpn-:Nmpv-
funkcija.N@Inv/jedne/promen&lx;ive:N	disjunktnih/skupova, disjunktni/skupovi.:Nmpg-
funkcija.N/neprekidna.A@Inv/u/ta&cy;ki:N	disjunktnim/skupovima,
geometrija.N@Inv/Loba&cy;evskog:N	disjunktni/skupovi.:Nmpd-;Nmpi-:Nmpl-
grani&cy;na.A/vrednost.N:N	diskretan/prostor,.:Nmsa-:Nmsn

the same inflective and derivational features. These variations influence only the root morpheme.

An extensive description of all these variations in the electronic dictionary would unnecessarily multiply its size. Besides that, the dialect variants would be consistently reproduced at the level of the lexical unit by representing the same lexical unit with several different lexical entries.

One solution may be found in the concept of the lexicographeme as a means of dictionary normalization (Krstev, 1997). The concept will be illustrated with one example. For the lexical unit *mesec* (Eng. moon) several variant forms exist according to different pronunciations and dialects: *mesec*, ek./*mjesec*, jek./*misec*, ik./*mljesec*, dial.jek. All these forms are recorded in the dictionary (SANU, 1959). The normalized form of this lexical unit could be *m#esec*, where *#e* represents the lexicographeme having the following properties: in written text it can be realized as one of the following sequences: – *e, je, i* – depending on the pronunciation. Furthermore, it can affect the preceding grapheme – that is, palatalize the preceding consonant – in the case of certain dialects.

This concept leads to the development of a system of meta-dictionaries from which the particular system of dictionaries can be realized which correspond to a certain dialect or orthographic practice. On the level of text processing, this system enables the different forms of text tuning – transformation from one alphabet to another – as well as the conversion from one pronunciation to another, etc.

Bilingual lexicography is often unable to define the precise translation equivalents for the given reasons (Krstev and Vitas, 1998). Thus, the bilingual lexicography, as well as comparative language studies, are burdened with the same problems that aggravate the processing of Serbo-Croatian from which different defects in multilingual communications can arise. In this way, the normalization of electronic dictionaries can contribute to some extent to the improvements of bilingual lexicography.

These concepts will be illustrated with a few examples that show the improvements of electronic publishing techniques by underlying text with the electronic dictionary.

Electronic edition of Vuk's Serbian proverbs

This collection, comprising about 7,000 proverbs, has been assembled by the language reformer Vuk Stefanović Karadžić. Its first edition dates from the year 1849. All the later editions reproduce this first edition in all aspects. The inventory of proverbs has not changed either, although there were references to nonexistent proverbs and a lot of identical proverbs differing only in word ordering are listed, etc. Nevertheless, this text

New Library World
Volume 100 · Number 1149 · 1999 · 171–178

is an essential part of any corpus of contemporary Serbian and Croatian.

In 1987 the distinguished Belgrade publisher NOLIT started the project of re-editing this collection of proverbs that ended in 1996 with the publication of the new paper edition. Besides the removal of some deficiencies of the old edition, the new edition is distinguished by the comprehensive index that contains all the lexical words – nouns, verbs, adjectives and numbers – that occur in proverbs. The presence of numerous variations has, however, encumbered the index significantly: the index, formatted in small script, represents a third of the whole book.

