Bilingual Dictionaries and IDM DPS: The Development of a Corpus-driven Slovenian-English Pocket Dictionary and its Implementation in the IDM Dictionary Publishing System

Polonca Kocjančič, Simon Krek
Trojina, Zavod za uporabno slovenistiko
Partizanska cesta 5, SI – 4220 Škofja Loka, Slovenia
polonca.kocjancic@trojina.si
simon.krek@trojina.si – www.trojina.si

Philippe Climent
IDM
27, rue Albert Einstein
BP 117 Champs sur Marne
77423 Marne la Vallée Cedex 2, France
climent@idm.fr – www.idm.fr

Abstract
The paper gives an overview of the situation in corpus linguistics and bilingual lexicography in Slovenia, followed by a description of one of the recent projects, the compilation of the English-Slovenian and Slovenian-English Pocket Dictionary (DZS, 2006). The Slovenian-English part of this dictionary is entirely corpus-driven, which is a novelty in Slovenian bilingual lexicography. In spite of being a pocket dictionary, it has been a pilot project in many aspects, one of them being the implementation of its database in the Dictionary Publishing System, a dictionary editor that is being continuously developed for specific lexicographic customers by IDM, Paris. Thus the project has demonstrated new ways of managing bilingual dictionary data, and has shown the possibilities that IDM offers to meet the particular needs of various publishing houses.

1 An overview of Slovenian bilingual lexicography and recent developments in Slovenian corpus linguistics

Computer technology has brought many positive changes to the lexicographic treatment of the so-called smaller languages like Slovenian, a language which now enjoys a better sociolinguistic status than ever before.

After World War II, Slovenian bilingual dictionary production began to flourish; in the following decades, a number of new dictionaries were compiled. However, only a handful of dictionaries have been revised since then, which is why new uses and roles of the Slovenian language, developed after the change in the political system, have revealed major lacunae in

the existing general bilingual dictionaries. The process of replacing these dictionaries with those compiled by groups of authors using modern compilation techniques has been relatively slow. In the field of pocket dictionaries, there has been some competition to cover certain language pairs, but in practice the new dictionaries have been compiled using traditional and author-centred techniques.

What all of these dictionaries have in common is the constant problem of getting reliable data on the Slovenian language. The large monolingual dictionary *Slovar slovenskega knjižnega jezika* (Bajec 1970-1991) has been a primary reference for many users, but it is becoming increasingly outdated. A recent publication in a related area has been the *Slovenski pravopis* (ZRC SAZU 2001), but this is, in essence, a manual of style, although also comprising an extensive dictionary section mostly derived from the *Slovar slovenskega knjižnega jezika.*

Since early 1990s, Slovenian bilingual lexicography and corpus linguistics have been developing hand in hand: back in 1994, the comprehensive English-Slovenian Dictionary (*Veliki angleško-slovenski slovar Oxford-DZS*, DZS 2005) was initiated. As there were practically no recent Slovenian reference works to rely on during the compilation of this dictionary, the 100-million-word *FIDA reference corpus* (http://www.fida.net) was built, later followed by the *Nova Beseda* corpus (http://bos.zrc-sazu.si/s_beseda.html). Today, the *FIDA corpus* continues in an upgraded version, the *FidaPLUS reference corpus* (http://www.fidaplus.net). Upon completion, this corpus will contain 300 million words, and will reflect improvements in balance, lemmatisation and parsing.

2 The English-Slovenian and Slovenian-English Pocket Dictionary

2.1 General presentation

This dictionary has been designed as one in a series of pocket dictionaries to be published by DZS. In 2001, the English-Slovenian part was published as a separate dictionary, and preparations for the subsequent Slovenian-English part commenced. At various stages of its production, the dictionary, though small (15,000 entries), served as a pilot project, due to the initial decision that it would be the first dictionary in Slovenian to be corpus-driven, and that it should reflect the encoding needs of the Slovenian-speaking audience.

