Straddling the Boundaries of Traditional and Corpus-Based Lexicography: A Latvian-Czech Dictionary

Michal Škrabal
Institute of the Czech National Corpus, Charles University in Prague
e-mail: michal.skrabal@ff.cuni.cz

Abstract

The aim of the presentation is to introduce an ongoing project of a Latvian-Czech dictionary, the first bilingual lexicographical description for this combination of languages. My quite logical tendency to employ corpus methods faces, however, a significant problem – a limited data base. Thus, I necessarily find myself straddling the boundaries between the traditional, "card-based", and "new", corpus-based, way of lexicographical work. However, this should not be seen as a disadvantage – the dictionary is being built on the solid foundations provided by traditional, time-proven lexicographical practise. Nevertheless, whenever possible, I try to employ modern-age tools and methods (corpora, internet, software, etc.) which save a substantial amount of work and time.

Several sample entries on the poster demonstrate the sought-after priorities of the dictionary. These are – at least in the field of Czech-Baltic lexicography – innovative features: a modern, user-friendly approach; a rich description of collocations; descriptiveness rather than prescriptiveness; the "emancipation" of certain lemmas (often excluded from the main list of entries, taboo-words, or words traditionally nested – e.g. feminine derivates, etc.); usage notes. If the existing, albeit prevailing, lexicographical description is inconsistent with the language reality, it should be – contrary to the tradition – discarded and replaced with another description.

Keywords: Latvian-Czech dictionary; traditional vs. corpus lexicography; bilingual lexicography; limited data base

The forthcoming Latvian-Czech dictionary, the first bilingual lexicographical description for this combination of languages (cf. Czech-Latvian dictionaries published in 1988 and 2006), should represent the most versatile tool possible aiding in the translation of a wide variety of texts, including e.g. classic literary works, contemporary journalistic or scientific texts. The goal is not to fill the current gap with a hastily-created, truncated compilation of other dictionaries that will only meet the basic needs of users. The dictionary should instead be based on real contemporary Latvian language usage, as is represented by corpus data.

The dictionary primarily aims to be a decoding one and is above all designed for Czech users, though allowing for a partial extension to Latvian users as well (it includes basic grammar information for Czech equivalents). Both the macro- and microstructure are primarily of a traditional sort (with exceptions mentioned below), including a conventional description of Latvian grammar and sample paradigms that will be referred to in the entries. The dictionary aspires to be medium-sized, with approximately 45,000 headwords.

My quite logical tendency to employ corpus methods (as a matter of fact, is it possible to foster lexicography today in any other way?) faces, however, a significant problem: a limited data base. The representative Latvian corpus *Līdzsvarots mūsdienu latviešu valodas tekstu korpuss 2013* (LVK2013)

has only 5.5 million tokens, I also utilise the Czech-Latvian part of the parallel corpus InterCorp,[1] web corpora[2] which are, however, rather problematic as regards their representativeness and quality of data. Thus, I necessarily find myself straddling the boundaries between the traditional, "card-based", and "new", corpus-based, way of lexicographical work. Figuratively speaking: the forthcoming Latvian-Czech dictionary will be a kind of a "bridge" between these two approaches to lexicography. However, this should not be seen as a disadvantage – the dictionary is being built on the solid foundations provided by traditional, time-proven lexicographical practise. After all, tradition plays a significant role in lexicography and it is not easy to cast it aside.[3] Nevertheless, whenever possible, I try to employ modern-age tools and methods (corpora, internet, software, etc.) which save a substantial amount of work and time.

These, together with traditional resources – namely monolingual, bilingual and specialized dictionaries (both printed and electronic ones) – and, additionally, excerpts, comprise a varied range of sources that ought to be mutually dealt with, especially in terms of the commonness and frequency of each individual lemma. Obsolete, generally peripheral lexemes are often artificially kept alive by many a dictionary. I want to avoid this. My aim is to describe, as precisely and faithfully as possible, the *contemporary* Latvian lexicon, both formal and informal, dating from the time since Latvia regained its independence in 1991, with an increasing number of English loanwords but all the while preserving older lexical layers (e.g. lexis from the Soviet era, which is still alive in contemporary Latvian).

