Cultural Impacts on Electronic Publishing: Experience in Serbia*

Duško VITAS, Cvetana KRSTEV

Abstract

The paper discusses the linguistic influences on an Electronic Publishing infrastructure in an environment with a low-level of linguistic standardization. Essentially, in Serbia in the last half of the century (at least) publishing has been based on the following facts:

1. Two alphabetic systems are regularly in use with the possibility to mix both alphabets in the same document;
2. The various dialects are accepted as a part of a linguistic norm;
3. Orthography is unstable—presently, several linguistic attitudes that have different views of the orthographic norm are under discussion;
4. 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. Due 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. In this paper we discuss a general frame based on a specifically encoded system of electronic dictionaries that makes the electronic texts independent from some of the mentioned constraints. The objective of such a frame is: (1) to enable the linguistic normalization of texts on the level of their internal representation, and (2) to establish bridges for the communication with other language societies. We underline some aspects of electronic text representation that ensures its correct interpretation in different graphical systems and in different dialects. This also allows text indexing and retrieval using the same techniques that are available for languages not burdened with these problems.

1. 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 fulfill the real needs of publishers, but the equipment itself proved not to be

* **Editors' notes:** The original title was "Cultural Impacts to Electronic Publishing: Experience in Serbia". In addition to conventional editing (including word substitution) some extra text has been included where the Editors felt it clarifies the Author's intended meaning. Any such added text is indicated by being surrounded by square brackets [...]. The Editors would like to acknowledge the help of Alex Robinson (University of Kent at Canterbury) in the editing of this paper.

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 others show that the importance of a technology is not sufficient to prevent its 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 efficient information flow.

Even in the recently adopted strategy for development of information technology in Yugoslavia (July 1997) the importance of electronic publishing is recognized but, unfortunately, the projects that are proposed continue with 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 investigates tools that would, at least at the linguistic level, enable efficient information processing [and communication].

2. 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, a great effort has been made to formally describe the structure of text, namely its logical and graphical layout. 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 it [ie, this description]. (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 the linguistic standardization is achieved these facts can be hidden. However, in a case of a language system such as Serbo-Croatian, the lack of this information can cause serious problems in every step of its processing.

3. 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 are defined: Serbian, Croatian, Bosnian, maybe also Montenegrian. We will confine ourselves in this article to the use of the language on the territory of Serbia as primarily relevant to our work.

The source of this diffuse situation can be found in a orthographic reform dating from the middle of the XIXc [nineteenth century] that introduced a phonetically based orthography. However, the cultural and historical conditions did not enable the support of this reform by an appropriate language standardization. The consequences were twofold: on a cultural level, this reform produced the rough separation from the former cultural heritage, and on the linguistic level, it enabled the reproduction of many pronunciations in a written message. The latter phenomenon yielded the situation in which the variations in a contemporary text resemble the problems encountered today by the 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 as a systemic feature of the language.

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, the recent attempts, for instance in the frame of ISO TC46, to define only the part of Serbo-Croatian that uses Cyrillic alphabet as Serbian were not justifiable. The corpus of daily newspapers published in Serbia shows that in some of them Cyrillic alphabet prevails, in the others Latin alphabet prevails, with the exceptional example of "Naša Borba" in which the both alphabets are equally in use: if one page is typeset using one alphabet then the opposite page uses the other. In this way, for purpose of automatic processing, dialects and subdialects used in particular text as well as the alphabet in which it is encoded are the essential attributes of text that should be explicitly encoded.