At the same time the preparation of the electronic edition has started at the Faculty of Mathematics, purely as a scientific, non-commercial project. The electronic edition is based on the encoding scheme proposed by the Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard, 1995). Besides that, for the electronic edition, the text of proverbs has been underlined by electronic dictionaries. For instance, the proverbs *Ja kad viđeh zelen drijen predadoh mu vas moj drijem i lijen* (Eng. When I saw the green dogwood I gave him over my slumber and my laziness) and *Ja mu kažem adumac sam a on pita koliko đece imam* (Eng. I say to him I am a eunuch and he asks how many children I have) are represented in the following way in the electronic edition:

```
<divp id=P1770 n=1769>
  <w a=',.ProN01:*sn**'>Ja</w>
  <w a=',.Adv*,.Con*'>kad</w>
  <w a=',vi|eti.V37.25.4J%vide-
    ti#E4.38:A1s'>vi&dx;eh</w>
  <w a=',.A08.01*:p#msn*:p#msa-'>ze-
    len</w>
  <w a=',.N08.01-J%dren#E3.07.01:msn-
    :msa-'>drijen</w>
  <w a=',predati.V06.50.2:A1s'>predadoh</
    w>
  <w a=',ono.ProN06:nsd*,on.
    ProN05:msd*'>mu</w>
  <w a=',vi.ProN04:*pg*:*pa*'>vas</w>
  <w a=',.ProA06:msn*:msa-:mpn*'>moj</
    w>
  <w a=',.N08.01-J%drem#E3.03:msn-
    :msa-'>drijem</w>
  <w a=',.Con*,.Par*'>i</w>
  <w a=',.A06.51J%le-
    n#E2.03:p#msn*:p#msa
    -'>lijen</w>
</pv>
<divp id=P1781 n=1780>
</pv> <w a=',.ProN01:*sn**'>Ja</w>
<opt.ph rend='bold'> <w a=',ono.-
ProN06:nsd*,on.ProN05:msd*'>mu</
w>
</opt.ph> <w a=',kaza-
ti.V21.05.4*:P1s'>ka&zx;em</w>
  <w a=',.N17.61+*%hadumac#H1.07:ms-
    n+'>adumac</w>
  <w a=',.ProA02:msn*:msa-,
    jesam.V99.00*:P1s'>sam</w>
  <w a=',.Con*,.Adv*,.Int*'>a</w>
  <w a=',.ProN05:msn*'>on</w>
  <w a=',.N70.01-*:fsn-:fpg-,pita-
    ti.V01.00.2*:P3s:A2s:A3s'>pita</w>
  <w a=',.Adv*,kolik.ProA02:nsn*:nsa*:'>
    koliko< /w>
  <w a=',|eca.N70.61+J%deca#E2.39'
    >ece</w>
  <w a=',imati.V04.00.2*:P1s'>imam</w>
</pv>
</divp>
```

In this electronic form every textual word in a proverb has been tagged with a SGML tag <w> whose attribute describes its possible lexical words, the particularities of its pronunciation and possible grammatical information. For instance, to textual word *pita* two lexical words can be associated: *pita* (Eng. pie) and *pitati* (Eng. to ask). The textual word *đece* is associated with the lexical word *đeca* from eastern *jekavian* pronunciation whose *ekavian* variant is *deca* (Eng. children). This form of electronic texts enables many text transformations, such as automatic lemmatization or indexing. Text can also be transformed so that it uses one chosen pronunciation or orthographic norm, as is shown by the transformation of the same two proverbs into the *ekavian* pronunciation, which has been done automatically using the information in the *a*-attribute and the appropriate information from the underlying dictionaries:

```
<!– East jekavian pronunciation (dialect)
–>
1770 Ja kad vi&dx;eh zelen drijen, pre-
dadoh mu vas moj drijem i lijen.
<!– Ekavian pronunciation –>
1770 Ja kad videh zelen dren, predadoh
mu vas moj drem i len.
<!– East jekavian pronunciation (dialect)
–>
1781 Ja mu ka&zx;em adumac sam, a on
pita koliko edx;ece imam.
<!– Ekavian pronunciation; contempor-
ary orthography concerning the use of
h –>
```

New Library World
Volume 100 · Number 1149 · 1999 · 171–178

1781 Ja mu ka&zx;em hadumac sam, a on pita koliko dece imam.

Other applications

In addition to the hypertextual connections deduced from the graphical and logical layout of text, it seems natural to enable the running through the text according to the lexical units. If the approximation of such a navigation in English can be accomplished with a kind of find (or search) command, in Slavic languages that are characterized by their reach morphological and derivational system, a special kind of support has to be developed. In order to investigate this possibility an experiment was done using a sample of mathematical textbooks (Nenadić, 1997). In contrast with the previous example, texts were parsed by the dictionaries of compounds and appropriate local grammars. The investigation is limited to the noun phrases. For instance, the phrase *niz intervala* (Eng. sequence of intervals) at the level of simple words can be recognized as a sequence:
(1) Pre Ns(g);
(2) Pre Np(g);
(3) Ns(na) Ns(g); and
(4) Ns(na) Np(g).

