Contents and evaluation of the first Slovenian-German online dictionary

Birte Lönneker
Institute for Romance Languages
Hamburg University
Von-Melle-Park 6
20146 Hamburg, Germany
birte.loenneker@uni-hamburg.de

Primož Jakopin
Corpus Laboratory
F. R. Institute of Slovenian language
ZRC SAZU, Gosposka ulica 13
1000 Ljubljana, Slovenia
primoz.jakopin@uni-lj.si

Abstract

This paper presents the first Slovenian-German and German-Slovenian online dictionary and contains evaluation figures for its Slovenian part. Evaluations are based on coverage of a Slovenian newspaper corpus as well as on user queries.

1 Introduction

The first Slovenian-German and German-Slovenian online dictionary is available at http://www.stud.uni-hamburg.de/users/birte/slo. Its current version, which was completed in November 2002, contains more than 4,800 entries covering the content of a beginners' textbook for Slovenian.

Section 2 gives some information about the contents and structure of the dictionary. Section 3 evaluates the Slovenian part, based on its coverage of lemmas in DELO, a Slovenian newspaper contained in the *Nova beseda* corpus[1] at ZRC SAZU (Beseda, 2000) as well as on its ability to fulfill user requests. Section 4 is the conclusion.

2 Contents and structure of the dictionary

In November 2002, the Slovenian-German-Slovenian online dictionary contained more than 4,800 entries which correspond to words, selected word forms and expressions appearing in the textbook *Odkrivajmo slovenščino*, a beginners' textbook for Slovenian (Čuk et al., 1996). The entire content of the textbook is covered.

The textbook is used in teaching Slovenian especially in German, French or Italian speaking communities.[2] It contains Slovenian-only text, explanations and exercises. As there is neither an index nor a vocabulary list, the online dictionary is a valuable completion of this educational material. However, it can be – and actually is – used independently of the textbook.

The current version of the online dictionary contains the following elements, based on *Odkrivajmo slovenščino*:

- all lemmas appearing in the texts, explanations, instructions and exercises;
- irregular inflected forms as well as the first person singular form of verbs;
- common conversational phrases and multiword expressions as well as some contextual examples of words and grammatical forms.

Grammar information for both languages and information on stressed syllables for the Slovenian entries are contained in additional fields. For both languages, three different kinds of search are possible:

[1] http://bos.zrc-sazu.si/a_beseda.html

[2] Personal communication from Meta Lokar, Centre for Slovenian as a Second/Foreign Language, University of Ljubljana.

1. exact match[3];

2. match a text string as a part of the dictionary entry;

3. match a text string at the beginning of the dictionary entry.

In the current version, the internal structure of the dictionary is a table containing one-to-one correspondences of words, word forms, and phrases. If an item has more than one equivalent in the other language, as many entries as necessary are created.

3 Evaluation

The evaluations of the Slovenian part of the dictionary concern its coverage of **a)** the corpus of the Slovenian newspaper DELO, as included in the ZRC SAZU corpus by the end of November 2002 (cf. Subsection 3.1); **b)** user queries to the dictionary which have been logged since the publication of the first trial version in April 2002 (cf. Subsection 3.2). Based on these analyses, some qualitative remarks about the most frequent missing items will be made (cf. Subsection 3.3).

The evaluation is based on the Slovenian part of the complete "vocabulary" (words, inflected forms, expressions) of the dictionary. The 4,841 entries actually contain 4,354 distinct Slovenian entries; this number is smaller than the overall number because some words are polysemous, or some expressions can have different translations. The minimum number of Slovenian lemmas in the dictionary can be approximated by counting those entries which either contain no space in both languages, or which are reflexive verbs (ending on "_se" in Slovenian or starting with "sich_" in German): There are 2,428 such entries.

3.1 Newspaper corpus coverage

To evaluate the coverage of texts by the Slovenian side of the dictionary, we chose the wordform list with frequencies of DELO, the main Slovenian daily, from January 1998 to August 2002.

[3]Exact match is case insensitive. Some characters or character combinations are treated in a special way in order to achieve matching of characters which might be difficult to enter, as German and Slovenian use different character sets.

words	60,843,505
distinct word forms	25,598
distinct lemmas or lemma sets	10,250

Table 1: Lemmatized word list for evaluation.

lemma(s)	distinct possible lemmas	word form	corpus frequency
absoluten:P	1	absolutni	797
absolutno:A;absoluten:P	2	absolutno	1,709
...	...	...	...

Table 2: Lemmatizer Output.