Several sample entries on the poster demonstrate the sought-after priorities of the dictionary. These are – at least in the field of Czech-Baltic lexicography – innovative features: a modern, user-friendly approach; a rich description of collocations; descriptiveness rather than prescriptiveness; the

[1] In its current version (8.0 from June 2015), it contains more than 32 million words: the original, manually aligned core of works of fiction (currently 1,336,000 words) has been expanded by several collections of automatically processed texts: namely EU legal texts from the *Acquis Communautaire* corpus (18,744,000 words), minutes of the EP meetings in 2007–2011 from *Europarl* (11,688,000 words), and movie subtitles from the *OpenSubtitles* database (280,000 words). For general information about InterCorp, see Čermák & Rosen 2012.

[2] Corpus *Latviešu valodas tīmekļa korpuss* (LVTK), compiled from Latvian web pages, contains 122,628,720 words, but unfortunately it is not lemmatized and is only partially tagged. Its quality is further reduced – in some cases beyond eligibility – by a significant proportion of texts being without diacritics, numerous foreign fragments, boilerplate, and other frequent defects (divided words or contrary ligatures, improper formatting, etc.). Despite the proclamation that the duplications in LVTK had been eliminated, identical or minimally differing sentences appear too often, which greatly distorts the frequency statistics.

The largest web corpus of Latvian currently available – *lvTenten* (total size: 667,668,379 words) – was crawled in 2014 by SpiderLing tool (Suchomel & Pomikálek 2014) as a member of the TenTen corpora family (Jakubíček, et al. 2013). The initial version was converted into the UTF-8 format, cleaned and de-duplicated. The corpus remains provisionally untagged, but provides the Word Sketch feature not offered by other corpora. I see this as its biggest asset, although my experience with it is rather small as of now, limited to just a few probes.

[3] "The current practice is largely a result of a slowly accreting tradition" (Gleason 1967: 90). Cf. L. Zgusta's dictum (1971: 191): "To understand what is the best in the tradition, to understand the general trend of future development, and to combine this knowledge in a dictionary which is well founded on facts whose roots are in history but whose future development has been foreseen and fostered, that is the real art of the accomplished lexicographer."

"emancipation" of certain lemmas (often excluded from the main list of entries, taboo-words, or words traditionally nested – e.g. feminine derivates, etc.); or usage notes. If the existing, albeit prevailing, lexicographical description is inconsistent with the language reality, it should be – contrary to the tradition – discarded and replaced with another description. In this way, I have tried to incorporate the results of one of my case studies (Škrabal 2016) into the poster, when the semantic division of lemma *biedrs* offered by the newest Latvian monolingual dictionary (MLVV) had to be rejected: 1. fellow, friend, colleague; 2. member; 3. comrade. On the basis of a manual analysis of corpus data (776 occurrences of the lemma in LVK2013), an overlooked sense[4] (yet, incidentally more frequent than the third one, historically-marked) was discovered; the rank of the first two senses was adjusted by frequency as well. The resulting semantic framework (1. member – 497 hits in LVK2013, i.e. 64 %; 2. fellow, friend, colleague – 204 hits, i.e. 26 % 3. deputy – 50 hits, i.e. 6 % 4. comrade – 25 hits, i.e. 3 %) was then enriched by numerous collocations, without any doubt the most numerous among all the Latvian translation dictionaries.

