A Workflow for Supplementing a Latvian-English Dictionary with Data from Parallel Corpora and a Reversed English-Latvian Dictionary

Daiga Deksne[1], Andrejs Veisbergs[2]
[1]Tilde, [2]University of Latvia
E-mail: daiga.deksne@tilde.lv, andrejs.veisbergs@lu.lv

Abstract

The lexicon of contemporary languages is changing rapidly, mostly by acquiring new loans and derivations. The change in lexicon is best reflected in the corpora of contemporary languages. Nowadays many collections of parallel-aligned texts are available electronically. To satisfy user needs for a modern, complete, up-to-date dictionary, we created a workflow for enriching the existing Latvian-English dictionary with data from parallel corpora containing lexis commonly used in contemporary language, as well as data from the reversed English-Latvian dictionary. While revising the existing Latvian-English dictionary, we identified some issues, for example, missing feminine forms of the nouns naming nationalities and occupations, representation of the words with optional parts or spelling variations. The task of dictionary improvement was done semi-automatically by the joint work of a lexicographer, computational linguists and programmers. Such natural language processing tools as a tokenizer, part-of-speech tagger, lemmatizer and spell-checker were used to reduce the manual work. As a result, the number of entries has increased by 32%, and the number of translations by 28%.

Keywords: electronic dictionaries, parallel corpora, NLP tools, XML format

1 Introduction

Electronic dictionaries are undergoing a huge expansion, both as concerns their production as well as their use. Though relatively new they also have to be updated regularly (Lorentzen & Trap-Jensen 2016) as the lexicon of contemporary languages is in a constant flux, with new items (especially loans and derivations) proliferating. Updating and expanding of dictionaries is a laborious and time-consuming process. However, in contrast to printed dictionaries, production of which also presumes considerable time for proofreading and printing, electronic dictionaries can be edited, supplemented and corrected promptly.

Moreover, electronic dictionaries are much less affected by space limitations (and mostly by screen size with regard to this issue). This affects some macrostructure issues, e.g. while regular derivatives (participles, verbal derivatives with prefix *non-*, agent nouns, occupation and nationality nouns in feminine and other categories) are generally avoided in printed Latvian dictionaries, these entries can be introduced in electronic one.

In order to create an electronic bilingual dictionary that corresponds to the users' current needs we created a workflow for merging three different data sources: an electronic version of the largest Latvian-English dictionary (Veisbergs 2016), the automatically reversed English-Latvian dictionary, and new entries from aligned bilingual parallel corpora. The dictionary in question is a large, general one, aimed at a relatively advanced Latvian user of English. It is mono-directional (aimed at the Latvian user) with no explanations for the Latvian part, while explanations for the English part are provided

where possible: labels, register, nuance markers, and bracketed semantic explanations. The entry structure consists of meanings, examples, and collocations subdivided by meaning, while the idiom block comes at the end of the entry.

There are different views as to the results (Newmark 1998; Geisler 2002; Veisbergs 2004; Krek, 2008) and efficiency (Geisler 1999; Tamm 2002; Veldi 2010) of bilingual dictionary reversing. Some studies and experiments are positive, others point to too much "noise" and extensive editing and proofreading that is too laborious.

The team has experience in creation of electronic dictionaries and dictionary-browsing environments on different platforms, enabling users to search in several dictionaries simultaneously. Integration of spell-checking and morphological analysis allows looking for inflected forms or finding the translation even for misspelled words. A uniform XML format for dictionary entry representation has been developed, and all dictionary resources are parsed and stored internally using this format (Deksne et al. 2013).

2 The Drawbacks of the Existing Latvian-English Electronic Dictionary

There were some drawbacks in the existing dictionary-browsing environment concerning representation of entries, comprehension of dictionary content by the user, and missing content.