About 75% of the text of the Monday–Saturday edition is sent in ASCII format every day via e-mail to a small list of handicapped ("DELO for the blind") and to research users (*Nova beseda*). DELO is a good source for modern Slovenian, the text is spell-checked and proof-read, the error-rate is low (Jakopin, 2002). The results of our evaluation will give an approximation of how well the lexical knowledge represented in the dictionary – which can be interpreted as that of a learner after finishing the study of the textbook – overlaps with the lexical content of newspaper text.

The word list of the DELO newspaper corpus at ZRC SAZU in its November 2002 version contains 930,977 distinct word forms with an overall occurence of 73,412,302. Using the Corpus Laboratory lemmatizer[4] (Jakopin, 2002), the 30,000 most frequent word forms (with an overall occurrence of 64,465,582 and a coverage of 87.8% of the whole corpus) were lemmatized. 25,598 out of these 30,000 word forms were recognized by the lemmatizer. The recognized word forms, which cover 82.8% of the entire DELO corpus (cf. Table 1), will serve as the basis of our evaluation.

Word forms of each single lemma that corresponds to an entry in the dictionary will be counted as covered. For ambiguous word forms, the procedure is more complicated: In this case the lemmatizer output will consist of a set of possible lemmas (cf. Table 2). As only a part of the corpus is POS-tagged (Jakopin and Bizjak, 1997), these sets cannot be disambiguated. We decided to evaluate them by marking with an asterisk all those lemmas that are not covered by the dictionary; if at

[4]http://bos.zrc-sazu.si/dol_lem.html

lemma(s)	word form	corpus frequency
aids:S	aids	527
aids:S	aidsa	466
aids:S	aidsom	391
ali:Č,V	ali	198,399
...	...	...
avto:S	avto	7,450
avto:S;*avt:S	avta	2,576
avto:S;*avt:S	avtom	1,948
avto:S;*avt:S	avtu	1,519
...	...	...

Table 3: Covered lemmas and lemma sets.

lemma(s)	word form	corpus frequency
*absoluten:P	absolutni	797
*absolutno:A;*absoluten:P	absolutno	1,709
*absurdno:A;*absurden:P	absurdno	388
*administracija:S	administracija	833
...	...	...

Table 4: Not covered lemmas and lemma sets.

least one of the alternative lemmas is unmarked, the underlying word form will be counted as covered. Tables 3 and 4 show parts of the sorted result of the marking procedure.

Inflected forms of lemmas that appear in Table 3 are counted as covered by the dictionary. For example, all occurrences of *avta*, *avtom* and *avtu* will be counted as covered because the lemma *avto* ('car') is in the dictionary, even if the alternative lemma *avt* ('out', in sports contexts) is missing. We believe that this method is a good approximation of how much a dictionary user can understand of the lexical content of the newspaper text. In the case of non-related lemmas, one of them is usually much more frequent (as with *avto* and *avt*), whereas in the case of related lemmas, the meaning of the missing one can be inferred from the other (as with *bogat* 'rich' and *bogatiti* 'to enrich'; only *bogat* corresponds to an entry). Table 4 shows some lemmas and lemma sets which are not covered by the dictionary.

By this method, we find that 68.3% of the words in the lemmatized list from the corpus are covered (for detailed results, cf. Table 5). We notice, however, that not all lemmas in the dictionary (which were approximated to 2,428 lemmas) are actually among the most frequent ones of the corpus; for example, the textbook lemmas *kozolec*

	Lemmatized corpus	Covered by dictionary	Percentage covered
Words	60,843,505	41,564,382	68.3%
Word forms	25,598	6,640	25.9%
Lemmas	10,250	2,083	20.3%

Table 5: Corpus evaluation results.

	All queries	Top 100
Number	14,030	1,754
Distinct	7,188	105
Number covered	3,764	1,298
Distinct covered	1,068	73
Percentage covered (overall)	26.8%	74.0%
Percentage covered (distinct)	14.6%	69.5%

Table 6: Query evaluation results.

'hayrack' and *potica* (a special cake), which are introduced in order to present the Slovenian culture, but also *meduza* 'jellyfish' and *vedeževalka* 'fortune-teller', around which some textbook stories are centered, are not among those derived from the frequent word forms in the corpus.

3.2 Query coverage

By 15 November 2002, the trial version of the dictionary logged more than 34,000 requests. For evaluating the coverage of user queries, we compare the dictionary entries to the log file containing the 14,030 requests asking for a translation from Slovenian into German. The results for all queries as well as for the 100 most popular queries are shown in Table 6.

