From Monolingual to Bilingual Dictionary: The Case of Semi-automated Lexicography on the Example of Estonian–Finnish Dictionary

Margit Langemets[1], Indrek Hein[1], Tarja Heinonen[2], Kristina Koppel[1], Ülle Viks[1]

[1] Institute of the Estonian Language, Roosikrantsi 6, 10119 Tallinn, Estonia

[2] Institute for the Languages of Finland, Hakaniemenranta 6, 00530 Helsinki, Finland

E-mail: margit.langemets@eki.ee, indrek.hein@eki.ee, tarja.heinonen@kotus.fi, kristina.koppel@eki.ee, ylle.viks@eki.ee

Abstract

We describe the semi-automated compilation of the bilingual Estonian–Finnish online dictionary. The compilation process involves different stages: (a) reusing and combining data from two different databases: the monolingual general dictionary of Estonian and the opposite bilingual dictionary (Finnish–Estonian); (b) adjusting the compilation of the entries against time; (c) bilingualizing the monolingual dictionary; (d) deciding about the directionality of the dictionary; (e) searching ways for presenting typical/good L2 usage examples for both languages; (f) achieving the understanding about the necessity of linking of different lexical resources. The lexicographers' tasks are to edit the translation equivalent candidates (selecting and reordering) and to decide whether or not to translate the existing usage examples, i.e. is the translation justified for being too difficult for the user. The project started in 2016 and will last for 18 months due to the unpostponable date: the dictionary is meant to celebrate the 100th anniversary of both states (Finland in December 2017, Estonia in February 2018).

Keywords: bilingual lexicography; corpus linguistics; usage examples; GDEX; Estonian; Finnish

1. Background

The idea of compiling the new Estonian–Finnish dictionary is about 15 years old: it first arose in 2003, after the successful publication of the voluminous Finnish–Estonian dictionary, a project which benefitted from necessary and sufficient financing, adequate time (5 years) and staff (7 lexicographers), and effective management. The dictionary was printed in two 1000-page volumes.

Times tend to change: it was not immediately financially possible to start the vice versa dictionary and the 'electronic' manuscript (text file with XML-like mark-up) was filed away. Then, in late 2015, times changed once more, this time for the better. Finland and Estonia, both approaching their 100th anniversaries of the state (Finland in December 2017, Estonia in February 2018), decided to jointly celebrate their anniversaries by offering each other, as a gift, two dictionaries: the Finnish–Estonian dictionary (2003) will be made available for free on the web in 2017, and the new and long-awaited Estonian–Finnish dictionary will be compiled and published electronically in February 2018.

The project was initiated in 2016, the agreement of the joint project was signed by the Institute for the Languages of Finland (Helsinki) and the Institute of the Estonian Language (Tallinn) in June 2016, leaving 18 months for the partners to achieve their mission.

2. Generation of the EST-FIN database

The database for the Estonian–Finnish Dictionary (henceforth EST-FIN) was generated combining two databases: the source language part was formed from the database of the monolingual general Dictionary of Estonian (ESTDic, to appear in 2018/2019), and the target language part from the database of the Finnish–Estonian dictionary (2003, 2 vols, 90,000 lemmas; henceforth FIN-EST). The EST-FIN database was created in August 2016 with a list of 80,000 lemmas. Since the Dictionary of Estonian is an ongoing project, we only managed to operate with four-fifths of the complete manuscript (ca 100,000 lemmas). The remaining fifth will be added to the database in June 2017 (after finishing the compilation of ESTDic) following the same principles.

The database structure of the monolingual dictionary was transformed into the structure of the bilingual database thus: we added the elements for the target language information (Finnish translations, grammatical information, e.g. government etc.). Specific tuple- or triple-groups of adverbs and adverbial phrases denoting state, place, direction etc. (functioning in 2-3 internal/external local cases only) were divided into separate entries due to different translation equivalents in the target language. (Technicalities have the bad habit of lasting forever...)

Reversing a bilingual dictionary has already been described in numerous papers (e.g. Maks, 2007; Viks, 2008). However, the otherwise trivial process of extracting lemma-translation pairs, reversing them and re-sorting according to the position of the translation in the article, had some unexpected hurdles.

The FIN-EST, as a typical paper dictionary, used tilde as a replacement for the non-variable start of the lemma (see example a). Restoring full-blown textual representation required, in rare cases, manual editing. Variable parts in Estonian phrases were recursively split into primitives: thus, the pair FIN *aavistuksen verran suolaa* ('pinch of salt')–EST *raasuke (natuke, sipake) soola* resulted in three potential pairs (example b).

(a) FIN loik|ata ('to hop') – FIN hän ~kasi ojan yli, ... FIN ~ata vihollisen puolelle

→ FIN hän loikkasi ojan yli, FIN loikata vihollisen puolelle

(b) FIN aavistuksen verran suolaa (‘pinch of salt’) – EST raasuke (natuke, sipake) soola