The existing dictionary-browsing environment allowed searching for entries in several dictionaries at once; the results from different sources were presented to the user in a sequential order; translations or examples (as in Figure 1) in the direct and in the reversed dictionary tend to overlap; a user is obliged to scroll through a long list of identical results. Sometimes one source contained a hyphenated form of a compound, while another contained a non-hyphenated form; for example, there were translations 'tom-cat' and 'tomcat' in the different data sources for the same headword. There were around 11,000 entries with the same headword in the existing Latvian-English dictionary and in the reversed English-Latvian dictionary, making users wonder where the differences lay. They were thus merged in the new version of the Latvian-English dictionary.

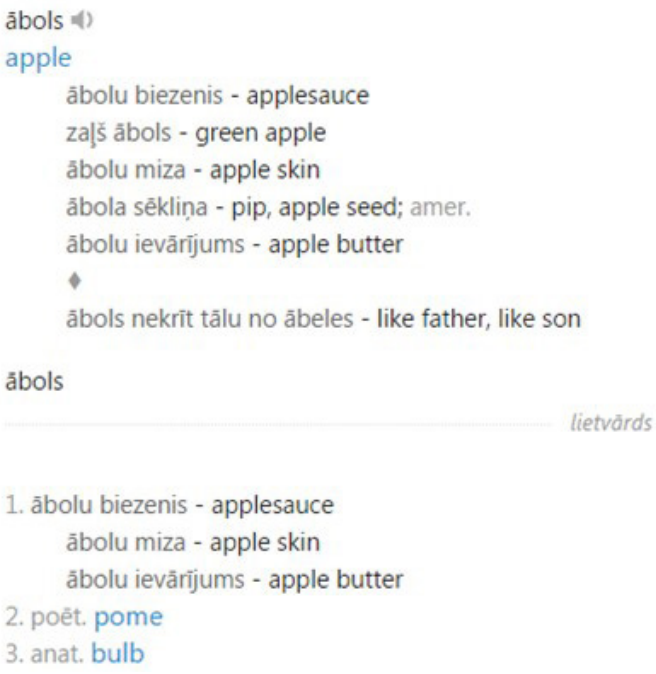

Figure 1: Entries from two different data sources in the dictionary-browsing environment.

- Latvian bilingual dictionaries traditionally do not contain words created by regular derivation rules, e.g. participles used as adjectives, negative forms of verbs or adjectives, feminine forms of nouns naming occupation or nationality. Frequently headwords having masculine and feminine forms are given with an additional ending, as in printed dictionaries. Examples (1) and (2) show the masculine forms with additional feminine endings for the nationalities 'Swede' and 'German'. Example (3) represents the masculine and feminine forms of the occupation name 'telephone operator'. As the root of the word in the dictionary is not marked, the feminine form cannot be expanded automatically. Some words were given with a spelling variant in parentheses (see (4), Eng. 'activation'). Such a format did not allow the user to find both forms of the word using a search engine in the dictionary-browsing environment.
  (1) zviedrs/iete
  (2) vācietis/e
  (3) telefonists/e
  (4) aktiv(iz)ēšana
- There were examples containing variations of one or several words separated by slash '/' symbol. Example (5) shows a phrase which expands to four different phrases (6), (7), (8), and (9). This format does not take an extra space but is not suitable for search, and is hard to comprehend for a person who does not know the foreign language very well.
  (5) to book/engage a season ticket/pass
  (6) to book a season ticket
  (7) to book a season pass
  (8) to engage a season ticket
  (9) to engage a season pass
- Another discrepancy was the representation of Latvian words translated in English as adjectives with an attributive meaning.
  (10) **pilsēta** town borough; (liela) city
  (11) **pilsētas** urban; municipal; town (attr.); towny
  (12) olveida
- There were about fifteen hundred such entries. The Latvian word is a noun in the genitive case with or without an ending. Sometimes in Latvian dictionaries a hyphen character is used to depict the genitive. Some compounds are used only in a genitive (see (11), Eng. 'egg-shaped') but for the most nouns, the base form is nominative. Using the genitive case of the noun for the main entry without additional information may often be confusing, as the genitive case may coincide with the plural nominative or accusative (for example, the bold formatted headword in the entry (11) is a single genitive form of the headword in (10), but the same form is also plural nominative or accusative). A label *gen*. was added to solve this homographic issue.
- Besides, as new words proliferate any dictionary is lagging behind. There are numerous words which have appeared in English in the last few decades, such as 'geocaching', 'flash mob' or 'raw-foodist'. Their Latvian counterparts and corresponding English equivalents were also added to the dictionary.

