The Retrieval of Data for Slovene-X Dictionaries

Mojca Šorli
Trojina, Institute for Applied Slovene Studies

The present article reflects on the linguistic issues concerning the preparation of text for a new Slovene-English dictionary. Discussion is based on concrete examples from the reversed database of the Oxford-DZS Comprehensive English-Slovene Dictionary and lexicogrammatical data from a corpus-based Slovene lexical database in the making. The key question is how to proceed with the compilation of a new, bilingual, dictionary database, using both sources but avoiding a distorted lexical analysis of Slovene in use, while also ensuring a thorough contrastive analysis of the relationships between the two languages.

1. Introduction

It is yet to be established how successful retrieval of data from a reversed bilingual database is. However, the first attempts to use information from the reversed Oxford-DZS Comprehensive English-Slovene Dictionary (2005/2006), more precisely from the Reversed English-Slovene Database (henceforth RESD), for the purposes of the compilation of a new Slovene-English dictionary indicate that the automatically generated database is a vast fund of information on the contrastive relations between English and Slovene, which should at no cost be overlooked.[1] The user has instant access to the potential direct translation candidates, and to the more contextually-bound potential translations, many of which would have been inaccessible to a non-native speaker without an insight into what could be called the mirror image of the language. On saying that, it is important to stress that a reversed database as we understand it is in no way to be confounded with the actual 'reversed' dictionary itself, but merely to be seen as a bilingual framework in which no solution is automatically transferred to a Slovene-English dictionary. We come to the conclusion that while a corpus-based monolingual database is needed to provide a fresh and authentic image of the source language, it is also important to explore and exploit the data obtained in the reversed bilingual database because that will add an extra dimension to the Slovene-X dictionary text.

1.1. Suggestions and solutions

Ideally, taking into account the advantages brought about by applying both monolingual and bilingual perspectives simultaneously, the process of compiling a Slovene-X dictionary would follow the stages listed below:

1st stage: SLOVENE LEXICAL DATABASE (corpus-based lexical data on the headword as recorded in its monolingual environment)
2nd stage: TRANSLATED SLOVENE LEXICAL DATABASE (translation of the material into the target language using information in the reversed database, TL corpora and other sources)
3rd stage: NEW SLOVENE-X DICTIONARY DATABASE (selection and partial bilingual data organisation with respect to the TL perspective)
4th stage: NEW BILINGUAL DICTIONARY

[1] The reversing procedure was aimed at maximizing the information potential of the detailed XML structure and at rendering the final result in a user-friendly manner (for more detail see Krek et al. 2008).

2. Slovene Lexical Database (henceforth SLD)[2]

Currently, a new Slovene lexical database of 2500 entries is being compiled, with two major goals in view:

- serving the purposes of lexicography, both for general monolingual and bilingual dictionaries for Slovene, which are strategically the most important, as well as for specialised reference books,
- meeting the needs of NLP and human language technologies in the context of Slovene

In order to be able to achieve these goals the following principles have been adopted:

- first and foremost, the database will include that which in language is typical: the core vocabulary, core meanings, syntactic structures and patterns, collocations, phraseology, and, to support all of these categories, corpus examples;
- the focus is on the detailed sense discrimination of the core vocabulary;
- besides semantic data, syntactic and other grammatical data is of interest in so far as it is relevant for lexicographical and grammatical descriptions.

2.1. Lexical data organisation

Lexicogrammar is the key organising principle of the database structure; each headword (lemma) is analysed on three levels: semantic, syntactic and collocational or syntagmatic. Lexical units are semantically disambiguated (senses and subsenses) by semantic indicators. Subsequently, each sense and subsense is ascribed a semantic frame, to a certain extent inspired by FrameNet.[3] The organising principle of the so-called semantic frame is that of a 'meaning scenario,' which defines the range of semantic and syntactic combinatory possibilities. However, the focus of interest for the SLD remains an individual lexical unit rather than semantic relationships entered into by groups of lexical units. The syntactic level displays syntactic patterns in which individual senses are realised. This part of the SLD is 'machine-readable' and primarily designed for natural language processing. For lexicographical purposes, syntactic patterns are also listed as syntactic structures, indicating concrete grammatical restrictions or so-called colligations.

At the syntagmatic level, typical collocations for individual (sub)senses are listed, entering into relationships with the corresponding syntactic patterns. Extended collocations are also recorded, as well as instances of text that are typical but not lexicalised to the point of requiring an explanation. Phraseological units are listed in a separate section at the end of an entry. Each phraseological unit contains meaning information in the form of a semantic indicator accompanied by all of its syntactic realisations and by corpus examples. Understandably, the structuring of information in the SLD has implications for the design of a potential bilingual lexical database.

---

[2] The operation is partly financed by the European Union, the European Social Fund, and the Ministry of Education and Sport of the Republic of Slovenia (2008-2013). The two basic sources of data are the FidaPLUS Slovene Reference Corpus (http://www.fidaplus.net), and Word sketches, integrated into the Sketch Engine (SkE) corpus query tool (Kilgarriff et al 2004) (http://www.sketchengine.co.uk).