2.2 Specific topics

2.2.1 The corpus

The lexicographers were provided with the following data from the *FIDA corpus*:

- a wordlist containing 20,000 most frequent lemmas
- inflected forms of each lemma along with their frequencies
- collocates sorted in two ways: by frequency and by MI3
- concordance lines for each lemma

With slight variations, all of these are nowadays quite standard sets of data with which lexicographers are provided (Hanks 2004, McEnery et al. 1997: 229). For the compilation of the dictionary, the MI3 was used. The reason for this is that although the corpus was lemmatised and morpho-syntactically tagged, all the tagging was performed automatically and

without the possibility of removing ambiguities in cases where two, three or even more lemmas were possible. Because Slovenian is a morphologically complex language, statistical data from the corpus can be somewhat unreliable. Furthermore, there is quite a large number of non-lemmatised words. As the corpus concordancer enables both MI and MI3 statistical values to be implemented, analyses showed that the MI3 score was more relevant for the purpose of compilation (Gorjanc and Krek 2001). The fact is that non-lemmatised words are attributed high MI scores, while the MI3 score neutralises the effects of the low frequency of certain collocates in the corpus on account of more frequent lexical units. The following table illustrates the difference between MI and MI3 in the case of the noun *čaj* (tea) (3082 hits):

	MI score			MI^3 score		
1	=nefermentiran	1	3	skodelica	271	1483
2	=superčaj	1	2	kava	336	4902
3	=bančo	1	2	čaj	200	3082
4	=neslajenega	1	2	piti	184	5617
5	=oolong	2	3	zeliščen	74	480
6	=koprivnega	5	5	in	1169	2729097
7	=čren	4	4	kamiličen	31	51
8	0902	4	4	pitje	90	1254
9	=koprivin	3	3	metin	29	56
10	=luštrekov	3	3	biti	1473	7749214

= - nonlemmatized

Figure 1. Table of MI and MI3 scores for the lemma *čaj* (tea);
the first column following the collocate is the frequency of the combination of the collocate with the node, while the second column represents the absolute frequency of the collocate.

2.2.2 The compilation and presentation of dictionary material

To achieve the necessary consistency in the lexicographic treatment of corpus material, lexicographers were provided with a styleguide and a source SGML/XML database with an underlying DTD created specifically for the purpose.

The macrostructure of the dictionary was determined by analysing the initial wordlist against the corpus to confirm that a lemma was not overrated; the lexicographers' work on the macrostructure consisted mainly of clearing it of corpus noise.

The microstructure of the dictionary reflects the initial decision regarding target users: meaning discrimination is based on the principle of translation equivalents in the target language, and there has been abundant use of semantic indicators, collocates, typical structures and corpus-based examples of use.

Below we describe four categories that represent either new data in Slovenian lexicography or a difference in the compilation approach. Two of these belong to the level of macrostructure and two to the level of microstructure:

1) Macrostructure:

a) As a new feature in Slovenian lexicography, corpus-derived information on frequency is presented graphically in front of each dictionary entry with zero to three diamonds representing four levels of frequency (up to 199 hits; 200-999 hits; 1,000-9,999 hits; over 10,000 hits).

b) Another new feature is the dictionary treatment of those Slovenian adjectives which can have two forms:

- one representing quality, with zero ending
- one representing sort or kind, ending in *-i*

Thus, the adjective "white" has two forms, namely *bel* (white), as in *bel pulover* (a white pullover), and *beli* (white), as in *belo vino* (white wine). Traditionally, the canonical form of representing such adjectives in dictionaries has been the first one, with both forms treated within a single entry, in spite of the fact that very often the form in *-i* is prevalent, or even the only form possible. In the Slovenian-English pocket dictionary, *bel* is still given the status of a main entry, while *beli* is treated as a subentry.

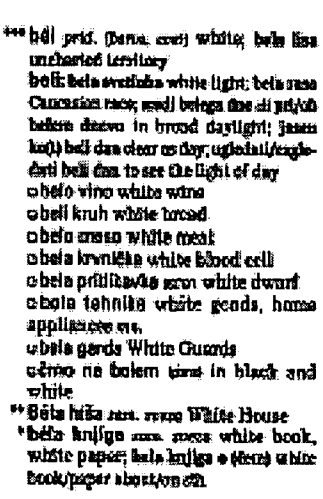

Figure 2. Treatment of the adjective "white" in the Slovenian-English pocket dictionary

2) Microstructure:

a) Treatment of collocators has been given special attention due to the statistical information provided by the corpus. Depending on the word class, a system of collocator listing has been developed. In the case of nouns, for example, there can be three kinds of collocates:

- adjectives: *[lažna, predvolilna] obljuba* ((false, pre-election) promise)
- noun complements: *oddaja [zemljišča, prostora]* (renting out (premises, a place))
- verbs: *[kotirati, trgovati] na borzi* ((to be listed, to trade) on the stock exchange)

b) The presentation of examples of use is another feature where the corpus makes an enormous difference compared to the traditional approaches. There are two major groups of examples:

• frequent and/or typical structural corpus-based examples like *dobiti brco* (to be fired/sacked) and *biti (pravi) magnet za (koga/kaj)* (to be a (powerful) magnet for (sb/sth));

• extended corpus examples, slightly edited where necessary, are used to represent a broader context; for example, the structure *marati za (koga/kaj)* (to care for sb/sth) is illustrated with *ni posebno maral za ženske* (he didn't especially care for women).