3 Compilation of Lists Containing Translation Hypotheses

A parallel corpus is a valuable resource when looking for new entries for dictionary supplementing. We compiled a corpus for possible translation extraction from several sources. The first part is a proprietary collection of parallel data used in different projects. The second part is formed by some components of an open source parallel corpus OPUS – a collection of translated texts from the web

(Tiedemann 2012). We use a collection of EU Translation Memories, documentation from the European Central Bank, documents from the European Medicines Agency, proceedings of the European Parliament and some other collections.

The Latvian text (as in (14)) was part-of-speech tagged and lemmatized (as in (15)) using NLP tools created by company *Tilde* (Deksne 2013; Pinnis & Goba 2011) while the English text (as in (13)) was left unchanged. Such a parallel-aligned corpus was passed to the next step of processing.

(13) account creation guide
(14) konta izveides norādījumi
(15) kontsN izveideN norādījumsN

The Moses toolkit used for statistical machine translation (Koehn et al. 2007) was employed for building the phrase tables. Each line in a phrase table contains a pair of Latvian and English word/phrase. These are hypothetical translations which have been obtained automatically using statistical methods. The pairs occurring only once and the stop words were filtered out. We made a list of word and translation pairs which were already present in the existing Latvian-English dictionary, and these were filtered out from the phrase table too. The rest of the lines were grouped by part-of-speech of the Latvian word. As a result of this process, we acquired lists of nouns, verbs, adverbs, adjectives, and interjections with hypothetical translations (see Table 1). Among the nouns and adjectives, there were many deverbalized derivations and compounds. Among the verbs, there were many negative verbs as well as the prefixed verbs (in the Latvian language, verb prefixation process is very productive). There are 11 prefixes used in verb formation. The verbal prefix can formally change some features of an aspect of the verb; it can also modify or change the lexical meaning of the base verb (Holvoet 2001).

Table 1: The number of entries and their hypothetical translations extracted from parallel corpora.

Word class	Number of entries	Number of hypothetical translations
noun	8,609	41,242
verb	4,928	29,617
adverb	1,483	7,847
adjective	1,025	4,247
interjection	64	523

We prepared several files in a simple tab separated format containing verbs, nouns, adjectives, adverbs and interjections. The second column contained the lemma of the word in Latvian, the third column contained the possible English translation, and the fourth column contained the frequency of the word pair in the corpus. The first column was reserved for the lexicographer's marking of possible inclusion in the dictionary (see the example in Figure 2). The lexicographer was asked to put a meaning number in the first column of the line valid for inclusion in a dictionary. The lexicographer was also instructed to manually add a comment in parentheses or some usage samples at the end of the line if necessary.

1	punktots	dotted	75
1	punktots	spotted	3
	punktots	background	2
	punktots	dotted lines	2
	punktots	green	2
1	punktots	polka-dot	2
	punktots	text	2

Figure 2: The adjective *punktots* with statistically acquired hypothetical translations from the corpus.

A total of 5,995 pairs or 7.18% of hypotheses (4,082 unique Latvian words, i.e., dictionary entries) were marked as suitable for inclusion in a dictionary.