[3] *International Journal of Lexicography*, Vol. 16, No. 3, September 2003, Special Issue: Framenet and Frame Semantics.

3. The relationship between the RESD data and the SLD data

3.1. Monodirectional or bidirectional translation equivalence

Typically, the unit exposed in the source language is semantically complete and it should, ideally, be rendered in the target language with an equal degree of naturalness and/or semantic fixedness. Due to the fact that the left side is determined by the source language, translation differences are quite common. The central issue is not so much how to achieve equivalence at the level of lexical meaning, but primarily how to generate equivalence at the level of the typical or of the frequent. About the asymmetry in translation dynamics, or about why it is that *denar je vir vsega zla* equals *money is the root of all evil* and vice versa, and why it is not that *everything is in apple pie order* equals *vse je čisto tako, kot mora biti* and vice versa; about why *tub-thumper* (informal) equals *bombastičen govorec/bombastična govorka*, but *bombastičen govorec/bombastična govorka* does not equal *tub-thumper* (informal) etc., it is possible to read more in Krek et al. (2008).

The question as to extent to which it is possible to describe a target language using the logic of reversal is complemented by the question of how to treat the contrastively relevant information in the RESD that does not occur in the currently compiled SLD. This can be the case either because a word or phrase is insufficiently represented in the corpus of Slovene, or because it does not feature in the Word sketch because the grammatical relation that would convey the word's typical contexts of use has not (yet) been defined in terms of Sketch grammar, such as the following phrases: *vsake toliko časa = once in a while*; *čez nekaj časa = after a (short) while*; *(kar) nekaj časa = for (quite) a while, for a (good) while*; *za časa (svojega, njihovega, njegovega) življenja = in (one's, their, his) lifetime*. It is justifiable and desirable to include these phrases in the SLD as the cross-cultural perspective shows that they are to a larger degree context-independent and/or semi-idiomatic. They all prove to be adequately represented in the corpus. In the RESD quite a few phraseological units are recorded that are not particularly frequent in the corpus, but in some cases it is to be assumed that this is due to the particular taxonomy of the texts of the selected corpus. On the other hand, and conversely, it is possible to overlook contrastively relevant instances of text when focusing exclusively on the source language situation. Understandably, this data varies according to the target language in question. Amongst the collocates [razpolaganje, ravnanje, razmetavati, razpolagati, financirati] = *[managing, handling, to waste, to have at one's disposal, to finance]*, which for the entry 'money' are found in the SkE under the gramrel 'prec z-d', from the perspective of Slovene it is quite easy not to recognise *razmetavati z denarjem* as a unit potentially translated into English idiomatically, with *to splash (money) around*. Maybe even more hidden under the gramrel 'post verb' amongst the collocates [porabiti, nakazati, nameniti, *vrniti*, zapraviti] = *[to spend, to pay into account, to allocate, to return, to waste]* is *to return*, which at first sight shows no special bilingual relevance (we presuppose: *to return the money, to pay back the money*), but a more detailed examination of the RESD and its broader contexts of use shows: *v nobenem primeru kupcem ne bomo vrnili denarja = in no case will customers be refunded*. It is even harder to predict which of the collocates could potentially form a compound, e.g., *pretok denarja = cashflow*. The above examples demonstrate how difficult (or impossible) it is to select and order collocates within a monolingual lexical database with respect to bilingual relevance.

For bilingual purposes, at stage 3 it is possible to benefit considerably from the data in the RESD by comparing it to the data in the SLD acquired through corpus analysis. The RESD is of particular value in searching for the appropriate English collocates – by entering a word or

a collocation, such as *(stisniti) pest*,[4] we get *[to ball, to bunch, to clench] one's fist*. Particularly relevant are those collocates in the SLD that alone or in combination with the headword produce an unpredictable, idiomatic, often single-unit, translation, e.g., *isker konj= courser.* In the case of the headword *konj* (=*horse*), these are practically all amongst its 'modifier' collocates[5] [dirkalni = *racing*, jahalni = *riding*, vprežni = *harnessing*, vlečni = *pulling*, tekmovalni = *racing*, kasaški = *trotting*] and [divji = *wild*, čistokrven = *pure blood*, plemenit = *noble*, isker = *lively*]. It is precisely looking for single-word equivalents of the source language multi-word lexical units that is one of the most challenging translation tasks. The equivalent with an unrelated morphemic base is virtually impossible to locate, no matter how good the monolingual sources, including the corpus. There is simply nothing in the collocation *isker konj* that, given our general knowledge of English, would lead us to the idiomatic expression *courser*. Similarly, if somewhat more predictably, we find the translation *to gallop (a horse)* for *pognati (konja) v galop*. In accordance with the information in the RESD, fixed collocations bordering on compounds and their equivalents include: [dirkalni] konj = *racehorse, racer, runner*, [jahalni] konj = *riding horse, saddle horse*, [vprežni] konj = *draught horse, carthorse, coach-horse*, [vlečni] konj = *driving horse*, (fig.) *workhorse*, [kasaški] konj = *trotter*, [divji] konj = *bronc* (informal), *bronco*, or [čistokrven] konj = *blood horse, purebred* in [plemenit] konj = *bloodstock*. A high degree of bidirectional equivalence is demonstrated at the level of idiomatic expressions. This is also true of compounds that are frequently (semi)terms, such as 'povodni konj' = (Zool.) *hippopotamus, hippo* (informal), 'trojanski konj' (Computers) = *Trojan (horse)* etc. Like information contained in usage examples, information on collocates is manifold: less contextually bound lexemes – such as those denoting something from the material world, e.g., horse, pneumonia, sun – form various relations with the headword, frequently featuring as examples of use or compounds. Indeed, often a collocate-headword combination is in itself an example of usage, that is if the minimal context is sufficient for a clear and authentic translation, e.g., *dobljena tekma = a won match, obeta se nam dež = we're in for more rain*.