→ FIN aavistuksen verran suolaa – EST raasuke soola

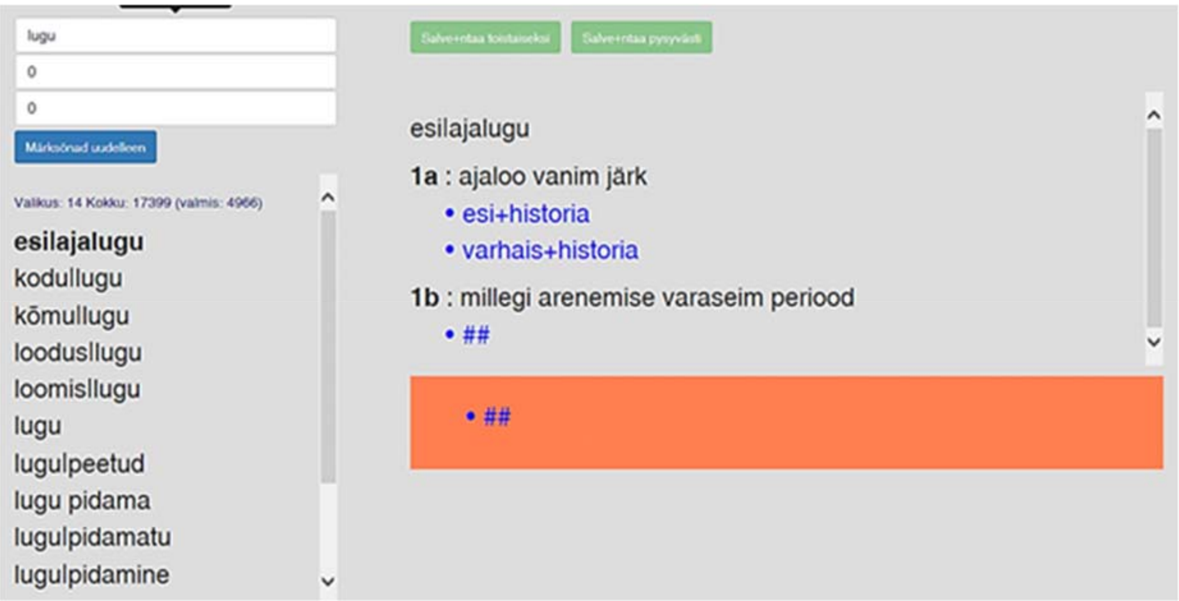
→ FIN aavistuksen verran suolaa – EST natuke soola

→ FIN aavistuksen verran suolaa – EST sipake soola

Estonian morphology subtly differs from Finnish in the usage of verb infinitive forms. The dictionary tradition in Estonian is to present verb lemmas in ma-infinitive, whereas the common da-infinitive is well understood by speakers of both languages and used for Estonian translations in the FIN-EST database. However, the EST-FIN database has lemmas in ma-infinitive and the reversing process had to make use of a morphological analyser for non-phrasal translations. Analysing phrases and collocations was not attempted for several reasons. The bulk of all pairs consist of nouns and noun collocations; verbs are relatively few. Also, many of the phrases would not be used, as they are constructed Estonian translations and not typically used as lemmas, e.g. Finnish *lautailla* (‘to surf’)–Estonian *rulaga sõita*, *lainelauaga sõita*, *lumelauaga sõita.* Applying morphological analysis to simple collocations without much context would also have produced many meaningless candidates. As the task at hand was not to compile a reverse dictionary, but rather to save as much time as possible by keeping routine tasks to a minimum, usage examples containing verb forms were ignored.

Extracting all translation pairs from the FIN-EST database (90,000 lemmas) resulted in 330,000 pairs, which could be inserted into pre-defined slots in the EST-FIN template. If the template Estonian dictionary entry had the same phrase that was used as translation, the corresponding Finnish phrase was filled in. This worked well for colloquialisms like EST *tere hommikust* (‘good morning’)–FIN *[hyvää] huomenta.* Most of the phrases (i.e. anything consisting of two or more words) still went unused. Where the EST-FIN database had a matching lemma, all the found Finnish counterparts were added as candidates for the translation equivalents under the first sense subdivision of the first homonym. The best unused candidates (ca. 40,000 words in Estonian) were used to form a complementary dictionary volume with skeleton articles filled in. These data will be used to grow the main dictionary in the future.

Instead of trying to guess the most appropriate homonym and sense from the limited data from the FIN-EST database, a second tool was provided as a first step in compiling the reversed dictionary. Our dictionary management system EELex does not support drag-and-drop editing, and voluminous dictionary entries require scrolling and copy-paste functions when one wants to move part of an entry to another position. To quickly delete, reorder and relocate generated Finnish translation candidates, we made a special tool that only displays enough information in the dictionary entry to indicate if, and under what sense, any of the translations belong (Figure 1). It only lists articles with several senses (to move) and/or several provided translations (to reorder) and keeps track of finished articles. Essentially it still is just an alternative user interface alongside EELex, giving the opportunity to simultaneously do other tasks the traditional way while automating the tedious task of distributing the translation candidates between senses.



Valikus: 14 Kokku: 17399 (valmis: 4966)

esilajalugu
kodu|lugu
kõmu|lugu
loodus|lugu
loomis|lugu
lugu
lugu|peetud
lugu pidama
lugu|pidamatu
lugu|pidamine

esilajalugu

1a : ajaloo vanim järk

- esi+historia
- varhais+historia

1b : millegi arenemise varaseim perood

- ##

- ##

Figure 1: Drag-and-drop editing tool as an alternative user interface alongside EELex

3. The compilation of the dictionary

Due to the lack of time—the unpostponable date of the 100th anniversaries—we have to impose on using automated lexicography as effectively as possible. At the same time, we are trying not to brush aside the substantial principles of dictionary-making.