In this sequence Pre denotes preposition, Ns noun in a singular form and Np noun in a plural form. After underlying the text with the electronic dictionary and applying constraints of agreement, the first two possibilities are rejected as the preposition *niz* (Eng. down) requires the noun in accusative.

This first attempt proved that such a text indexing is fruitful, especially in respect to eliminating ambiguity. The goal of further research will be to establish an appropriate semantic network.

In parallel to the electronic dictionary construction the production of the corpus of parallel texts is on its way. It is done in cooperation with the TELRI (Trans-European Language Resources Infrastructure) project which is funded by the European Commission. TELRI has launched the work on production of a corpus of parallel texts for European languages. As a result, electronic versions of Plato's *Republic* for more than 20 languages, including Serbo-Croatian, has been produced, fully SGML encoded according to the TEI guidelines. Alignment with English, French and German has been produced. In cooperation with the MULTEXT-East project, an electronic version of Orwell's *1984* for ten languages, including Serbo-Croatian has been produced, fully SGML encoded according to CES1 guidelines[4]. Alignment has been done for all the language pairs.

As an example, the passage from Orwell's *1984*: "'I was passing,' said Winston vaguely. 'I just looked in. I don't want anything in particular'", which is translated in Serbo-Croatian as "'*Samo sam prolazio', neodređeno reče Vinston, 'pa sam pogledao. Nisam tražio ništa posebno*'", is presented in the following way in the parallel corpus which has been aligned to the level of the sentence[5].

*** Link: 1-2 ***
<<Oshs.1.9.63>> <Oshs.1.9.63.1>
"Samo sam prolazio", neodre&dx;eno re&cy;e Vinston, "pa sam pogledao."
.EOS
<<Oen.1.8.62>> <Oen.1.8.62.1> "I was passing," said Winston vaguely.
<Oen.1.8.62.2> "I just looked in." .EOS
*** Link: 1-1 ***
<Oshs.1.9.63.2> "Nisam tra&zx;io ni&sx;ta naro&cy;ito." .EOS
<Oen.1.8.62.3> "I don't want anything in particular." .EOS
.EOP

On the basis of the experiences obtained during the work on these two projects the production of a parallel corpus has started in which one language will be Serbo-Croatian with the intention of aligning it with as many languages as possible, especially with languages with direct contact with Serbo-Croatian. It is expected that underlying such a parallel corpus with electronic dictionaries will enable, to a certain extent, the automatic establishment of translation equivalents.

Conclusion

In this article the problems encountered in one unstable linguistic system were illustrated. In the scope of traditional publishing these problems were disguised due to human understanding of language in which text is typeset. Promoting electronic publishing requires the explicit representation of at least a

New Library World
Volume 100 · Number 1149 · 1999 · 171–178

part of this knowledge through support for the processing of linguistic data.

Notes

1 (For instance, the daily newspaper *Danas* is published in the Latin alphabet and it uses degraded Latin for its WWW presentation (http://www.danas.co.yu/). The daily newspaper *Večernje Novosti* is published in Cyrillic but also uses degraded Latin for its WWW presentation (http://www.vnovosti.co.yu/). On the other side, weekly newspapers *Vreme* (paper version in Latin) and *Ilustrovana Politika* (paper version in Cyrillic) both use the Latin alphabet coded in Windows 1250 code page for their WWW presentations (http://www.danas.co.yu and http://www.politika.co.yu/ilustro/, respectively). The radio station B92 presents some of its news in the Latin alphabet coded in ISO 8859-2 and some in the degraded Latin alphabet (http://www.opennet.org/).