The above two phenomena—reproduction of dialects in a written text and the use of two alphabets—are two propositions [requirements] that have to be taken into consideration before processing some text. The lack of appropriate support that would answer to these propositions [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, as is shown by the examples of the daily newspaper "Naša Borba" whose paper version regularly uses two alphabets and the quarterly of the Serbian Association of Literary Translators "Mostovi" which uses the Cyrillic alphabet for its paper version. Diacritics are, however, distinctive as is shown by the case of the following word forms:

reci, dative singular of *reka* (engl. *river*)
reci, imperative second person singular of *recí* (engl. *to say*),
recći, infinitive (engl. *to say*),
rečći, nominative singular of *rečˇ* (engl. *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 the simple string matching methods: even the transliteration between Cyrillic and Latin alphabet is not unique [not one-to-one], and pronunciation variants are standardized neither on orthographic nor on 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* (engl. *communication*) becomes *саопштење* instead of *саопштење*, and *саопштење* becomes *saopštewe* instead of *saopštenje* if transformation is done the other way round. Moreover, digraphs can be ambiguous as is shown by the case of the noun *konjugacija* (engl. *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 the Serbo-Croatian dictionary of literary language and vernacular through the 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 the linguistic problems that arise in the field of electronic publishing is therefore often found in the *ad hoc* orthographic guidelines. As a consequence of the separation of Serbian and Croatian language standard, the key interest of most of the orthographers during the last few years in Serbia was, and still is, to stress the differences between these two language norms. In spite of many differences, most orthographers agree on two points:

- The Serbian language consists of two pronunciations: ekavian and jekavian. (Jekavian pronunciation is also the base for Croatian language norm. No language norm has the Ikavian pronunciation as its base.);
- The Serbian language uses both Cyrillic and Latin alphabets.

However, there is not a full agreement about the 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 (engl. *president*), encompassing both *predsednik, ek.* and *predsjednik, jek.*
r(ij)eč (engl. *word*), encompassing both *reč, ek.* and *riječ, jek.*

From the formal point of view, this solution introduces parenthesis into the alphabet and substantially complicates the recognition of formal words as sequences 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. The automatic linguistic support for the 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 the lexicographic and linguistic theory, and particularly due to the unstable terminology, generates additional difficulties in the transfer of information and knowledge.

4. Electronic dictionary

One possible solution to the mentioned problems is found in the theoretical framework of the lexicon-grammar (Gross 1975) which gives one answer to the problem of distribution of

information between grammar and dictionary. The basis 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 this model is a *simple word* defined as a sequence of characters between two separators. The components of the system are: (1) a dictionary of simple lexical units (DELAS) with an accompanying dictionary of the corresponding inflected forms (DELAF); (2) a dictionary of compounds (DELAC) with an accompanying dictionary of corresponding inflected forms (DELACF); (3) the 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 disambiguation (Roche 1997). These components, or some part of them, are developed for several European languages, namely for French, English, German, Italian, Spanish, Portuguese, Polish, 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

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 can not be deduced from the syntactic properties of its constituent simple words. The 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 syntagmatic constructions are also precisely and extensively described.

The system of electronic dictionaries including the local grammars integrated in the system INTEX enables the transformations of texts that are based on its natural language organization, e.g. automatic lemmatization or document indexing with disambiguation. This system is a resource for different applications rather 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 the following example the short excerpts are given from the constructed dictionaries[1].

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/tačka.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

Taking into account the state of affairs in traditional lexicography and problems outlined in section 3, the construction of the system of electronic dictionaries of Serbo-Croatian has to take care of the following:

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

[1] 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 real application.

The research has shown that the variations of dialect or orthographic origin do not influence the morphological behaviour of lexical units. For instance, in the mentioned examples *reč, ek.* and *riječ, jek* (engl. *word*) and *hleb* and *leb* (engl. *bread*) all the variations of the same lexical unit have on the morphological level the same inflective and derivational features. These variations influence only part of the root morpheme.

The extensive description of all these variations in an electronic dictionary would unnecessarily multiply its size. Besides that, the dialect variants would be consistently reproduced on 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 *lexicographeme* as a mean of dictionary normalization (Krstev 97). The concept will be illustrated on one example. For the lexical unit *mesec* (engl. *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 59). 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 phoneme—that is, palatalize the preceding consonant—in the case of certain dialects.

This concept leads to development of the 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 (Krstev 1998). Thus, bilingual lexicography, as well as the 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.