3.1 Automatically-compiled entries

First, we tried to estimate how many entries might be 'ready' from the very beginning. Inspection of the initial EST-FIN database (80,000 entries) gave us a preliminary picture (Figure 2). Estonian is a predominantly agglutinative language, which, when creating new senses, mostly makes use of morphologic derivation. Polysemy applies to about every tenth Estonian word (Langemets, 2010: 269). Roughly the same is true of Finnish, which belongs to the same group of Uralic languages. The total entries with a single meaning in the EST-FIN database is 73,000. The 'ready'-quality was assumed for the simplest words only, i.e. words with a single meaning, preferably with no subsenses, and with no more than three translation equivalents or examples to be translated. There were 14,389 such 'ready' entries in the EST-FIN database (Figure 2), incudingl 9,784 with one translation equivalent (e.g. EST *hambapasta* ‘toothpaste’, EST *kuupmeeter* ‘cubic meter’); 3,053 with two equivalents (e.g. EST *sisepoliitika* ‘home affairs’); and 902 with three equivalents (e.g. EST *hormoon* ‘hormone’ (see Figure 4), EST *ajaleht* ‘newspaper’).

The remaining part comprising entries with a single meaning (58,611) still needs further editing: the lexicographer’s task is to select the proper equivalents, and select, edit and translate the examples. The most curious case is 69 (!) equivalents for the EST *pritsima* ‘to splash’ in the initial EST-FIN database. More than half of the entries (41,475) received no translation equivalents at all when reversing the database (e.g. EST *digitelevisioon* ‘digital TV’, EST *grammatikareegel* ‘grammar rule’, EST *alatähtsustama* ‘understate’).

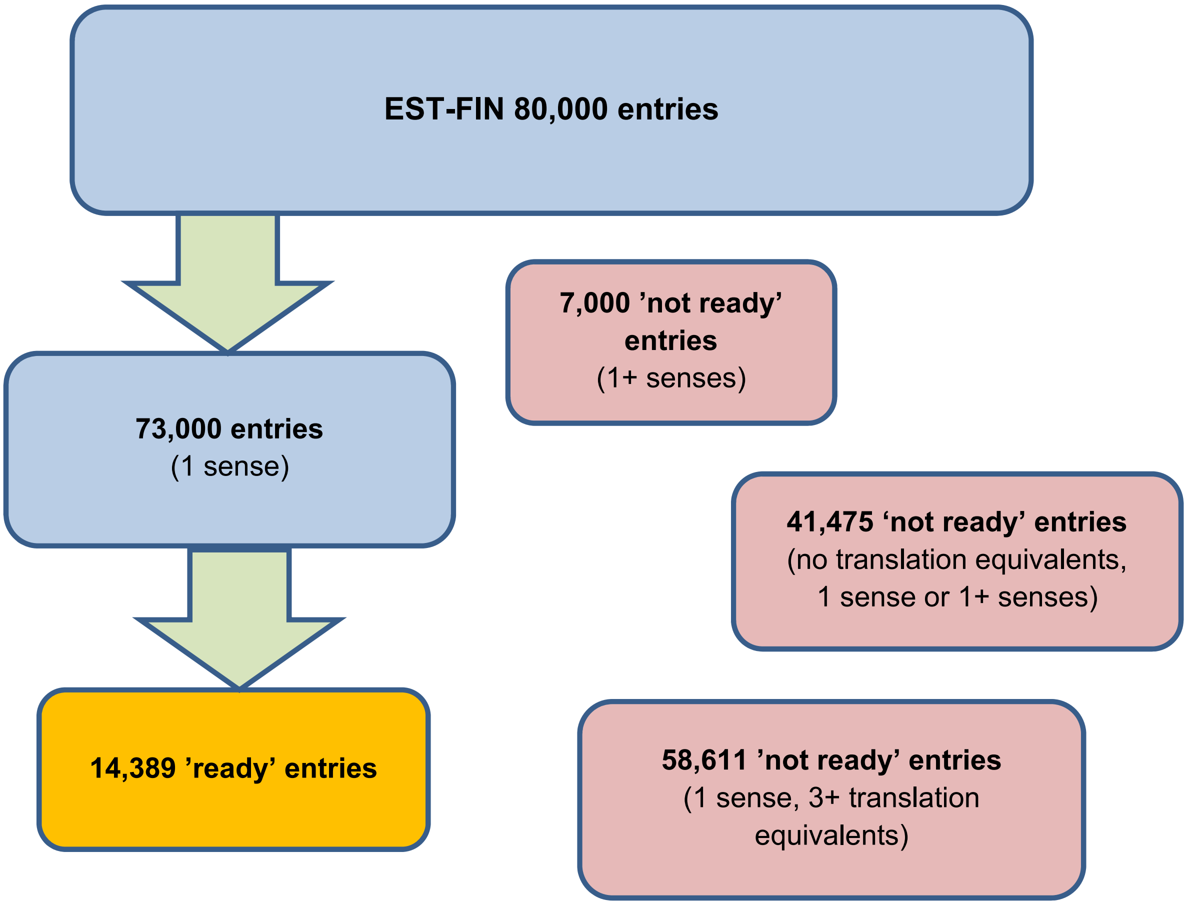

EST-FIN 80,000 entries
7,000 ’not ready’ entries (1+ senses)
73,000 entries (1 sense)
41,475 ‘not ready’ entries (no translation equivalents, 1 sense or 1+ senses)
14,389 ’ready’ entries
58,611 ’not ready’ entries (1 sense, 3+ translation equivalents)

Figure 2: The initial EST-FIN database: automatically completed ‘ready’ entries vs. ‘not ready’ entries

Secondly, we have refined the initial 80,000-lemma list by acquiring frequency information from a corpus (5 frequency groups for approx. 50,000 top frequent lemmas). The lexicographers edit the articles along those groups starting with the most frequent words, i.e. the 5,000 lemmas of the Basic Dictionary of Estonian (2014).

