The Funny Mirror of Language: The Process of Reversing the English-Slovenian Dictionary to Build the Framework for Compiling the New Slovenian-English Dictionary

Simon Krek
Jozef Stefan Institute

Mojca Šorli
Trojina

Polonca Kocjancic
University of Ljubljana

The article describes the process of reversing the English-Slovenian dictionary database in XML format to create the framework for compiling the Slovenian-English dictionary. The aim was to maximize the abundance of information in an extensive dictionary database with a complex and detailed structure. The process involved lemmatization and POS-tagging of both source and target languages, construction of routines to form the preliminary list of possible headwords and their translation equivalents, as well as routines which enabled the grouping of numerous dictionary examples available in the original dictionary under the appropriate translation equivalent. The result is the reversed dictionary database in XML format with the DTD and XSL file to control the layout for viewing the database in Internet browsers or other XML-aware-dictionary-editors. The article presents the process of reversing the dictionary and the features of the final database. It also reflects on the linguistic issues concerning the fact that the database represents only the mirror image of the English-Slovenian contrastive relation and argues that the contrastively undistorted lexical information from a monolingual Slovenian reference corpus has to be taken into consideration when compiling the new Slovenian-English dictionary.

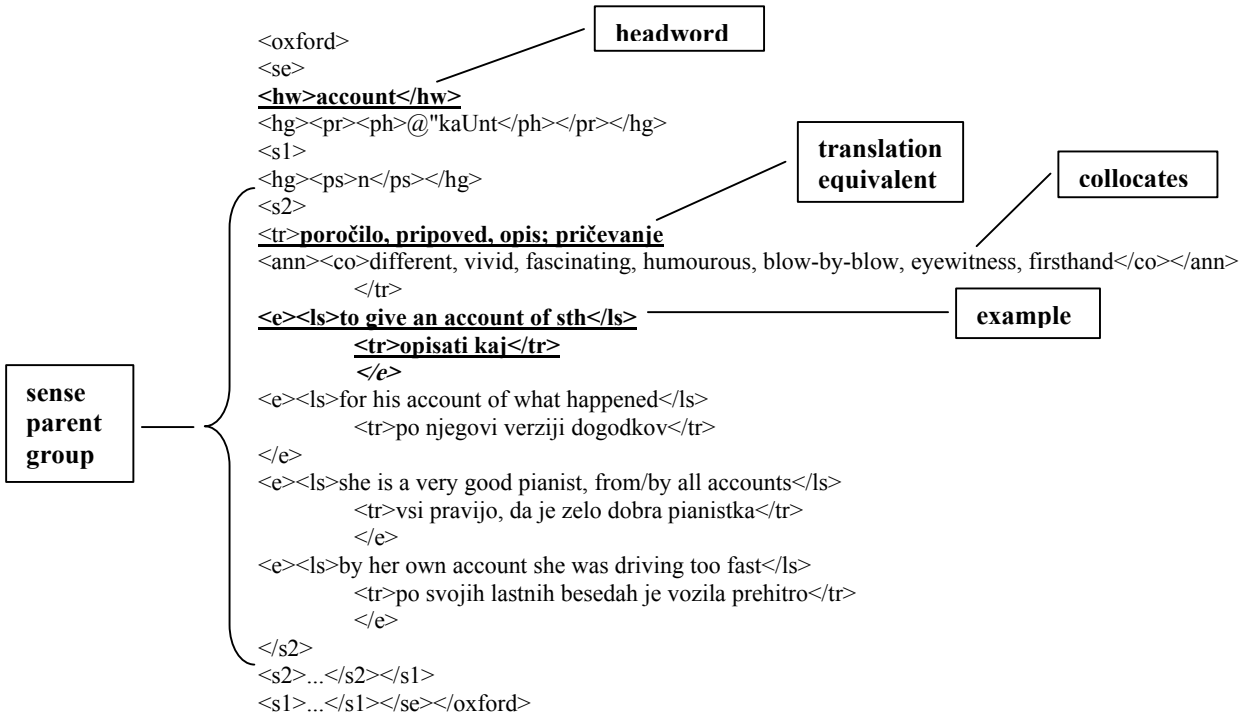
1. The source

The source of the dictionary data involved in the process is the SGML/XML database of the printed Oxford-DZS Comprehensive English-Slovenian Dictionary (Krek 2006)[1]. The dictionary is the most comprehensive English-Slovenian dictionary to date—a corpus-based work with approximately 120,000 entries. It is based on the Oxford-Hachette French Dictionary, expanded and designed as a monodirectional bilingual tool for the Slovenian native speaker community. Perhaps lexicographically the most interesting feature of the dictionary is its monodirectional concept resulting in particular choices regarding the headword list, microstructural features and the layout design of the printed dictionary.

