R{j}ecnik.com: English—Serbo-Croatian Electronic Dictionary

Vlado KEŠELJ
Faculty of Computer Science
Dalhousie University, Halifax
Canada, vlado@cs.dal.ca

Tanja KEŠELJ
Korlex Software
Bedford NS, Canada
tanja@keselj.net

Larisa ZLATIĆ
Larisa Zlatic Language Services
Austin, Texas, USA
larisaz@serbiantranslator.com

Abstract

The features of R{j}ecnik.com dictionary, as one of the first on-line English-Serbo-Croatian dictionaries are presented. The dictionary has been on-line for the past five years and has been frequently visited by the Internet users. We evaluate and discuss the system based on the analysis of the collected data about site visits during this five-year period. The dictionary structure is inspired by the WordNet basic design. The dictionary's source knowledge base and the software system provide interfaces to producing an on-line dictionary, a printed-paper dictionary, and several electronic resources useful in Natural Language Processing.

1 Introduction

The dictionaries, monolingual, bilingual, or multilingual, are the standard way of collecting and presenting lexicographic knowledge about one or more languages. The electronic dictionaries (EDs) are not merely a straightforward extension of their printed counterparts, but they entail additional purely computational problems.

ED as marked-up text. An ED may be seen simply as a long, marked-up text. The important computational issues arise around the problem of efficient keyword search and appropriate presentation of the dictionary data. The search is performed in the context of a markup scheme, such as SGML or XML, and the query model has to provide expressibility for search queries within this scheme; e.g, searching for a keyword within a certain text region. An example of such research is the OED project conducted from 1987 through 1994 (Tompa and Gonnet, 1999; OED, 2004). One of the achievements of the OED project was that the search software was able to retrieve all occurrences of words and phrases within the dictionary corpus of size 570 MB in less than a second (Tompa and Gonnet, 1999).

Knowledge-base Structure of an ED. The second aspect of EDs is the structure of information represented in them. This structure is of interest to linguists, lexicographers, and various dictionary users, but it is of chief interest to computational linguists. A major computational challenge is how to design the dictionary structure in order to make its maintenance manageable and efficient. Various lexical resources that were developed in the last few decades have become invaluable in Natural Language Processing (NLP), most notably the WordNet. Another reason why efficiency in dictionary maintenance is important is that natural languages change dynamically and good ED should track these lexical innovations. Different domains need to be covered, and the parts of the dictionary that are becoming old and archaic need to be time-stamped and archived as such.

In this paper, we present a bilingual bidirectional on-line Serbo-Croatian (SC)-English dictionary that has been available on the Internet since 1999. This is the first published report describing this resource. The dictionary internal structure is motivated by the WordNet structure, and it provides a way of producing monolingual SC and bilingual SC-English wordnet.

2 Related Work

The OED project (Tompa and Gonnet, 1999; OED, 2004) is a related project that was discussed in section 1. There are many on-line dictionaries on the Internet: monolingual, bilingual, and even multilingual. Probably the most comprehensive list is given at the site YourDictionary.com[1], collected by Robert Beard from the Bucknell University, which lists on-line dictionaries for 294 languages, including two entries for sign languages (ASL and Sign).

There are not that many on-line SC-English[2]

[1]http://www.YourDictionary.com

[2]Under language name "Serbo-Croatian" (SC) we assume labels Serbo-Croatian, Serbian, Croatian, or Bosnian.

dictionaries. YourDictionary.com lists about five such dictionaries. Most of them are narrow-domain dictionaries. The Google directory[3] lists seven dictionaries. Rjecnik.com[4] is the oldest one in these language pairs and is still active and expanding. Tkusmic.com[5] was created in 2003 and has a very similar interface. One of the most popular dictionaries is Krstarica.com[6]. A long list of dictionaries is given at Danko Šipka's web site.[7] Many of those are not active any more, or they are textual dictionary files with a limited domain.

The WordNet (Miller, 2004) project is relevant to our work, since we propose a dictionary structure based on the building blocks that follow the WordNet structure. As a result, a direct by-product of our ED is an SC WordNet. The task of creating a Serbo-Croatian WordNet is already underway within the Balkanet project (Christodoulakis, 2002).