4 Process Workflow

The workflow for enriching a Latvian-English dictionary consisted of several steps. Some tasks where automatized, and some involved manual work of a computational linguist or a lexicographer. The lexicographer regularly updates the dictionary in an MS Word document using rich formatting, e.g., a font style for entry title must be bold, a font style for comments or usage and grammatical information must be italic, specialized meanings have different punctuation symbols: commas, semicolons, colons, dashes. The computational linguist created scripts for transforming the content of the dictionary from the MS Word format (see Figure 4) to the proprietary dictionary XML format (see Figure 3). Specific XML tags allow marking of all parts of an entry. Special tags are reserved for headwords, translations, samples, sample translations, idioms, meaning numbers, usage information, comments, grammatical information, and punctuation symbols. It is important to scrupulously comply with formatting rules in the MS Word document, as errors in formatting can invoke errors in XML representation.

The next step was to merge the XML representation of the Latvian-English dictionary with the data from parallel corpora and from a reversed dictionary. Data from TAB separated lists of translation hypotheses from parallel corpora was converted to the XML format. A special color attribute was appended to the title and the translation tags to mark this entry as coming from a different source. As the structure of TAB separated lists is very simple, this step was easy to implement. The new entries where appended to the XML representation of the Latvian-English dictionary and all entries were sorted alphabetically.

```
<entry title="aisbergs">
<title>aisbergs</title>
<transl>iceberg</transl>
<idiom />
<from_sample>aisberga redzamā daļa</from_sample>
<to_sample>tip of the iceberg</to_sample>
</entry>
```

Figure 3: Sample xml entry for a Latvian word *aisbergs* (Eng. ‘iceberg’).

The reversed dictionary was first filtered by removing translations and usage examples which were already present in the Latvian-English dictionary. The filtered version of the reversed dictionary was then merged with the main dictionary by including a whole entry if an entry with such title word did not exist, or by adding the translations and the usage samples at the end of the existing entry. A different color attribute was appended to the title and the translation tags to mark this entry as coming from the reversed dictionary. A new MS Word document was generated from the internal XML format. The information coming from the XML tags with a color attribute was reflected in the MS Word document (see Figure 4). The prepared document was then passed to a lexicographer for post-editing. In such a format the lexicographer could distinguish the parts of the dictionary coming from different sources and make the necessary corrections, such as, for example, reordering the translations or grouping them in a separate meaning.

aizsargiepakojums protective bag, protective packaging
aizsargierīce safety device; protective equipment; *(ieroču)* safety-bolt; *(uz dzelzceļa)* safeguard; protection device; *tehn.* protector
aizsargjosla 1. *mil.* defence zone; **2.** *bot.* protective zone; meža a. – forest shelter belt, green belt
aizsargkārta coating, protective layer
aizsargkonstrukcija protection structure
aizsargkrāsa protective colouring; *mil.* camouflage colours; *poligr.* safety ink
aizsargkrēms barrier cream
aizsargķivere helmet, hardhat *amer.*

Figure 4: MS Word document with automatically merged entries.

5 Results

Editing the data extracted from parallel corpora involved the deletion of numerous entries that were grammatically erroneous, e.g. under adverbs many nouns in plural appeared, as both categories contained the ending *i.* There were also numerous gerund/participle entries that had fully predictable standard forms in both Latvian and English, which were considered not worth keeping. Likewise, Latvian verbs with a standard negative prefix which in translation would normally be equivalent to the English verb plus particle *not* were not included.

As a result, the new entries are mostly derivatives, chosen from the long list on the basis of two criteria – frequency and irregularity of English counterparts (that a Latvian user might not be able to surmise). There was also a considerable number of "missing" entries that for some reason had not been in the old dictionary. The combined version also yielded some "double entries" – either the result of wrong spelling of the Latvian word or in some cases parallel spellings. The latter were then joined in full form to the main entry. Very specialized terms or rare and obsolete words were avoided. Apart from the Latvian headwords, there was a huge number of additional English equivalents which were added to the English part, distributed among the senses or examples or added to the idiom block.

We also added gender differentiation in Latvian entries, thus having double Latvian entries for those English equivalents where gender does not differ (as in (16) and (17)), and separate entries for those where English has gender marked lexical units (the masculine forms as in (19), (20) and (23); the feminine forms as in (18), (21) and (22)).