As can be seen from the table, the coverage of all queries is quite low (26.8% overall coverage and 14.6% coverage of distinct queries). This is due to the fact that the dictionary contains general basic words and expressions; user queries, however, range from basic to specialised vocabulary and include all sorts of expressions, spelling mistakes and even queries in languages other than Slovenian. If we look at the top 100 queries, however, the results are much better: 74.0% of all queries and 69.5% of distinct queries are covered.

3.3 Qualitative remarks

A closer look at the most frequent corpus words and user queries not covered by the dictionary shows the most serious gaps in the vocabulary. Interestingly, the top 30 missing lemmas, as ac-

lemma	English	corpus frequency
zaradi	for; because of	119,864
namreč	namely	63,553
torej	well; therefore	42,806
poleg	beside; besides	34,043
glede	with regard to; as for	33,668
okoli	around; round	25,363
pač	indeed; surely	24,529

Table 7: Frequent not covered closed class items.

quired from the corpus analysis and from the user requests, hardly overlap. The results show that in order to enhance corpus coverage, the insertion of seven frequent closed class items (prepositions, particles and conjunctions, which cover 343,826 occurrences, cf. Table 7), is as important as the insertion of the top seven missing nouns, which cover 351,323 occurrences. In contrast to these findings, user queries mainly concern open class items: Only two of the top 30 missing lemmas from the query evaluation are neither verbs nor nouns. For details on the distinction between open and closed word classes, cf. Greenbaum (1996).

The politico-economical context of the newspaper is reflected by nouns like *predsednik* 'chairman', *zakon* 'law', *minister* 'minister' and *podjetje* 'company', which are among the most frequent missing ones. Out of these, only *zakon* is also among the 30 most popular and unmatched user queries. An analysis of the unmatched user queries shows popular missing lemmas in the general domain, like *pozdrav* 'greeting; regards', *postaja* 'station', *krava* 'cow' and *odpad* 'waste; rubbish'. Economical or legal terms of daily life are popular as well: *pogodba* 'contract', *potrditi* 'to confirm', *račun* 'bill' and *davek* 'tax' can be mentioned as missing from the dictionary.

4 Conclusions and future work

Using an approximated method of untagged corpus coverage evaluation, we found that a dictionary of general Slovenian based on a beginners' textbook covers nearly 70% of the 82.8% most frequent lemmatized words in a big newspaper corpus. The textbook also contains lemmas which are less frequent in the corpus. A comparison of the corpus evaluation results with an analysis of the most popular user queries shows differences in the distribution of word classes among the most frequent unmatched items. Corpus coverage can be quickly enhanced by the insertion of some closed class items, while dictionary users are more interested in open class items.

The dictionary will be enlarged based on these quantitative and qualitative corpus and query analyses. Using a tagset covering the grammar information necessary for the Slovenian language (cf. e.g. http://bos.zrc-sazu.si/cgi/ckb_oo.html, (Jakopin and Bizjak, 1997)), the grammar information about Slovenian lemmas and word forms will be completed.

Acknowledgements

The first author thanks the Corpus Laboratory at the Fran Ramovš Institute of Slovenian language, ZRC SAZU, for research facilities and hospitality. Her three-month stay at this institution was supported by grants from the Ministry of Education, Science and Sport of the Republic of Slovenia and from the DAAD (*DAAD Doktorandenstipendium im Rahmen des gemeinsamen Hochschulsonderprogramms III von Bund und Ländern*). The Slovenian-German-Slovenian online dictionary has been awarded the Laurence Urdang EURALEX Award and is published with the kind permission of the editors of the textbook *Odkrivajmo slovenščino.*

References

BESEDA and its texts. Available: http://bos.zrc-sazu.si/a_about.html.

Metka Čuk, Marjanca Mihelič and Gita Vuga. 1996. *Odkrivajmo slovenščino.* Ljubljana: Filozofska fakulteta, Seminar slovenskega jezika, literature in kulture.

Sidney Greenbaum. 1996. *The Oxford English Grammar.* Oxford: Oxford University Press.

Primož Jakopin and Aleksandra Bizjak. 1997. O strojno podprtem oblikoslovnem označevanju slovenskega besedila. *Slavistična revija*, 45(3–4):513–532.

Primož Jakopin. 2002. Extraction of lemmas from a web index wordlist. *Abstracts of the 7th TELRI seminar, September 2002, Dubrovnik*, 8–9.