The lexicographers' main tasks are to edit the translation equivalent candidates (selecting and reordering) or to find the candidates, if missing, and decide whether or not to translate the existing usage examples, i.e. whether the translation is justified for being too difficult for the user.

3.2 Bilingualizing the monolingual dictionary

Since the source language part of the EST-FIN database comes from the monolingual dictionary, the usual drawbacks of the bilingual dictionary (e.g. insufficient sense discrimination, lack of example sentences, to begin with, mentioned in Adamska-Salaciak & Kernerman, 2016: 276) should be avoided. Unfortunately, drawbacks of another type remain.

The monolingual dictionary of Estonian is not aimed at learners but at native speakers of Estonian. The dictionary describes current Estonian and focuses on sense discrimination. It is being compiled using etTenTen corpus and the Sketch Engine tool (Kilgarriff et al., 2004). The examples are meant to illustrate the senses, they are real as well as natural in Estonian, but in many cases, are definitely not good examples for learners. For another, the Estonian Collocations Dictionary (ECD, to appear in 2018, see Kallas et al., 2015) grammatical constructions (collocations) have been extracted from the corpus. Since the ECD work is in progress, the database is inadequate so far and not usable for the EST-FIN database.

There are 95,000 usage examples in the EST-FIN database. The average number of examples per entry is 1.1. Around half of the entries have no examples in the database. We are aware from the previous studies (Frankenberg-Garcia 2012, 2014, quoted in Lew, 2015: 5) that users find three examples per sense significantly more helpful than just a single example. The problems remain: how could we manage to translate all necessary examples? How could we obtain enough examples for all the entries? How might we obtain good examples for the bilingual dictionary?

Bilingualizing the dictionary involves bilingualizing the metalanguage (domain and style labels, grammatical information etc.) as well as—to some extent—the explanations. Bilingualizing explanation means first and foremost simplifying the definitions: the definitions in a monolingual dictionary are usually much longer and more complex than in a bilingual dictionary. We have decided to preserve the semantic structure of the treatment of the source language, but for better understanding of Estonian (as L2) we have shortened many definitions into glosses.

3.3 Directionality of the dictionary

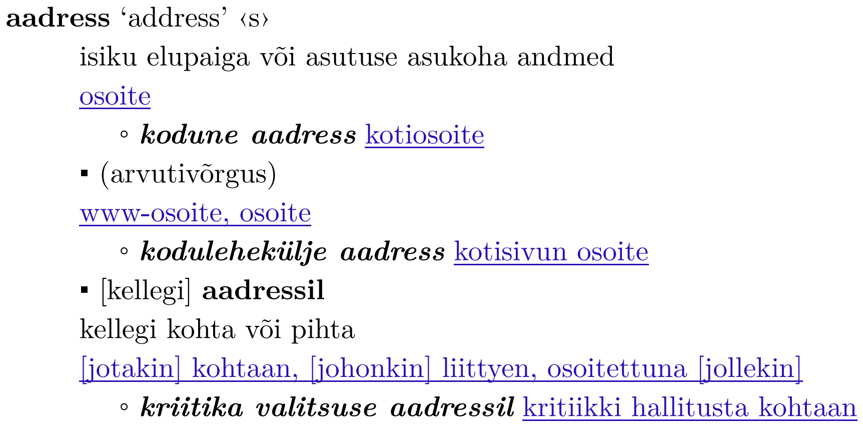
Traditionally all (paper) dictionaries have been compiled to fulfil the needs of all conceivable users trying to solve several different tasks. A good bilingual dictionary should enable both understanding L2 and producing L2, though the most important task seems to be the latter—an opinion supported by many researchers (see Adamska-Salaciak & Kernerman 2016). The best dictionary would be ‘addressed specifically to the native speakers of one of its two object languages’ (Adamska-Salaciak & Kernerman, 2016). The dictionary should somehow selectively separate information for different purposes, i.e. it should be monodirectional.

For Estonian users, the EST-FIN dictionary functions as a L1–L2 dictionary, helping L1 users to talk about specific phenomena of his/her own culture. There are several concepts (senses) that are not lexicalized in L2 (e.g. EST *akadeemiline tund* (‘(in the universities:) 45 mins’), EST *präänik* (‘a thick soft spicy biscuit’)).

For Finnish users, the EST-FIN dictionary functions as the receptive L2–L1 dictionary, helping to render the L2 meanings.

For production of L2 (Estonian or Finnish), it should contain first and foremost collocational information as well as good examples.

Modes of provision of semantic information. In the EST-FIN database, the semantic information explaining the lexical items of the source language is placed into definitions (EST, rarely FIN), equivalents (FIN) and examples (EST, rarely FIN). Estonian seems to dominate over Finnish (Figure 3, Finnish underlined) in the entries for single words (EST *aadress* ‘address’) but the phrasal verbs and idioms (EST *käsi peseb kätt* ‘one hand washes another’) are as carefully explained in Finnish as in Estonian. There are about 6,000 multiword units in the EST-FIN database and as part of the headword list they are treated likewise.



aadress ‘address’ ‹s›
isiku elupaiga või asutuse asukoha andmed
osoite
◦ ***kodune aadress*** kotiosoite
▪ (arvutivõrgus)
www-osoite, osoite
◦ ***kodulehekülje aadress*** kotisivun osoite
▪ [kellegi] **aadressil**
kellegi kohta või pihta
[jotakin] kohtaan, [johonkin] liittyen, osoitettuna [jollekin]
◦ ***kriitika valitsuse aadressil*** kritiikki hallitusta kohtaan

käsi peseb kätt ‘one hand washes another’