The SGML/XML structure of the dictionary was based on the near-SGML data from the original Oxford-Hachette French Dictionary and converted to the standard SGML format with elements and features added, discarded or changed to conform to the new dictionary structure according to a style guide that was, in a broader sense, defined in 1997.

An example from the English-Slovenian dictionary in the SGML/XML format:

[1] As described in Šorli et al. (2006), Grabnar and Šorli (2003) and Hočevar (2006).



Example 1: the Oxford/DZS Comprehensive English-Slovenian Dictionary

A detailed description of the structure would be quite historical as the majority of dictionary databases are now in the SGML/XML format. However, it seems important to show a minuscule example of the hierarchical information available to the designers of the system. Translation equivalents are separated either by a comma or a semi-colon within the same XML element regardless of their mono- or multi-word status. If parts of a multi-word expression are the same, a slash was used to avoid repetition (ex. Slo. translation: *aberantni/spremenjeni kromosomi* Eng. headword: aberrant chromosomes), but they can be resolved fully automatically as the rule was to use the slash only in cases where the left and right collocate can be combined uniformly with the head. On the other hand, headwords are always in separate elements, even if variant spellings or other headword alternatives are listed.

2. The intermediate database

The goal of the dictionary reversing exercise was to gain maximum benefit from the abundance of information in the dictionary database and to make use of the complex dictionary structure. The starting point of the reversing process was translation equivalents found in the <tr> element, separated by a comma or a semicolon, which were treated as the immediate candidates for the Slovenian-English dictionary headword list. However, several decisions had to be taken already in this first step. Namely, one of the decisions in the process of compiling the English-Slovenian dictionary was to distinguish clearly between translation equivalents insertable in the context and glosses that were not useful as translation equivalents as they only explain the meaning of English headwords. It is clear that only translation equivalents were considered as possible headwords in the reversed dictionary but, as every lexicographer knows, the line between a proper translation equivalent and a gloss is sometimes rather thin. Therefore, we had to decide whether to list all multi-word translations as possible headwords or to apply stricter criteria for the candidates. Analysis showed that it seemed more reasonable to list all translation equivalents, even multi-word candidates, since lexicographers can easily skip the obviously non-eligible headwords without the risk of losing a multi-word unit to be used as the future dictionary example under a single-word headword.

The other feature of the dictionary we wanted to capitalize on was its abundance of examples, particularly those that were only kept in the database and not used in the printed dictionary because they were considered to be more useful for the Slovenian-English contrastive analysis than for the English-Slovenian dictionary. However, in order to be able to put the examples to maximum use we had to consider the lemmatization and POS-tagging of the Slovenian translations. Slovenian is a

highly inflected language and, as opposed to the immediate translation equivalents, wordforms in the examples were mostly inflected. In order to have the possibility to list them under the corresponding Slovenian headword, they had to be lemmatized. As a result, both the Slovenian translations of the examples and the immediate translation equivalents (in the case of multi-word units) were lemmatized and POS-tagged in the entire dictionary. The process was done fully automatically by the proprietary tagger owned by Amebis, the leading software company in Slovenia involved in NLP.

Furthermore, in dealing with the dictionary examples it proved to be very important whether or not the original English dictionary example contained the future English translation equivalent already detected and used as the possible translation candidate in the Slovenian-English dictionary. Although English morphology is rather simple, it was also necessary to POS-tag the English part of dictionary examples in order to be able to detect the existence of the translation equivalents.

Here is one example from the resulting database with additional linguistic information and prepared for its use in the dictionary reversing process:

```
<se id="402">
<hw>account</hw><hg st="493"><pr><ph st="294">{#@}{#"}ka{#U}nt</ph></pr></hg>
<s1><hg st="494"><ps>sam.</ps></hg>
 <s2 1/9>
/*****
 <tr>
  <b+tr>poročilo, pripoved, opis; pričevanje
   <ozn>
    <w l="poročilo/S">poročilo</w>
          <g>poročilo/S</g>
    <w l="pripoved/S">pripoved</w>
          <g>pripoved/S</g>
    <w l="opis/S">opis</w>
          <g>opis/S</g>
    <w l="pričevanje/S">pričevanje</w>
          <g>pričevanje/S</g>
   </ozn>
  </b+tr>
 <ann><co>different, vivid, fascinating, humourous, blow-by-blow, eyewitness, firsthand</co></ann></tr>
 <e+>
  <ls>
   <a>to give an account of sth
    <ozn>
     <w l="to/R">to</w>
     <w l="give/G">give</w>
     <w l="an/Z">an</w>
     <w l="account/S">account</w>
     <w l="of/D">of</w>
     <w>sth</w>
    </ozn>
   </a>
  </ls>
  <etr>
   <b+etr>opisati kaj
    <ozn>
     <w l="opisati/G">opisati</w>
           <g->opisati/G</g->
     <w s="n" l="*kaj/Z kaja/S">kaj</w>
    </ozn>
   </b+etr>
  </etr>
 </e+>
 <e+>
  <ls>
   <a>by his account of what happened
    <ozn>
     <w l="by/D">by</w>
     <w l="his/Z">his</w>
     <w l="account/S">account</w>
     <w l="of/D">of</w>
```