Some collocate-headword combinations show strong semantic prosodies. Headwords such as the verb *obetati* require a particularly detailed analysis of their collocational behaviour. Unlike for most headwords, rather than just listing amongst collocates, say, statistically the most salient nouns within the determined span as listed in the Word sketch, it is virtually obligatory to record their entire collocational environment, particularly the adjective-noun collocational pattern. This is because most nouns are collocates of *obetati* only in restricted contexts, with particular adjectives: *obetajo se boljši časi* (=*better times are ahead*) – *čas* ('time') is a collocate for *obetati* almost exclusively in the context of *boljši* ('better'). In general, some headwords show a greater demand for wider collocational contexts, and in these cases writing a bilingual entry will presumably require additional corpus analysis to show which translations are the most commonly related to individual collocates, and whether analogous semantic prosodies can be detected in the target language.

4. Conclusion

The compilation of a Slovene-X dictionary requires an optimal exploitation of data both from a monolingual and bilingual perspective. The corpus-based Slovene Lexical Database provides an image of typical and idiomatic Slovene that should be complemented by the contrastively relevant information from the reversed bilingual database. If data from the latter

[4] The first, most literal translation would be *to squeeze one's fist*.

[5] Literal translations of all of the collocates are given, for true equivalents see below.

is applied without critical consideration of its nature there is a danger of morphosyntactic as well as semantic distortion in the lexicographic description of the source language. To avoid this the relevant data is double-checked in the corpus. On the other hand, if focus is entirely on the source language reality, it is quite easy to miss contrastively relevant information. Before the compilation of a Slovene-X dictionary it is mandatory to build an adequate bilingual database.

Bibliography

Atkins, B. T. S.; Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Fillmore, C.; Johnson, C. R.; Petruck, R. L. (2003). 'Background to Framenet.' In Bogaards, P. (ed.). *International Journal of Lexicography*. Oxford: Oxford University Press. 16/3. 235-250.

Gantar, P. et al. (2008–2013) (in the making): *The Slovene Lexical Database*. (Compiled as one of the activities within the »Communication in Slovene« project, co-financed by the European Social Fund and the Ministry of Education and Sport of the Republic of Slovenia).

Gantar, P. et al. (2009). *Specifications for the compilation of the Slovene Lexical Database* (http://www.slovenscina.eu/Media/Kazalniki/Kazalnik6/SSJ_Kazalnik_6_Specifikacije-leksikalna-baza_v1.pdf.)

Kilgarriff, A.; Rychly, P.; Smrž, P.; Tugwell, D. (2004). 'The Sketch Engine'. In *Proceedings / XI EURALEX International Congress*. Lorient, France. 105-116.

Krek, S.; Šorli, M.; Kocjančič, P. (2008). 'The Funny Mirror of Language: the Process of Reversing the English-Slovenian Dictionary to Build the Framework for Compiling the New Slovenian-English Dictionary'. In Bernal, E.; DeCesaris J. (eds.). *Proceedings / XIII EURALEX International Congress*. Barcelona. Institut Universitari de Linguistica Aplicada: Documenta Universitaria.

Krek, S.; Kilgarriff, A. (2006). 'Slovene word sketches'. In Erjavec T; Žganec Gros, J. (eds.). *Jezikovne tehnologije*. Institute 'Jožef Stefan', Ljubljana. 62-67.

Krek, S. (ed.). (2005-6). *Veliki angleško-slovenski slovar Oxford* [The Oxford-DZS Comprehensive English-Slovene Dictionary]. Ljubljana: DZS.

Šorli, M. et al. (2006). 'The Oxford-DZS Comprehensive English-Slovenian Dictionary'. In Corino, E.; Marello, C.; Onesti, C. (eds.). *Proceedings / XII EURALEX International Congress*. Torino: Edizioni dell`Orso: Università di Torino: Academia della Crusca. 631-637.