ütlus selle kohta, et teenele vastatakse teenega
ilmaus siitä, että palvelukseen vastataan palveluksella
käsi kättä pesee KUV, käsi käden pesee KUV

Figure 3: Providing semantic information in both languages (Finnish translations underlined, other symbols: ◦ usage example; ▪ subsense; FIN definition)

Modes of treating synonyms. If synonymy is presented for the concept (in the Estonian part) and is designated by different terms (lemma and its synonyms, marked with = in the database) and described by the same definition in the particular sense, then the semantic information (incl. Finnish translations) remains the same for all counterparts (Figure 4: EST *hormoon* ‘hormone’ = *sise|nõre* = *inkreet*). We have agreed that the domain label (in Finnish, e.g. FYSIOL) functions as a semantic gloss for the Finnish, so it would not be necessary to translate Estonian definitions.

hormoon ‘hormone’ ‹s›

aine, mis reguleerib inimese ainevahetust ning organismi talitlust
(= sise|nõre, inkreet)
hormoni FYSIOL, sisä+erite FYSIOL, umpi+erite FYSIOL

sise|nõre ‘hormone’ ‹s›

aine, mis reguleerib inimese ainevahetust ning organismi talitlust
(= sise|nõre, inkreet)
hormoni FYSIOL, sisä+erite FYSIOL, umpi+erite FYSIOL

inkreet ‘hormone’ ‹s›

aine, mis reguleerib inimese ainevahetust ning organismi talitlust
(= sise|nõre, inkreet)
hormoni FYSIOL, sisä+erite FYSIOL, umpi+erite FYSIOL

Figure 4: Treating synonyms in the dictionary (Finnish translations underlined)

Collocational information. Collocations and other lexical bundles are not systematically and explicitly treated in the EST-FIN database. As mentioned above, the work on the Estonian Collocations Dictionary is in progress.

Modes of provision usage examples. The examples are essential for all types of learners as well as L2 users. It was attested 20 years ago, that the dominance of bilingual dictionaries is greater for L1–>L2 translation (i.e. for producing L2) than for L2–>L1 translation (Atkins & Varantola, 1997). The examples of the EST-FIN database are meant to illustrate the senses of the monolingual dictionary, i.e. in the EST-FIN database they function for L2–>L1 translation.

Translation of dictionary examples has not always been seen as the best solution for a bilingual dictionary (Adamska-Salaciak, 2006; Hmeljak Sangawa & Erjavec, 2012). A corpus is needed to provide typical L2 examples in the (unidirectional) bilingual dictionary. Adamska-Salaciak (2006) favours presenting L2 examples only, with the exception of difficult cases (2006: 494):

> Naturally, even in dictionaries whose examples are normally left untranslated, exception must be made for sentences or parts thereof which might be too difficult for the average user to interpret on their own.

And another statement from a lexicographer (emailed to one of the authors, January 2017) keeping in mind producing L2:

> Personally, I use bilingual dictionaries pretty rarely. Google is often fine, especially for phrasal expressions: I test what I intend to say against data on the web.

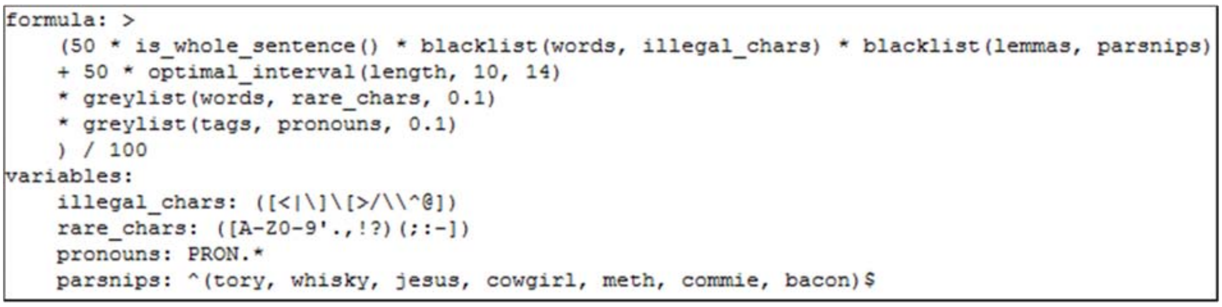
So, would it not be marvellous if our user could have real (authentic) material at hand? Since we cannot give the user access to a large high-quality bilingual (parallel) corpus as well as following Adamska-Salaciak (2006), we should instead provide our users with good L2 examples.

Next, we will discuss the possibilities of obtaining good examples for both languages.

4. Good examples for Estonian

Presenting authentic sentences in dictionaries is a common practice in modern lexicography. One of the possibilities for extracting authentic examples from corpora is to use GDEX (Kilgarriff et al., 2008)—a software part of Sketch Engine (Kilgarriff et al., 2004). GDEX evaluates syntactic and lexical features of sentences and sorts concordances according to how perfectly they meet all the relevant criteria. As a result, GDEX offers a list of sentences: the better candidates are at the top of the list and the not-so-good ones at the bottom. The theoretical framework for GDEX development is proposed in Kilgarriff et al. (2008) and Kosem et al. (2011, 2013).