5. Electronic edition of Vuk's Serbian Proverbs

This collection, comprising about 7000 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, a number of identical proverbs differing only in for word ordering, etc. Nevertheless, this text 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 of this collection of proverbs that ended in 1996 with the publication of the new paper edition. Besides the removal of the 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.

At the same time the preparation of the new edition 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 Text Encoding Initiative TEI (http://www-tei.uic.edu/orgs/tei). Besides that, the text of proverbs has been underlined [parsed] electronically with the dictionaries which yielded the results illustrated by the following two proverbs:

<divp id=P1770 n=1769>
<!-- When I saw the green dogwood I gave him over my slumber and my laziness -->
<pv> <w a=‘,.ProN01:*sn**’>**Ja**</w>
<w a=‘,.Adv*,.Con*’>**kad**</w>
<w a=‘,viđeti.V37.25.4J%videti#E4.38:A1s’>**vi&dx;eh**</w>
<w a=‘,.A08.01*:p#msn*:p#msa-’>**zelen**</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%len#E2.03:p#msn*:p#msa-’>**lijen**</w>
</pv>

<divp id=P1781 n=1780>
<!-- I say him I am eunuch and he asks how many children I have -->
<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=‘,kazati.V21.05.4*:P1s’>**ka&zx;em**</w>
<w a=‘,.N17.61+*%hadumac#H1.07:msn+’>**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-,pitati.V01.00.2*:P3s:A2s:A3s’>**pita**</w>
<w a=‘,.Adv*,kolik.ProA02:nsn*:nsa*:’>**koliko**< /w>
<w a=‘,đeca.N70.61+J%deca#E2.39’ >**&dx;ece**</w>
<w a=‘,imati.V04.00.2*:P1s’>**imam**</w>
</pv>
</divp>

Every textual word in a proverb has been tagged with SGML tag **<w>** whose attribute **a** describes its possible lexical words, the particularities of their pronunciation and possible grammatical information. 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, predadoh 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 **&dx;ece** imam.
<!-- Ekavian pronunciation; contemporary orthography concerning the use of *h* -->
1781 Ja mu ka&zx;em **hadumac** sam, a on pita koliko **dece** imam.

The full description of this project can be found at http://alas.matf.bg.ac.yu/~cvetana/proverb.

6. Hypertext with natural language links

In addition to the hypertextual connections developed [which represent or] form the graphical and logical layout of text, it seems natural to enable 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 rich 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, underlying [parsing] of text was done by dictionaries of compounds and appropriate local grammars. The investigation is limited to the noun phrases. The following example shows the fragment of mathematical text indexed by dictionaries DELAC and DELACF which enables the establishment of links between all the different forms of the same lexical unit.

&CY;i&NX;enica:Nfpg-:Nfsn- da u **kompletan/metri&cy;ki/prostor:Nmsd-:Nmsl-**
svaki **Ko&sx;ijev/niz:Nmsa-:Nmsn-** konvergira od
su&sx;tinski.AP@fpa:AP@fpn:AP@fpv:AP@fsg:AP@mpa
je va&zx;nosti u u mnogim **primena/funkcionalne/analize:Nfpd-:Nfpi:Nfpl,**
matemati&cy;ki. AP@fpa:AP@fpn:AP@fpv:AP@fsg:AP@mpa
disciplini.Nfpa-:Nfpn-:Nfpv:Nfsg u kojoj se, pored ostalog,
prou&cy;avaju **metri&cy;ki/prostor:Nmpn-:Nmpv-.**

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

7. Corpus of parallel texts

In the scope of the TELRI (Trans-European Language Resources Infrastructure) project, funded by the European Commission, work on the production of a corpus of parallel texts for European languages was launched. 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 MULTEXT-East project electronic version of Orwell's "1984"

for 10 languages including Serbo-Croatian has been produced, fully SGML encoded according to CES1 guidelines (http://www.cs.vassar.edu/CES/CES1.html). Alignment has been done for all the language pairs.