(16) kinoapmeklētājs, kinoapmeklētāja filmgoer, moviegoer amer.
(17) klasesbiedrs, klasesbiedre classmate
(18) dzejniece poetess; poet
(19) dzejnieks poet
(20) kinoaktieris film actor, screen actor
(21) kinoaktrise film actress, screen actress
(22) cariene tsarina, tzarina, czarina
(23) cars tsar, czar.

Apart from these, some common abbreviations as well as some proper names were included in the entry list. While the printed dictionary traditionally had a separate appendix for geographical names, the electronic version does not differentiate between such categories. The author/lexicographer has concluded that in future the printed version will also drop the appendix and provide the geographical names in the single entry list. This seems to be a more reader-friendly approach.

6 Collateral Ideas and Solutions

Though most of the supplementation focused on derivatives, new meanings and extra equivalents, some issues of collocations and idioms were also brought into focus. It is well known that dictionaries often tend to compartmentalize the information by linguistic categories. This is partly inevitable as a result of the essence of dictionary – providing isolated, generalized material, not contextual use (the latter being almost indescribable in its complexity). Yet compartmentalization tends to be affected by theoretical linguistic categories, which is a somewhat scholastic exercise, trying to draw precise borders for concepts such as idiom, collocation, compound, and word. This leads to "a tendency to circumscribe the research field for purposes of consolidation" (Burger 2007: 11), while corpus linguistics produces the opposite. While theoretically usually correct, these divides are often artificial, as the strictly defined linguistic concepts do not reflect the fuzzy, blurred and scalar reality of the language, especially in a multilingual setting. Clutching at the mandatory correspondence of categories (idiom for idiom, collocation for collocation, word for word) in dictionaries is not sensible, but is often practiced and also emphasized in research. Bo Svensén plainly states "idioms in the source language must as far as possible be paralleled in the target language by idioms with the same content" (Svensen 1993: 156), and thus presented idiom for idiom. However, this is not always possible, nor should it be mandatory: language structures are different, so are ideas about some linguistic concepts, like idiom and compound.

The reversal exercise showed that some idioms in the English-Latvian dictionaries were translated as words and some words as idioms, and these were full equivalents. Should we avoid the reverse – giving some idioms as an equivalent for some Latvian words, and some words as an equivalent for some idioms – just because there is a category divide?

Sometimes an idiom was the only adequate equivalent for a word, e.g.

(24) sastiķēt [to together stick] to chip in;
(25) apmuļķot [to around fool] to make* a fool (of), to fool, to take* for a ride;
(26) appuišot [to around boy] to fetch and carry, to wait on hand and foot.

In its turn, a Latvian idiom might have an English collocation or lexical counterpart that would be a better semantic match than an idiom with an analogous image, e.g.

(27) domu grauds [thought grain] aphorism, maxim;
(28) ziedu laiki [blossom days] heyday, highday, palmy days, prime, zenith;
(29) sarkanais gailis [the red cock] fire; ielaist sarkano gaili [to let the red cock in] – to set* fire (to).

Sometimes an idiom would have several equivalents: words, phrases, idioms:

(30) tukši vārdi [empty words] mere words, wind, hot air, lip service.

Finally, a simple entry that clearly illustrates the structural and semantic shifts between languages:

(31) galarezultāts [end-result] outcome, the end result; ◊ galarezultātā [in the end-result] – at the end of the day; in the end.

The Latvian compound, corresponding to an English compound or collocation, deserves an English idiom when used in a declined form.

This presented a more flexible approach to the idiom-word divide, tearing down the conventional barriers of lexicographical thinking and practice. We should think more in terms of equivalence of meanings, not structures, words or phrases (Atkins & Rundell 2008). We believe that dictionaries

should be "much more phrasal than they currently are" (Granger 2008: 1353), as it is well known that "idiomaticity facilitates communication" (Bejoint 2000: 216).