GDEX was developed as a set of classifiers for specific features and it was first used in the preparation of an electronic version of the Macmillan English Dictionary (Macmillan 2002, 2007). All features are quantifiable, e.g. sentence length, word length, presence or absence of certain words or non-words, the number of pronouns in the sentence etc. Each feature has its own individual value and GDEX counts them in a fixed way. It ranks the sentence with a score from 0 to 1. The specification of measured features and the way in which they are combined is defined in files called GDEX configurations (Figure 5).



```
formula: >
    (50 * is_whole_sentence() * blacklist(words, illegal_chars) * blacklist(lemmas, parsnips)
    + 50 * optimal_interval(length, 10, 14)
    * greylist(words, rare_chars, 0.1)
    * greylist(tags, pronouns, 0.1)
    ) / 100
variables:
    illegal_chars: ([<|\]\[>/\\^@])
    rare_chars: ([A-Z0-9'.,!?)(;:-])
    pronouns: PRON.*
    parsnips: ^(tory, whisky, jesus, cowgirl, meth, commie, bacon)$
```

Figure 5: Syntax of GDEX configuration files[1]

While some of the parameters apply to all languages (e.g. sentences start with a capital letter and end with a punctuation mark), some are language-specific (e.g. sentence length, keyword position etc.). Therefore, it is reasonable to modify parameters according to the languages that the GDEX configuration will be applied upon.

First, GDEX configuration for Estonian was developed in the Institute of the Estonian Language in 2014, in collaboration with Lexical Computing Ltd., for the automatic extraction of the Estonian Collocations Dictionary (ECD) database. ECD is a monolingual dictionary aimed at learners of Estonian as a foreign or second language at the upper intermediate and advanced levels. The Estonian configuration is being continuously developed. The latest version was set up in 2016 based on research carried out under ISCH COST Action IS10305 European Network of e-Lexicography during a short term scientific mission in the University of Ljubljana.

The Estonian corpus (560 mio words) contains of two parts: the Estonian Reference Corpus (biased heavily towards newspaper texts), and the web corpus, crawled by SpiderLing in 2013. Syntagmatic relations of content words are described as lexico-grammatical constructions defined by means of morphosyntactic categories (phrase type, part of speech, inflectional categories). The Estonian Sketch Grammar has been worked out by Kallas (2013). The corpus was tagged for sentences, clauses and morphology (POS-tag and inflections) by Filosoft Ltd (ESTMORF).

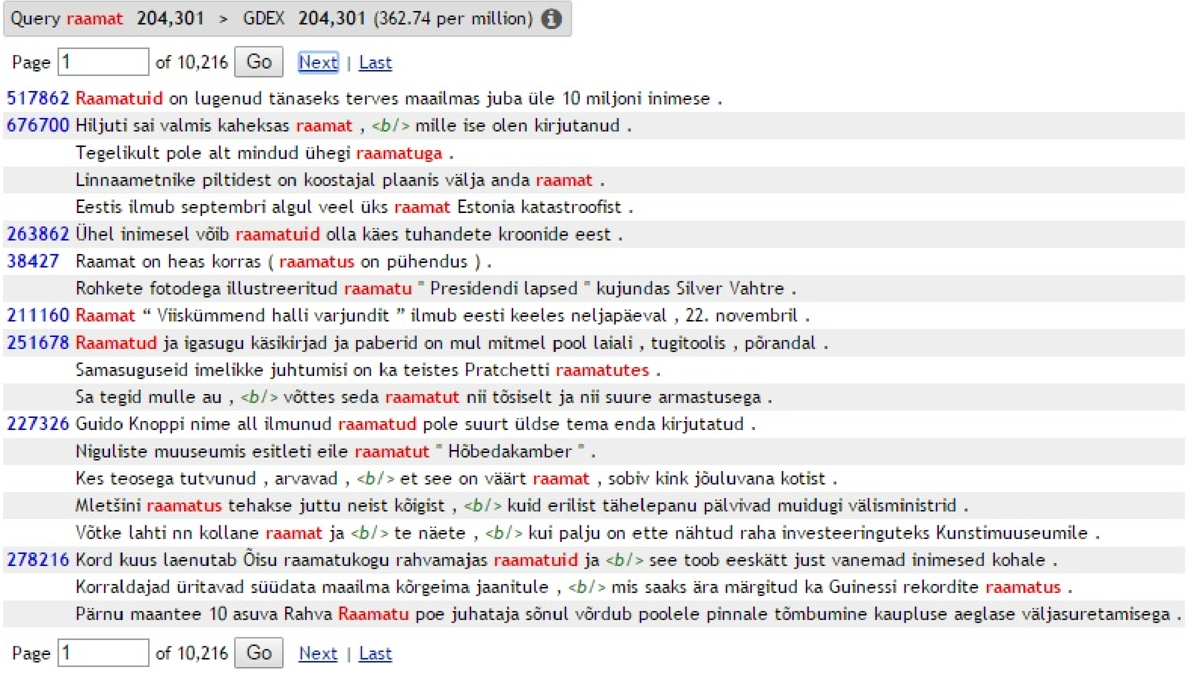
In developing GDEX for Estonian, the needs of language learners have always been considered. Sentences in learner dictionaries should ideally be short, not syntactically and grammatically complex, include frequent words, help the learner to understand the meaning of an unknown word, and/or show the collocation in its typical context.

[1] https://www.sketchengine.co.uk/syntax-of-gdex-configuration-files/ (25.5.2017).