As an example, one excerpt from the Serbo-Croatian and English version of Orwell's "1984" aligned to the level of sentence and hand-validated is presented.

*** 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

*** Link: 2 - 1 ***
<<Oshs.1.9.64>> <Oshs.1.9.64.1> "Ba&sx; dobro", re&cy;e antikvar. <Oshs.1.9.64.2> "Ne verujem da bih vam mogao udovo&lx;iti." .EOS
<<Oen.1.8.63>> <Oen.1.8.63.1> " It's just as well," said the other, " because I don't suppose I could have satisfied you." .EOS

*** Link: 1 - 1 ***
<Oshs.1.9.64.3> On okrete svoj meki dlan navi&sx;e, pokajni&cy;kim pokretom. .EOS
<Oen.1.8.63.2> He made an apologetic gesture with his softpalmed hand. .EOS

Both the corpora and supporting software will be produced on CD-ROM whose description can be found at http://www.ids-mannheim.de/telri/cdrom.html.

On the basis of the experience 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.

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

9. References

Birnbaum, D.J. 1995 “Informational and Presentational Units in Early Cyrillic Writing”. In: *Proceedings of First International Conference Computer Processing of Medieval Slavic Manuscripts.* Blagoevgrad, Bulgaria, 24 – 28 July 1995. 41–49.

Gross, M. 1975. “Méthodes en syntaxe”. Hermann, Paris

Gross, M. 1989. “La construction de dictionnaires électroniques”. In: *Annales des télécommunications* 44(1-2), 4-19.

Gross, M. 1989a. “The use of Finite Automata in the Lexical Representation of Natural Languages”. In: Gross, M.; Perrin, D. (eds.), *Electronic Dictionaries and Automata in Computational Linguistics*, Lecture Notes in Computer Science, no. 377, Springer-Verlag, Berlin, 34–50.

Krstev, C., Pavlović-Lažetić, G., Vitas, D. 1997. “Neutralization of Variations in a Dictionary Entry’s Structure in Serbo-Croatian”. *Formal Slavistik.* Junghanns, U. and Zybatow, G. (eds): 417−425.Frankfurt am Main: Vervuert Verlag.

Krstev, C., Vitas, D. 1998. “Morphological Normalization of Translation Equivalents”. In *Third European TELRI Seminar “Translation Equivalence—Theory and Practice”, Montecatini Terme, Italy, 16-18 October*. Institut für deutsche Sprache, Mannheim and Tuscany Word Center, Montecatini Terme, 1998. (to appear).

Krstev, C. 1997. “*One approach to text modeling and transformation”* (in Serbo-Croatian). Ph.D. Thesis, Faculty of Mathematics, University of Belgrade.

Nenadić, G. 1997. “Algorithms for Recognition of Compounds in Mathematical Text and its Applications” (in Serbo-Croatian). Master Thesis. Faculty of Mathematics, University of Belgrade.

Popović, Lj. 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*, no. 8: 209–240. Lausanne.

Roche, E., Schabes, Y. (ed.) 1997. “Finite-State Language Processing”. A Bradford Book, The MIT Press, Cambridge, Massachusetts, London, England.

SANU 1959–1990. *Rečnik srpskohrvatskog književnog i narodnog jezika, vol. 1–14 (A–-N)*, Beograd: Srpska akademija nauka i umetnosti i Institut za srpskohrvatski jezik.

Schwartze, C. 1985. “Text Understanding and Lexical Knowledge”. *Lecture at the International Pragmatics Conference*. Viareggio.

Silberztein, M. 1993: *Dictionnaires électroniques et analyse automatique de textes: le system INTEX*. Paris: Masson

Vitas, D. 1993. “Mathematical Model of Serbo-Croatian Morphology (Nominal Inflection)” (in Serbo-Croatian). PhD thesis, Faculty of Mathematics, University of Belgrade.

Vitas, D., Krstev, C. 1996. „Tuning the Text with Electronic Dictionary”. In: *Papers in Computational Lexicography*, COMPLEX’96. Budapest, 267-276