3 Project Description

Project history. The on-line dictionary R{j}ecnik.com has been active since 1999. One of its most visible characteristics, also noted by other users, is simplicity of the user interface. There is one search textual field in which the user enters the query and the dictionary reports all dictionary entries matching the query on either English or SC side. It provides an efficient search mechanism, returning the results within a second.

Lexical resources. As a lexicographic resource, this is a wide-coverage, up-to-date, bidirectional, and bilingual dictionary covering not only general, often used terms, but also over 8,000 computer and Internet terms,[8] as well as healthcare and medical vocabulary, including useful abbreviations. The entries are grouped by semantic meaning and part of speech, in the WordNet fashion. The English lexemes are associated with their phonetic representations, and the entries are marked by domain of usage (e.g., computers, business, finance, medicine). Colloquial and informal expressions are marked with special symbols so that they can be easily identified. In addition, the dictionary contains plenty of illustrative examples showing the language in use. A suitable text encoding for SC is used so that the software generates both Latin (Roman) and Cyrillic script versions. Dialectical and geographical differences are also marked.

Software overview. The dictionary software is developed in the Perl programming language. From the source dictionary file, the searchable on-line resource file is generated. It is in textual format and it is indexed through an inverted file index for searchable terms in English and SC. The searchable terms are chosen selectively. The tags and descriptions are not searchable since this would produce spurious search results.

Dictionary structure. Following the ideas from OED (Tompa and Gonnet, 1999), we adopted the philosophy of modern text markup systems that "*a computer-processable version of text is well-represented by interleaving 'tags' with the text of the original document, still leaving the original words in proper sequence.*" Additionally, we adopted the ideas from the WordNet project (Miller, 2004) in structuring our knowledge base around the basic entry unit being a meaning; i.e., one meaning = one entry. One source dictionary entry (vs. a printed, or on-line dictionary entry) corresponds to one synset in WordNet. It is represented in one physical line in a textual file, or it may be stored in several lines which are continued by having a backslash (\) character at the end of each line but the last one. An entry starts with the English lexemes separated by commas followed by an equal sign (=), and the corresponding SC lexemes, also separated by commas. Additional pertinent information is encoded using tags. This representation is conceptually simple and efficient in terms of manual maintenance and memory use. It is also flexible, since it allows tags to define features that refer to the whole entry or just individual lexemes. Such representation deviates from the commonly used XML notation because we find the XML notation to be more "machine-friendly" than user-friendly, but it can be automatically converted to XML. To illustrate the difference between TEI (Sperberg-McQueen and Burnard, 2003), the standard XML-based markup scheme, and our markup scheme, we adopt an example from (Erjavec, 1999), which is shown in in Fig. 1.

[3] http://directory.google.com

[4] http://rjecnik.com and http://recnik.com

[5] http://www.tkuzmic.com/dictionary/

[6] http://www.krstarica.com/recnik/

[7] http://www.public.asu.edu/˜dsipka/rjeynici.html

[8] A number of the terms was collected through public discussion at the e-mail list Serbian Terminology maintained by Danko Sipka (http://main.amu.edu.pl/mailman/listinfo/st-l).

(A)

```
<entry key="bewilder">
  <form>  <orth type='hw'>bewilder</orth>
          <pron>bI"wIld@(r)</pron> </form>
  <gramgrp><pos>vtr</pos></gramgrp>
  <sense orig='sem'>
    <trans><tr>zbuniti</tr>, <tr>zaplesti</tr>,
    <tr>zavesti</tr>, <tr>posramiti</tr>,
    <tr>pobrkati</tr></trans>
    <eg><quote>too much choice can bewilder a
             small child</quote>
      <trans><tr>prevelik izbor mo"ze zbuniti
             malo d{ij}ete</tr></trans>
    </eg>
</entry>
```

(B)

```
abash [\eb'ae"s], bewilder [biw'ild\er], \
  confound [kanf'aund], confuse [k\enfj'u:z]\
  = :v zbuniti, zaplesti, zavesti, posramiti,\
    :coll pobrkati :/coll, :eg too much choice\
  can bewilder a small child = prevelik izbor\
  mo"ze zbuniti malo d{ij}ete
```