2.2.3 Possible future applications of the project

Firstly, the dictionary is the first finished product of the analysis of Slovenian corpus data as proposed in the description of the Slovenian lexical database by Gorjanc and Krek (2001) and Gorjanc et al. (2005) and, as such, it is representative of a radically new approach to the treatment of Slovenian language data. Although the compilation of a comprehensive Slovenian lexical database remains a task to be completed in the future, since the extent of the project exceeds by far the range and resources of one or several pocket dictionaries, this experience serves as a solid starting point for further developments in the field.

Secondly, the dictionary has opened up the possibility of amplifying the database with additional corpus data, as well as appending other language data, in order to compile a much desired comprehensive Slovenian-English dictionary.

3 Bilingual data and the IDM Dictionary Publication System (IDM DPS)

3.1 Editorial programs – general overview

The question of dictionary editorial program choice and the final format of a dictionary database involves two quite different segments. Firstly, the program in which a lexicographer edits a database should be as comfortable as possible; it should be adapted to the specific needs of the compilation process. Traditional editorial programs are not sufficient for the task, because dictionary databases contain both text as well as strong internal structure. Furthermore, dictionary editorial programs should take account of the fact that the lexicographer needs quick access to certain closed sets of content elements, typical hierarchical structures of entries, or to be able to perform complex database searches. There is a second requirement that is quite independent from the first point: the ability of the editorial program to store and export the database in a hierarchically structured XML format.

To compile the Slovenian-English pocket dictionary, a simpler dictionary editing program was inherited from previous dictionary projects as the software tool that could meet the basic needs of dictionary compilation. As better solutions for future projects were scrutinised, IDM turned out to be one of the companies which provided solutions for numerous technical dilemmas raised in the course of the project. Therefore, it was agreed to test the bilingual database using the IDM DPS.

3.2 The IDM Dictionary Publishing System (IDM DPS)

The DPS (Dictionary Publishing System) comprises a comprehensive range of tools aimed at helping lexicographers, dictionary editors and publishers to undertake dictionary projects, whether updates or new creations, bilingual or monolingual. Its native support of the Unicode character set, along with the concept of editing task splitting, makes it very use-

ful for international projects involving freelance lexicographers nation-wide or worldwide (http://www.idm.fr/DictionarySolutions.htm).

DPS enables lexicographers to focus on the lexicographic process without being constrained by the "mechanics" of dictionary production. This means that the specifics of the database are in a way hidden to the lexicographer; he or she does, of course, have to be thoroughly familiar with the hierarchical features and demands of the database, but is spared much of the tiresome work with SGML/XML elements that is so typical of many other SGML/XML editors.

For dictionary editors, DPS provides much easier management of all kinds of editorial tasks via a web administration tool. These tasks include stage management and editing tasks management. Thus a dictionary editor can create, edit or delete the stages through which a database must pass in order to achieve the desired goal; these stages are then automatically ascribed to particular editing tasks. To name but one particular feature: if the editor has to give a lexicographer a certain number of entries to edit, he or she chooses the entries from a central database, marks them as a new editing task, adds the person in charge of editing, the deadline and other details. The lexicographer then downloads the editing task, and after completing the file uploads it again. Once uploaded, the editing task is merged back into the database or moved into the next editing stage.

3.3 The Implementation of the Slovenian-English part of the dictionary in the IDM DPS

A sample of the newly compiled database was incorporated into the IDM DPS as an example of handling bilingual dictionary material. To perform the test, a sample of the database and the DTD was handed to IDM, who created editorial server accounts, converted the source SGML database into a more stringent XML format and made other necessary adaptations involving the dictionary layout.