7 Conclusions

The results of our work: the former Latvian-English dictionary contained 36,608 entries with 96,066 translations and 22,090 usage samples. The reversed dictionary contained 20,253 entries, a large part of entries partly or fully overlapped with the former Latvian-English dictionary. The new enlarged Latvian-English dictionary contains 53,867 entries, 132,481 translations and 22,277 usage samples including 4,082 new entries which were added after processing parallel aligned corpus. Despite protracted post-editing work, the accomplished end result is impressive. It not only considerably increased the number of entries, senses and equivalents, but also yielded interesting theoretical insights in the practical lexicography, like idiom treatment, genitive – attributive words, among others. One should also consider the benefits of a massive increase in the number if items for digital use and machine translation purposes.

The dictionary is available online at https://www.letonika.lv/groups/default.aspx?g=2&r=10621033&f=1.

8 References

Atkins, B.T.S. and Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: OUP.

Bejoint, H. (2000). *Modern Lexicography*. Oxford: OUP.

Burger, H. (2007). Semantic aspects of phrasemes. In D. Burger. D. Dobrovol'skij, P. Kuehn, N.R. Norrick (eds.) *Phraseologie. Vol. 1.* Berlin, New York: Walter de Gruyter, pp. 90-109.

Deksne, D., Skadina, I., & Vasiljevs, A. (2013). The modern electronic dictionary that always provides an answer. In *Electronic lexicography in the 21st century: thinking outside the paper: proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. pp. 421-434.

Geisler, C. (2002). Reversing a Swedish-English dictionary for the Internet. In L. Borin (ed.) *Language and Computers, Parallel Corpora, Parallel Worlds*. Amsterdam: Rodopi. pp. 123-133.

Granger, S. & Paquot, M. (2008). "From dictionary to phrasebook?" In E. Bernal, J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada, pp. 1345-1355.

Holvoet, A. (2001). *Studies in the Latvian Verb*. Kraków: Wydawnictwo universitetu Jagiellońskiego.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. and Dyer, C., (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, pp. 177-180.

Krek, S.; Šorli, M.; Kocjancic, 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 E. Bernal, J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, pp. 535-542.

Lorentzen, H. and Trap-Jensen, L. (2016). What, When and How? – the Art of Updating an Online Dictionary. In T. Margalitadze, G. Meladze (eds.) *Proceedings of the 17th EURALEX International Congress*. Tbilisi: Ivane Javakhishvili Tbilisi University Press, pp. 138-145.

Newmark, L. (1998). Reversing a One-Way Bilingual Dictionary. In Th. Fontanelle et al. (eds.) *EURALEX 1998 Proceedings*. Liege: University of Liege.

Svensen, B. (1993). *Practical Lexicography*. Oxford, New York: OUP.

Tamm, A. (2002). Reversing the Dutch-Estonian Dictionary to Estonian-Dutch. In A. Braasch, C. Povlsen (eds.) *Proceedings of the Tenth EURALEX International Congress. Vol. 1,* Copenhagen: CST, pp. 389-399.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *LREC Vol. 2012*, pp. 2214-2218.

Veisbergs, A. (2004). Reversal as Means of Building a New Dictionary. In G. Williams, S. Vessier (eds) *Proceedings of the Eleventh EURALEX International Congress*. Lorient: UBS. Vol. I, pp. 327-332.
Veisbergs, A. (2016). *The New Latvian English Dictionary*. Riga: Zvaigzne ABC.
Veldi, E. (2010). Reversing a Bilingual Dictionary: a mixed blessing? In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress*. 6-10 July 2010. Leeuwarden/Ljouwert: Fryske Akademy – Afûk. pp. 861-865.

Acknowledgements

The research has been supported by the European Regional Development Fund within the project "Neural Network Modelling for Inflected Natural Languages" No. 1.1.1.1/16/A/215.