Figure 1: Comparative example with TEI

The entry (A) in Fig. 1 shows an entry with TEI markup, in (B) we give our corresponding entry. The tags are preceded with a colon (`:`). English lexemes are associated with their phonetic representations within the square brackets. The phonetic representation is encoded using the *vfon* encoding.[9] All changes to the dictionary can be easily tracked down using the key `:id` tag and the standard CVS (Control Version System) system. The encoding *ipp* is used to encode SC text fragments, since they include additional letters beside the standard 7-bit ASCII set. The on-line version of the dictionary is encoded using the *dual1* encoding for simplicity and efficiency reasons. The input query can be entered using the ipp encoding, and is translated into the dual1 encoding before matching. The *krascii* encoding[10] is additionally accepted in the input query as the most common transcribing scheme, although it inherently leads to some incorrect matches.

A very systematic variation in SC is ekavian vs. ijekavian dialect; for example: mleko/mlijeko (milk) and primeri/primjeri (examples), but also hteo/htio (wanted). The text is converted via the following regular expression substitutions for ekavian and ijekavian: `s/\{(([^\|\}]*)\|)?([^\}]*)\}/$2/g` and `s/\{(([^\|\}]*)\|)?([^\}]*)\}/$3/g`.

The list of tags used in the dictionary is given in Fig. 2.

[9] The details about different encodings such as *ipp, vfon,* and *dual1* are provided in (Kešelj and others, 2004).

[10] Krascii is a simple transcribing scheme that ignores diacritics.

POS tags: noun (n), verb (v), adjective (a), adverb (adv), article (art), preposition (prep), conjunction (conj), interjection (interj), pronoun (pron), numeral (num), noun phrase (np), verb phrase (vp), symbol or special character (sym), and idiom (idiom).
Morpho-syntactic features: diminutive (dim), feminine (fm), imperfective (ipf), intransitive (itv), masculine (m), neuter (nt), past participle (pp), perfective (pf), plural (pl), preterite or past tense (pret), singular (sl), and transitive (tv).
Dialect tags: American (am), Bosnian (bos), British (br), Croatian (hr), Serbian (sr), and Old Slavic (ssl).
Domain tags: agriculture (agr), archaeological (archl), architecture (archt), biology (bio), botany (bot), computer (c), diplomacy (dipl), electrical (elect), chemistry (chem), culinary (cul), law (law), linguistic (ling), mathematics (mat), medicine (med), military (mil), mythology (myt), music (mus), religion (rel), sports (sp), and zoology (zoo).
Computer science subareas, cob *tag* (e.g., cob pl): internet (int), programing languages (pl), computational linguistics (cl), graph theory (gt), cryptography (crypt), data structures (ds), formal languages (fl), computer networks (cn), information retrieval (ir), and object oriented programming (oop).
Misc.: abbreviation (abb), abbreviation expansion (abbE), colloquial (coll), description (desc), example (eg), obsolete (obs), see (see), unique entry identifier (id), and vulgar (vul).

Figure 2: List of tags

Year	Avg.visits per day	Avg.time b/w visits	Len. of the longest query
1999	106	13m 34s	953
2000	249	5m 47s	710
2001	402	3m 34s	1556
2002	662	2m 10s	2492
2003	1018	1m 25s	4958
2004	2158	40s	1249

Figure 3: Site visit statistics

4 Dictionary and Usage Statistics

The dictionary has been on-line for five years (since 22-Jul-99). As of 28-Apr-2004, it has 60,338 lexemes, organized in 20,911 entries. The average system response time is 0.4 sec. Some site statistics are given in Fig. 3. The interface is supposed to be used only for short-word queries, but long queries are also submitted in hope that the system would do machine translation. As can be seen from the figure, the longest submitted query had the length of 4958 bytes. Still, the majority of the queries are below 100 bytes: in 1999 there were 0.03% queries sub-