2 An overview of this approach can be found at http://www.ladl.jussieu.fr/.

3 In all Serbo-Croatian examples diacritics and digraphs will be represented by following SGML entities: &cy; (č), &cx; (ć), &sx; (š), &zx; (ž), &dx; (đ), &lx; (lj), &nx; (nj), and &dy; (dž). It does not necessarily mean that the same representation was used in a real application.

4 This project is described at http://www.cs.vassar.edu/CES/CES1.html

5 Both the corpora and supporting software have been produced on CD-ROM whose description can be found at http://www.telri.de/.

References and further reading

Birnbaum, D.J. (1995), "Informational and presentational units in early Cyrillic writing", *Proceedings of the First International Conference on Computer Processing of Medieval Slavic Manuscripts*, Blagoevgrad, 24-28 July, pp. 41-9.

Gross, M. (1975), *Méthodes en Syntaxe*, Hermann, Paris.

Gross, M. (1989), "La construction de dictionnaires électroniques", *Annales des Télécommunications*, Vol. 44 Nos 1-2, pp. 4-19.

Gross, M. (1989a), "The use of finite automata in the lexical representation of natural languages", in Gross, M. and Perrin, D. (Eds), *Electronic Dictionaries and Automata in Computational Linguistics*, Lecture Notes in Computer Science, No. 377, Springer-Verlag, Berlin, pp. 34-50.

Krstev, C. (1997), "One approach to text modeling and transformation", (in Serbo-Croatian), PhD thesis, Faculty of Mathematics, University of Belgrade.

Krstev, C. and Vitas, D. (1998), "Morphological normalization of translation equivalents", *Third European TELRI Seminar: "Translation Equivalence – Theory and Practice"*, Montecatini Terme, 16-18 October 1997, Institut für deutsche Sprache, Mannheim and Tuscany Word Center, Montecatini Terme, pp. 117-23.

Krstev, C., Pavlović-Lažetić, G. and Vitas, D. (1997), "Neutralization of variations in a dictionary entry's structure in Serbo-Croatian", in Junghanns, U. and Zybatow, G. (Eds), *Formal Slavistik*, Vervuert Verlag, Frankfurt am Main, pp. 417-25.

Nenadić, G. (1997), "Algorithms for recognition of compounds in mathematical text and its applications", (in Serbo-Croatian), Masters Thesis, Faculty of Mathematics, University of Belgrade.

Popović, L.J. (1996), "Deux approches idéologiques de la vernacularisation de la langue littéraire chez les Serbs à la fin du 18e et dans la première moitié du 19e siècle", *Langues et Nation en Europe Centrale et Orientale du 19e Siècle à Nos Jours*, Cahiers de l'ILSL, Lausanne, No. 8, pp. 209-40.

Roche, E. and Schabes, Y. (Ed.) (1997), *Finite-state Language Processing, A Bradford Book*, The MIT Press, Cambridge, MA and London.

SANU (1959-1990), *Rečnik Srpskohrvatskog Književnog i Narodnog Jezika, Vol. 1-14 (A-N)*, Srpska akademija nauka i umetnosti i Institut za srpskohrvatski jezik, Beograd.

Schwartze, C. (1985), "Text understanding and lexical knowledge", lecture at the International Pragmatics Conference, Viareggio.

Selimović, M. (1987), *Za i protiv Vuka*, BIGZ, Beograd.

Silberztein, M. (1993), *Dictionnaires Electroniques et Analyse Automatique de Textes: Le System INTEX*, Masson, Paris.

Sperberg-McQueen, C.M. and Burnard, L. (Eds) (1995), *Guidelines for Electronic Text Encoding and Interchange*, Text Encoding Initiative, Chicago-Oxford.

Vitas, D. (1993), "Mathematical model of Serbo-Croatian morphology (nominal inflection)", (in Serbo-Croatian), PhD thesis, Faculty of Mathematics, University of Belgrade.

Vitas, D. and Krstev, C. (1996), "Tuning the text with electronic dictionary", *Papers in Computational Lexicography*, COMPLEX'96, Budapest, pp. 267-76.