Parameters for Estonian are as follows (Koppel 2017):

- Starts with a capital letter and ends with a punctuation mark;
- Sentence length is 4 to 20 tokens;
- Optimal sentence length is 6 to 12 tokens;
- Contains a verb;
- Maximum word length is 20 characters;
- Certain characters (e.g. <|\]\[>/\\}{ˆ@• · *#=_~) are prohibited and certain characters (e.g. ;:“„”«»"’×…§-) are penalized;
- Certain words (e.g. *pigem* ‘rather’, *teisisõnu* ‘in other words’, *seetõttu* ‘for that reason’), word pairs (e.g. *seda enam* ‘even more’, *teiste sõnadega* ‘in other words’, *teisest küljest* ‘on the other hand’) and sentence initial tags (e.g. conjunction, abbreviation, interjection) are prohibited from appearing in the beginning of the sentence;
- Words with a frequency of less than 5 are prohibited;
- Lemmas with a frequency of less than 1000 are penalized;
- Keyword repetition is prohibited;
- Sentences including pronouns, words from graylist (e.g sensitive words, profanities), abbreviations, proper names, certain non-finite constructions are penalized;
- Sentences containing more than 2 verbs, more than 1 adverb, more than 1 pronoun, more than 1 conjunction, more than 1 proper name, more than 1 numeral and more than 1 comma are penalized.

Figure 6 shows the GDEX output after the latest version for Estonian is implemented.



```
Query raamat  204,301  >  GDEX  204,301 (362.74 per million)

Page 1  of 10,216  Go   Next | Last

517862 Raamatuid on lugenud tänaseks terves maailmas juba üle 10 miljoni inimese .
676700 Hiljuti sai valmis kaheksas raamat , <b/> mille ise olen kirjutanud .
       Tegelikult pole alt mindud ühegi raamatuga .
       Linnaametnike piltidest on koostajal plaanis välja anda raamat .
       Eestis ilmub septembri algul veel üks raamat Estonia katastroofist .
263862 Ühel inimesel võib raamatuid olla käes tuhandete kroonide eest .
38427  Raamat on heas korras ( raamatus on pühendus ) .
       Rohkete fotodega illustreeritud raamatu " Presidendi lapsed " kujundas Silver Vahtre .
211160 Raamat “ Viiskümmend halli varjundit ” ilmub eesti keeles neljapäeval , 22. novembril .
251678 Raamatud ja igasugu käsikirjad ja paberid on mul mitmel pool laiali , tugitoolis , põrandal .
       Samasuguseid imelikke juhtumisi on ka teistes Pratchetti raamatutes .
       Sa tegid mulle au , <b/> võttes seda raamatut nii tõsiselt ja nii suure armastusega .
227326 Guido Knoppi nime all ilmunud raamatud pole suurt üldse tema enda kirjutatud .
       Niguliste muuseumis esitleti eile raamatut " Hõbedakamber " .
       Kes teosega tutvunud , arvavad , <b/> et see on väärt raamat , sobiv kink jõuluvana kotist .
       Mletšini raamatus tehakse juttu neist kõigist , <b/> kuid erilist tähelepanu pälvivad muidugi välisministrid .
       Võtke lahti nn kollane raamat ja <b/> te näete , <b/> kui palju on ette nähtud raha investeeringuteks Kunstimuuseumile .
278216 Kord kuus laenutab Õisu raamatukogu rahvamajas raamatuid ja <b/> see toob eeskätt just vanemad inimesed kohale .
       Korraldajad üritavad süüdata maailma kõrgeima jaanitule , <b/> mis saaks ära märgitud ka Guinessi rekordite raamatus .
       Pärnu maantee 10 asuva Rahva Raamatu poe juhataja sõnul võrdub poolele pinnale tõmbumine kaupluse aeglase väljasuretamisega .

Page 1  of 10,216  Go   Next | Last
```

Figure 6: GDEX output for Estonian lemma *raamat* ‘book’

5. Good examples for Finnish

In Finnish, the goal is to find good examples of translation equivalents of Estonian headwords with the help of GDEX configuration (Heinonen, 2015). The configuration for Finnish is based on the one developed by Kristina Koppel for Estonian (see above).

In general, the same configuration works well for both languages. We prefer short, simple and context-free examples: conjunctions, sentence-initial connectives and anaphoric elements are unfavourable. At first sight, one might think that there is not much else to do except replace the original Estonian lexical items by Finnish words. For instance, an Estonian connective expression *teiste sõnadega* ‘in other words’ would be replaced by its Finnish equivalent *toisin sanoin.* However, a considerable part of the Finnish data used in this project represents informal register. Finland’s strict copyright legislation is partly to blame for this since it strongly favours the use of freely-accessible web-pages for corpora. In any case, stylistic variation poses a problem which is not as noticeable on the Estonian side. It is possible that Finnish speakers tend to write more informally than Estonians, or that there is a bigger difference between the standard and the vernacular in Finnish than in Estonian. Whichever is the case, it is more confusing than helpful if the examples contain words and expressions that are not even recorded in a standard Finnish dictionary. An ideal solution would be a grammatically augmented dictionary that could deal with variation in words and inflectional affixes.

The Finnish corpus is tagged morphologically but not syntactically, and there is no way to fix its colloquial syntactic patterns. However, what can be done, is to ban most common colloquial word forms by including them in the list of “bad words”, which is one of the parameters of GDEX and is originally used for excluding inappropriate words.