immediate Slovenian translation equivalents

possible candidates for Slovenian headwords, lemmatized

the English part of the example

lemmatization of the English part of the example

the Slovenian translation of the example - lemmatized and POS-tagged

```
   <w l="what/Z">what</w>
   <w l="happen/G">happened</w>
  </ozn>
 </a>
</ls>
<etr>
 <b+etr>po njegovi verziji dogodkov
  <ozn>
   <w s="n" l="po/D">po</w>
   <w s="n" l="njegov/Z">njegovi</w>
   <w l="verzija/S">verziji</w>
          <g->verzija/S</g->
   <w l="dogodek/S">dogodkov</w>
          <g->dogodek/S</g->
  </ozn>
 </b+etr>
</etr>
</e+>
<e+>
 <ls>
  <a>she is a very good pianist, from
   <ozn>
    <w l="he/Z">she</w>
    <w l="be/G">is</w>
    <w l="a/Z">a</w>
    <w l="very/P">very</w>
    <w l="good/P">good</w>
    <w l="pianist/S">pianist</w>
    <w l="from/D">from</w>
   </ozn>
  </a>
 <ann>/by</ann>
  <a> all accounts
   <ozn>
    <w l="all/Z">all</w>
    <w l="account/S">accounts</w>
   </ozn>
  </a>
 </ls>
 <etr>
  <b+etr>baje je zelo dobra pianistka
   <ozn>
    <w l="baje/L">baje</w><g->baje/L</g->
    <w s="n" l="*biti/G jesti/G">je</w>
    <w l="zelo/R">zelo</w><g->zelo/R</g->
    <w l="dober/P">dobra</w><g->dober/P</g->
    <w l="pianistka/S">pianistka</w>
          <g->pianistka/S</g->
   </ozn>
  </b+etr>
 <ann>; vsi pravijo, da je zelo dobra pianistka; </ann></etr>
</e+>
 </s2>
```

Example 2: the annotated English-Slovenian database

3. The final database

The final database consists of 138,369 possible candidates for Slovenian headwords in the future Slovenian-English dictionary, together with their translations and dictionary examples, grouped into four distinct categories:

1. **group "one to one"**: in this case the new Slovenian headword appears as a one-to-one translation, separated either by a comma or a semicolon, of the new English candidate for the translation equivalent; the corresponding examples from the entire dictionary database where the one-to-one translation appears in the English part of the example – either in its base or inflected form – are grouped under each equivalent

2. **group "one to multi-word + base form"**: in this case the new Slovenian headword appears as a part of the multi-word Slovenian translation equivalent in the English-Slovenian dictionary and is used in its base form
3. **group "one to multi-word + inflected form"**: in this case the new Slovenian headword appears as a part of the multi-word Slovenian translation equivalent in the English-Slovenian dictionary and is used in one of its inflected forms
4. **group "no translation"**: the new Slovenian headword is used in the Slovenian part of the example but none of the one-to-one or one-to-multi-word English translation equivalents is used in the English part of the example.

The last group is seen as particularly useful since it exposes contrastively interesting cases where in the English-Slovenian dictionary lexicographers had to find a solution that did not include the most common translation equivalents for a particular headword. Below we present a minuscule example from the database in the XML format:

```
<geslo id="040502"><slo>krčne žile</slo>
        <status><num>1</num><num>0</num><num>0</num><num>0</num></status>
        <nadskupina tip="neposredni"><skupina><prevod>varicose veins</prevod>
                        <enota tip="0-11"><tr><f>krčne žile</f>, varice</tr>
                        <info>0=ce id="93261" hw="varicose veins" ps="mn. sam." la="med."</info>
                        </enota>
        </skupina></nadskupina>
</geslo>
```

Example 3: the reversed Slovenian-English database

XML format was chosen as the default database format because of its standard and widespread use. However, an XSL file has also been created to enable lexicographers to visualize the data in a user-friendly form. The final result is the following:

Picture 1: Slovenian-English database

According to the system described above, in the source English-Slovenian dictionary database the Slovenian translation *žezlo* appears three times as the immediate translation equivalent of the English headwords: "mace", "rod" and "sceptre". Additionally, in the entire English-Slovenian database it appears in two usage examples as the translation of "sceptre" and these two are listed accordingly under the new English translation "sceptre". The second category encompasses the cases where *žezlo* is found as the immediate translation equivalent in its base form in a multi-word expression, and there are two such cases: the noun "bauble" and the verb "to sceptre" where a two-word translation had to be used to give the appropriate verbal meaning. The third category renders translations where one of the listed immediate translation equivalents is used in a multi-word expression but not in its base form. There is one such case in the dictionary – the adjective "sceptered" where *žezlo* is used in the instrumental case. The last category provides the information about the English headwords where none of the detected immediate translations was used as the translation of *žezlo* in usage examples. Under the original English-Slovenian entry "attribute", the American spelling "sceptre" was used and thus the procedure did not detect the equivalence between *žezlo* and "sceptre". Had we also applied variant spellings the usage example would be listed in the first category under the translation "sceptre". The conversion procedure, once established, was fully automated and did not involve any manual work.

4. The "funny mirror" from the lexicographer's perspective

The key purpose of the reversed database is to give the compiler of the new Slovenian-English dictionary excellent accessibility to the existent potential translation equivalents, therefore the organizational principles of the material are of paramount importance. In the compiling process, the compiler must be able to locate the most adequate translations as early on as possible, especially given the vast amount of (secondary) data s/he is faced with. The reversed dictionary database is one of the tools to be used in compiling Slovenian-English entries. The distorted image of Slovenian as determined by the English side has to be adjusted by sources and tools for the analysis of Slovenian. To that end, this framework will be combined with the FidaPLUS Reference Corpus of Slovenian (*http://www.fidaplus.net*)[2] and the Word Sketch Engine (*http://www.sketchengine.co.uk*)[3], a corpus tool that analyses a word's grammatical and collocational behaviour.

The reversed database can only be used to the benefit of a reversed dictionary if the user acknowledges a simple truth about the dynamics of the translation process: *L1 (Source Language) → L2 (Target Language) does not equal L2 (Source Language) → L1 (Target Language)*. There can be no doubt that the abundance of contrastively relevant data contained in a reversed dictionary database is of enormous help in compiling a reversed dictionary. However, there are some specific issues that have to be taken into account if we are to avoid falling into traps set continually by the reversed perspective. Typically, the L1 content will be a self-contained semantic unit and, ideally, rendered into L2 with an equally natural and/or lexically frozen semantic unit. However, in many cases the levels of this naturalness and fixedness differ, sometimes considerably. The key problem is not so much that of semantic equivalence but rather that of equivalence in terms of the typicality/frequency. Distortions of the image of Slovenian occur on the morphosyntactic as well as on the semantic levels of lexicographic description. Below are some examples in various categories:

A) Single-word lexical units (as entry headwords):

gledalec = **(on)looker, spectator, (tele)viewer, watcher, bystander, beholder (formal or dated), filmgoer (filmski gledalec), ringsider (gledalec v prvi vrsti), standee (gledalec na stojišču).**

Corpus analysis shows that of the above potential translations only two candidates, namely *spectator* and *(tele)viewer*, correspond to *gledalec* in isolation, whereas the others correspond to translations of certain multi-word units, e.g., *filmski gledalec = filmgoer*, *gledalec na stojišču = standee*. Nevertheless, they are all valid translations for the purposes of decoding.

[2] Described in Arhar and Gorjanc (2007) and Arhar et al. (2007).

[3] Described in Krek and Kilgarriff (2006), Kilgarriff et al. (2004).

B) Multi-word lexical units (as entry headwords):

bombastičen govorec, bombastična govorka = **tub-thumper (informal)**

While acceptable as a kind of explanatory equivalent, 'bombastičen govorec', let alone 'bombastična govorka' (feminine), is inappropriate as an entry headword given that it has no reality in the FidaPLUS corpus. *Bombastičen* (=bombastic) is relatively frequent in the corpus (851 hits) collocating, in descending order, with *naslov* (=headline), *izjava* (=statement), *napoved* (=forecast), *novica* (=news), *članek* (=article), *začetek* (=start, beginning), etc. Amongst these collocates, there is no near synonym for *govorec* (=speaker, 3963 hits), which in turn typically collocates with *dober* (=good), *rojen* (=born), *slavnosten* (=ceremony, in attributive use), *uradni* (=official). Perhaps the closest to the original, and yet still too distant from it, are near synonyms *strasten/ognjevit/goreč* (=passionate). So the compiler is left to chose between an explanatory equivalent of the type *govorec, ki* /…/ (=a speaker who /…/) and a compromise solution, an artificially construed collocation – the above 'explanatory translation'.