biedr|s m_1 (Vsg ~i!) **1.** (politické strany, organizace, klubu, komise ap.) *člen* • sociāldemokrātu partijas b. *člen sociální demokracie* • arodbiedrības b. *člen odborů, odborář* • kultūras komisijas b. *člen kulturní komise* • ~a karte *členská legitimace* • ~a nauda *členský poplatek* • ~u sapulce *členská schůze* • kļūt par goda ~u *stát se čestným členem* • biržas b. <ekon> *člen burzy* (cenných papírů) ‖ **2.** *kolega, druh, společník, kamarád* • darba b. *spolupracovník, kolega z práce* • skolas / klases / studiju b. *spolužák (kolega) ze školy / z třídy / ze studií* • dzīves b. *životní partner, druh* • komandas b. *spoluhráč, týmový kolega* • viņš ir lielisks sarunu b. *skvěle se s ním povídá, je to skvělý společník* • cīņu b. *spolubojovník, kamarád ve zbrani* • domu b. *ideový souputník* • ceļa b. *spolucestující, souputník* • istabas b. *spolubydlící* ‖ **3.** <polit> *zástupce, náměstek* • ministru prezidenta b. *místopředseda vlády, vicepremiér* • Saeimas priekšsēdētāja b. *místopředseda parlamentu* ‖ **4.** <sov> *soudruh* • b. Kalniņš *soudruh Kalniňš* • cienījamie ~i *vážení soudruzi* • ~u tiesa *soudružský soud*

Figure 1: Entry for *biedrs* in the forthcoming Latvian-Czech dictionary.

Similarly, two feminine derivates, *biedre* and *biedrene*, have been processed: they are no longer a mere "appendage" to the basic entry *biedrs* but form separate entries of their own. The manual analysis of corpus data (especially left-side collocates) also showed that they are not synonyms, as it might seem from the current description in the Latvian dictionaries. Their collocation profiles are quite different: most of the left-side collocates found in LVTK[5] are complementarily divided between

[4] This sense is not a new one, just an updated one from the inter-war period.

[5] The lexeme *biedre* (the accusative singular form *biedri*, which is homonymous with the accusative singular/nominative plural form of the lemma *biedrs*, is disregarded) can be found here minimally 528 times, the lexeme *biedrene* 261 times, cf. 31:21 ratio in LVK.

- left-side collocates typical for the lexeme *biedre* (superscript preceding the collocate refers to the relevant sense): [1]*apvienības* (10), [1]*asociācijas* (61), [1]*biedrības* (14), [1]*goda* (12), [1]*kolēģijas* (6), [1]*komisijas* (19), [1]*organizācijas* (8), [3]*priekšsēdētāja/priekšsēdētājas* (10), [1]*PSKP* (9), [1]*savienības* (59);
- for the lexeme *biedrene*: [2]*darba* (13, collocation *d. biedre* only 2), [2]*domu* (4), [2]*istabas/istabiņas* (20), [2]*klases* (22, collocation *k. biedre* only 2), [2]*komandas* (16, collocation *k. biedre* only 1), [2]*kursa* (25), [2]*skolas* (23, collocation *skolas biedre* only 1), [2]*sola* (4), [2]*studiju* (9);
- for both lexemes: [2]*ceļa* (*biedre* 4 / *biedrene* 5), [2]*cīņu* (5/2), [2]*dzīves* (19/13), [2+1]*grupas* (2/3), [1]*kluba* (19/5), [2]*sarunu/sarunas* (14/29).

these two forms, only a few being common for both of them. These findings significantly influenced the final form of both entries (which in most of Latvian dictionaries do not feature separately at all):

biedr|e f5 **1.** (politické strany, organizace, klubu, komise ap.) *členka* • Latvija ir šīs organizācijas b. kopš 2002. gada *Lotyšsko je členem této organizace od roku 2002* ‖ **2.** *kolegyně, kamarádka* • dzīves b. *životní partnerka, družka* ‖ **3.** <polit> *zástupkyně, náměstkyně* • ministru prezidenta b. *místopředsedkyně vlády, vicepremiérka* • Saeimas priekšsēdētāja b. *místopředsedkyně parlamentu* ‖ **4.** <sov> *soudružka*

biedre|ne f5 (Vsg ~n!, Gpl ~ņu) **1.** *kolegyně, kamarádka* • skolas b. *spolužačka* • istabas b. *spolubydlící* • darba b. *spolupracovnice, kolegyně z práce* ‖ **2.** <sov> *soudružka*

Figure 2: Entries for *biedre* and *biedrene* in the forthcoming Latvian-Czech dictionary.