The Dictionary Editor consists of three main panes:

- the Table of Contents, which comprises the dictionary's wordlist (left-hand column),
- the Treeview, which is the lexicographer's primary editing tool; it gives an overview of a document's structure and enables easy manipulation of SGML/XML elements, free text editing and adding closed-set texts (right-hand above)
- the Preview, which reflects the final image of the entries as they are to appear in the prospective end-product (right-hand below)

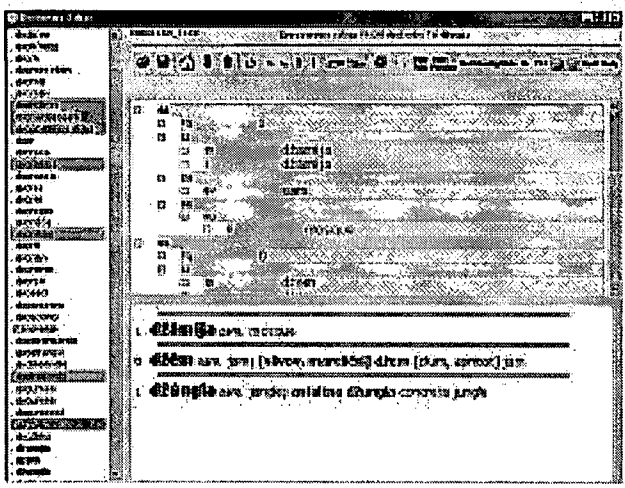

Figure 3. Sample bilingual entries in IDM DPS

To illustrate the centralised Web Administration tool, the administration task management page has been chosen; in this page, the administrator prepares a task file and assigns the lexicographer, deadline and editing stage.

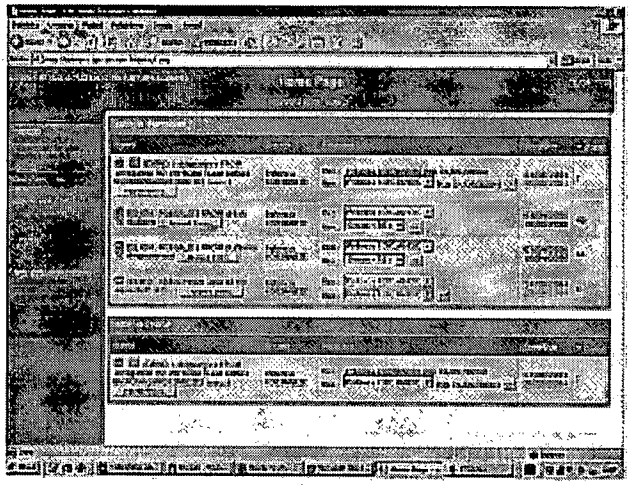

Figure 4. One of the server pages of IDM DPS

One of the questions necessarily raised when handling dictionary data is font support. IDM DPS does have the technical ability to support Unicode characters, but the complete integration of one or several language supports remains a task to be addressed within a larger project and a broader lexicographic and computational linguistics community.

4 Conclusion

The paper first describes the compilation of a corpus-driven Slovenian-English pocket-size dictionary. To give the project a broader perspective, current developments in Slovenian bilingual lexicography are described, followed by a more detailed discussion of certain macrostructural and microstructural aspects of the dictionary. In the continuation, the presentation focuses on general aspects of software support in the dictionary-making process, followed by a brief description of the implementation of the bilingual dictionary database in the IDM Dictionary Publishing System.

References

A. Dictionaries

Krek, S. (ed.) (2005), *Veliki angleško-slovenski slovar Oxford-DZS. A-K.* Ljubljana, DZS.

Krek, S. and Zaranšek, P. (eds.) (2006), *Mali angleško-slovenski in slovensko-angleški slovar.* Ljubljana, DZS.

Bajec, A. et al. (eds.) (1970-1991), *Slovar slovenskega knjižnega jezika.* Ljubljana, DZS.

Slovenski pravopis (2001), Ljubljana, Založba ZRC.

B. Other Literature

Gorjanc, V. (2005), *Uvod v korpusno jezikoslovje.* Ljubljana, Izolit.

Gorjanc, V. and Krek, S. (2001), 'A corpus-based dictionary database as the source for compiling Slovene-X dictionaries', in *Proceedings of the COMPLEX 2001 / 6th conference on Computational Lexicography and Corpus Research.* Birmingham, pp. 41-47.

Hanks, P. (2004), 'Corpus Pattern Analysis', in Williams, G. and Vessier, S. (eds.), *Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress, (Euralex 2004),* Volume I. Lorient, Université de Bretagne Sud, pp. 87-97.

McEnery, A., Langé, J.-M., Oakes, M., & Véronis, J. (1997), 'The exploitation of multilingual annotated corpora for term extraction', in Garside, R., Leech, G., McEnery, A. (Eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, London: Addison Wesley Longman, pp. 220-230.