bombardiranje medijev = **media blitz**

Apart from being quite atypical in Slovenian, the collocation *bombardiranje medijev* is also ambiguous. *Mediji* in this position can function as a subject or a direct object. "Media blitz", on the other hand, displays a high degree of semantic opacity and syntactic fixedness, to the point where it could be labelled as journalistic. Below are some more examples of disproportionately occurring, but nonetheless legitimate translations pairs: *tat grobov*, *tatica grobov* (feminine) = body-snatcher.

C) (In)sufficiently contextualised lexical strings (as illustrative examples):

The context in the English-Slovenian dictionary (D1) is determined by the English perspective. Therefore, in the reversed dictionary it needs to be adequately suited to the needs of the Slovenian situation: *bogato obdarjena blondinka* = a pneumatic blonde; the most typical and semantically accurate collocate would be *prsata blondinka* (=a busty blonde), but the chosen translation in D1 is quite acceptable for a passive user.

D) Idiomatic lexical strings (phraseology):

On the whole, ready-made translations are often available for the Slovenian-English dictionary (D2): *denar je vir vsega zla* = money is the root of all evil; *roka roko umije* = you scratch my back and I'll scratch yours; čas kislih kumaric = silly season (informal Br). Nevertheless, the translations in D1 can be too generic or explanatory, e.g., *vse je čisto tako, kot mora biti* = everything is in apple pie order, or sometimes too literal to be identified as the source idioms for D2, e.g., *spustiti duha iz steklenice* = to let the genie out of the bottle (variants: *duh je ušel iz steklenice* = the genie is out of the bottle; *spraviti duha nazaj v steklenico* = to put the genie back into the bottle). The corpus of Slovenian shows marginal usage of the idiom *duh iz steklenice*, conditionally collocating with *spustiti*, but with the derived variants as listed above it is virtually non-existent. The translation used in D1, however, is acceptable and legitimate insofar as it conveys the most precisely the content of the English idiom.

5. Conclusions

A complex technical procedure was applied to the English-Slovenian dictionary database to give the compilers of the future Slovenian-English dictionary access to the abundance of information about possible Slovenian-English translation equivalents contained in the database. The procedure aimed at maximizing the information potential of the detailed XML structure and rendering the final result in a user-friendly manner. However, the conversion process reveals a distorted image of the source language and lexicographers have to analyse other monolingual sources for Slovenian before using the information from the converted Slovenian-English database. Only after such analysis can this information be properly used to divide the semantic space of the source language items as a function of the target language.

References

Arhar, Š.; Gorjanc, V.; Krek, S. (2007). "FidaPLUS corpus of Slovenian: the new generation of the Slovenian reference corpus: its design and tools". In Davies, M. (ed.). *Proceedings of the Corpus Linguistics Conference, CL2007, University of Birmingham, UK, 27-30 July 2007*. Birmingham.

Arhar, Š.; Gorjanc, V. (2007). "Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa". *Jezik in slovstvo* 52 (2). 95-110.

Grabnar, K.; Šorli, M. (2003). "Novi veliki angleško-slovenski slovar Oxford-DZS". *Jezik in slovstvo* 48 (3/4). 126-133.

Hočevar, M. M. (2006). "Veliki angleško-slovenski slovar OXFORD-DZS dokončan". *Mostovi* 40 (1/2). 180-187.

Kilgarriff, A.; Rychly, P.; Smrž, P.; Tugwell, D. (2004). "The Sketch Engine". *Proceedings of the XI Euralex Conference, Lorient, France.* 105-116.

Krek, S. (ed.). (2005-6). *Veliki angleško-slovenski slovar Oxford*. Ljubljana: DZS.

Krek, S., Kilgarriff, A. (2006). "Slovene word sketches". In Erjavec, T.M; Žganec Gros, J. (eds.). *Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 62-67.

Šorli, M. et al. (2006). "The Oxford-DZS Comprehensive English-Slovenian Dictionary". In Corino, E.; Marello, C.; Onesti, C. (eds.). *Atti del XII congresso internazionale di lessicografia: proceedings XII EURALEX international congress, Torino, Italia September 6th-9th, 2006*. Torino: Edizioni dell'Orso: Università di Torino: Academia della Crusca. 631-637.