Summing up the results of this and other case studies reveals that the final appearance of the sample entries differ from the solutions offered by other monolingual and bilingual dictionaries of Latvian. I attribute this finding to the fact that only in this single case has corpus data been systematically used or, more precisely, the initial description was compared with them and duly taken into account (cf. also a typology of Latvian dictionaries as regards methodology on the poster).

The technical aspects of language material processing play an important role in today's lexicography. Specialized lexicographical software (in my case TshwaneLex, based on XML markup language), providing sophisticated lexical databases, is indispensable nowadays. I deliberately give special attention to this aspect of the project: it is vital for the success of my work and, more importantly, even extends beyond it. Most of the data recorded in the database which will serve as a basis for the development of the Latvian-Czech dictionary should be reflected in the finished work. In addition, however, there is also supplementary information serving my internal needs that can be easily expanded by simple modifications of the DTD structure. The database thus potentially becomes the knowledge base of the Latvian language in general – and also the source, as such, for other linguistic, translatological, pedagogical or NLP projects. The possibility of separating the content of the lexical description from its form (which is fully adaptable to the individual needs of both the lexicographer and the target recipient) can be seen as the milestone of lexicographical theory and practice (cf. also Vondřička 2011: 11ff.).

References

Čermák, F. & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. In *International Journal of Corpus Linguistics* 13(3), pp. 411–427.

Gleason, R.I. (1967). The Relation of Lexicon and Grammar. In F.W. Householder, S. Saporta (eds.). *Problems in Lexicography.* Bloomington: Indiana University, pp. 85–102.

Jakubíček, M. et al. (2013). The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL 2013*. Lancaster: UCREL, pp. 125–127.

Ķelpe, M. et al. (1988). *Čehu-latviešu vārdnīca*. Riga: Avots.

MLVV: *Mūsdienu latviešu valodas vārdnīca.* Accessed at: http://tezaurs.lv/mlvv/ [28/04/2016].

Nikuļceva, S. (2006). *Česko-lotyšský slovník. Čehu-latviešu vārdnīca.* Prague: Leda.

Suchomel, V. & Pomikálek, J. (2012). Efficient Web Crawling for Large Text Corpora. In A. Kilgarriff, S. Sharoff (eds.) *Proceedings of the seventh Web as Corpus Workshop (WAC7)*. Lyon,

pp. 39–43.
Škrabal, M. (2016). Srovnávací aspekty lotyšského a českého lexikonu: Materiály k sestavení lotyšsko-českého slovníku. Comparative aspects of Latvian and Czech lexicons: Materials for assembling a Latvian-Czech dictionary. PhD thesis. Charles University, Prague, Czech Republic.
Vondřička, P. (2011). Formalized contrastive lexical description: a framework for bilingual dictionaries. Formalizovaný kontrastivní popis lexikálních jednotek: deskriptivní rámec pro dvojjazyčné slovníky. PhD thesis. Charles University, Prague, Czech Republic.[6]
Zgusta, L. (1971). *Manual of Lexicography.* Prague: Academia.

Corpora:

LVTK: *Latviešu valodas tīmekļa korpuss* [versija 1.0]. LU Matemātikas un informātikas institūts, Rīga. Accessed at: http://www.korpuss.lv [28/04/2016].
LVK2013: *Līdzsvarots mūsdienu latviešu valodas tekstu korpuss 2013.* LU Matemātikas un informātikas institūts, Rīga. Accessed at: http://www.korpuss.lv [28/04/2016].
Rosen, A. & Vavřín, M. (2015): *Korpus InterCorp, verze 8 z 4. 6. 2015*. Ústav Českého národního korpusu FF UK, Praha. Accessed at: http://www.korpus.cz [28/04/2016].

[6] Also available online at: http://wanthalf.saga.cz/library/lexdesc_thesis.pdf [28/04/2016].