1999	2000	2001	2002	2003	*2004*
95 love	522 love	854 love	1252 hello	1977 hello	607 hello
95 hello	499 hello	756 hello	1205 love	1776 love	590 love
70 you	346 you	521 you	892 you	1287 you	416 you
57 devojka	215 good	324 good	487 i	707 good	259 i
38 i	170 *f. . . (en)*	278 i	453 good	705 i	216 good
34 *k. . . (sc)*	158 i	264 devojka	341 *f. . . (en)*	578 thank you	204 da
34 djevojka	154 I	254 *f. . . (en)*	335 thank you	573 *f. . . (en)*	191 se
30 djak	148 devojka	252 thank you	333 happy	551 beautiful	191 thank you
30 *f. . . (en)*	144 are	243 happy	330 beautiful	499 are	189 beautiful
28 word	141 thank you	218 I	319 I	486 i love you	185 volim

Figure 5: The most commonly asked queries (*f. . . (en)* and *k. . . (sc)* denote obscene words

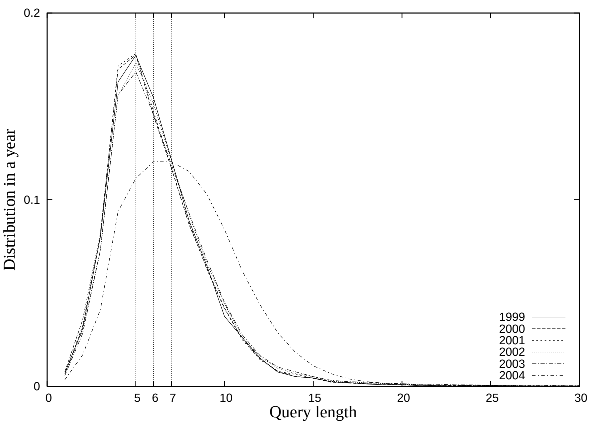

Figure 4: Distribution of query lengths

mited longer than 100 bytes, 0.05% in 2000 and 2001, 0.14% in 2002, 0.27% in 2003, and 0.12% in 2004. The distribution of query lengths less than 30 bytes is given in Fig. 4. The most commonly asked queries are given in Fig. 5.

5 Conclusions and Future Work

We have presented the features of an electronic English-SC dictionary. The dictionary is designed to be multi-functional, providing the interfaces to produce a printed dictionary copy and an on-line searchable lexicon. We propose a dictionary structure inspired by the WordNet, which is flexible and easy to maintain. We also report the site statistics of the on-line dictionary during the last five years.

Future work. The plan for future work includes incorporating a lemmatizer that would translate inflected word forms into their canonical representations. This is relevant for English, but it is a more important issue in SC, which is a highly-inflectional language. We do not know of any lemmatizer or stemmer currently available for SC. The software interfaces for producing a wordnet form, and a TEI-encoded form will be developed. An issue of long queries needs to be addressed. Currently, if a user submits a long query, which is usually a sentence or paragraph, the dictionary reports “zero entries found.” A fall-back strategy should be provided, which will consist of tokenizing the input and giving the results on querying separate lexemes.

6 Acknowledgments

We thank Danko Šipka, Duško Vitas, and anonymous reviewers for helpful feedback. The first author is supported by the NSERC.

References

D. Christodoulakis. 2002. Balkanet: Design and development of a multilingual Balkan WordNet. WWW.

T. Erjavec. 1999. Encoding and presenting an English-Slovene dictionary and corpus. In *4th TELRI Seminar*, Bratislava.

V. Keŝelj et al. 2004. Report on R{j}ecnik.com: An English — Serbo-Croatian electronic dictionary. Technical Report CS-2004-XX, Dalhousie University. *Forthcoming.*

G.A. Miller. 2004. WordNet home page. http://www.cogsci.princeton .edu/~wn/.

OED. 2004. Oxford English Dictionary. WWW. http://www.oed.com/, Apr. 2004.

C.M. Sperberg-McQueen and L. Burnard. 2003. Text encoding initiative. http://www.tei-c.org/P4X/index.html, accessed in May 2004.

F. Tompa and G. Gonnet. 1999. UW centre for the new OED and text research. http://db.uwaterloo.ca/OED/.

D. Vitas, C. Krstev, I. Obradović, Lj. Popović, and G. Pavlović-Lažetić. 2003. An overview of resources and basic tools for the processing of Serbian written texts. In D. Piperidis, editor, *First workshop on Balkan Languages and Resources*, pages 1–8.

D. Šipka. 1998. *Osnovi Leksikologije i Srodnih Disciplina.* Matica srpska.