In the Finnish GDEX configuration, this is achieved by listing such prevalent items as spoken forms of general verbs, pronouns and some other grammatical words:

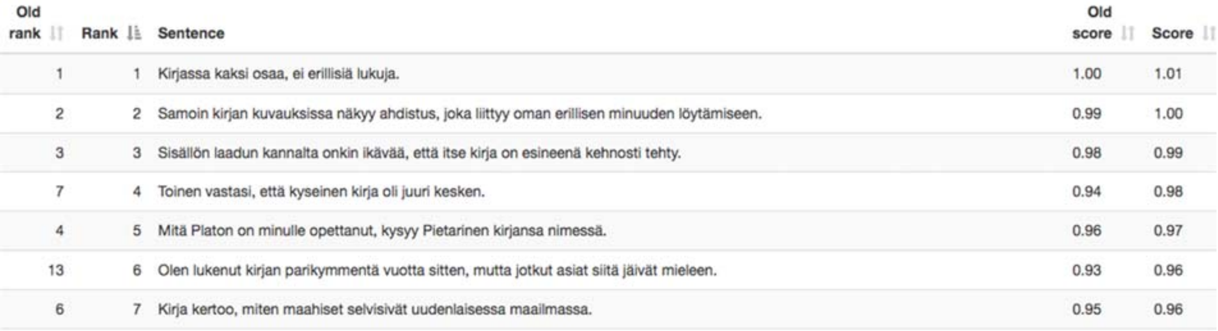
- *oon, oot, oo, tuu, paan, sais, vois*, etc. (informal forms of frequent verbs)
- *mä, sä, toi, noi, mulle, sulla, tolle*, etc. (informal forms of pronouns)
- *vaik, nii, niiku, ku, kans*, etc. (informal forms of conjunctions and adpositions)

Since these items tend to co-occur with other informal words and structures, this is in fact a rather efficient way of preventing unwanted example sentences to surface.

The task of obtaining good Finnish examples is complicated also by the fact that many seemingly fine sentences are hampered by morphosyntactic misanalysis. For instance, out of a test sample of 30 sentences intended to illustrate the use of the verb *kaupata* ‘to trade’, only seven were in fact occurrences of this lemma. The remaining sentences

(23/30) instead contained the noun *kauppa* in its various senses (‘a store’, ‘a deal’, ‘commerce’, etc.). The reason for this is that words *kaupata* and *kauppa* share the same stem, and some frequent inflectional forms are ambiguous. The same situation is also common in English: the form *shops* can be either a noun in the plural or a verb in the third person singular. Since the noun *kauppa* is more common than the verb *kaupata*, its occurrences dominate in the concordance. Furthermore, such ambiguities bring about mismatches in a bilingual context if the intended translation equivalents do not show up but are instead replaced with misanalyzed forms in corpus examples.

Figure 7 displays a list of top examples for the word *kirja* (‘a book’). The top score is given to a sentence *Kirjassa kaksi osaa, ei erillisiä lukuja.* This sentence should not score as highly since it does not have a finite verb. Its literal translation is: ‘two parts in the book, no separate chapters’. Again, the problem lies in an ambiguous word form: this time the word *osaa*, which has two readings, ‘part’ in the partitive singular, or ‘can, be able’ in the third person singular.



Old rank	Rank	Sentence	Old score	Score
1	1	Kirjassa kaksi osaa, ei erillisiä lukuja.	1.00	1.01
2	2	Samoin kirjan kuvauksissa näkyy ahdistus, joka liittyy oman erillisen minuuden löytämiseen.	0.99	1.00
3	3	Sisällön laadun kannalta onkin ikävää, että itse kirja on esineenä kehnosti tehty.	0.98	0.99
7	4	Toinen vastasi, että kyseinen kirja oli juuri kesken.	0.94	0.98
4	5	Mitä Platon on minulle opettanut, kysyy Pietarinen kirjansa nimessä.	0.96	0.97
13	6	Olen lukenut kirjan parikymmentä vuotta sitten, mutta jotkut asiat siitä jäivät mieleen.	0.93	0.96
6	7	Kirja kertoo, miten maahiset selvisivät uudenlaisessa maailmassa.	0.95	0.96

Figure 7: GDEX output for Finnish lemma *kirja* ‘book’

However, Figure 7 shows that even with some faulty analyses, the parameters succeed in ordering the sentences in the corpus in a reasonable way. Scrolling down the list, one encounters lengthy, complex, or fragmentary sentences.

The GDEX for Finnish is still being developed. This is being performed in the GDEX editor[2], which is a standalone tool of Sketch Engine developed by Lexical Computing Ltd. software developer Jan Michelfeit. The GDEX editor enables comparisons between two settings of parameters in parallel, and this is used so that the Estonian-based configuration has been taken as a starting point and it is modified step by step towards a configuration that arranges Finnish data in an optimal way. In Figure 7, the results from Estonian-based GDEX (with few lexical additions) are labelled as “Old rank” and “Old score”. Once a parameter is modified, the changes to its ranking and scores can be immediately calculated. As can be seen, the tops of the lists are, in any case, very

[2] https://beta.sketchengine.co.uk/gdex_editor

similar. After the screenshot in Figure 7 was taken, the word *toinen* (‘other’, ‘another’, ‘second’) was tentatively added to the list of prohibited words in a sentence-initial position, with the effect of dropping the rank of the sentence number 4 *Toinen vastasi, that...* ('The other replied that...') to number 25.

6. Conclusion

In the semi-automated compilation process of the Estonian-Finnish (EST-FIN) dictionary we have roughly followed the same steps as mentioned by Gantar et al. (2016: 201).

We reused the previous lexical databases. The database was generated combining two existing databases: the source language part was formed from the database of the monolingual general Dictionary of Estonian (DicEst, to appear in 2018/2019), and the target language part from the database of the Finnish–Estonian dictionary (FIN-EST, 2003). It is worth mentioning that the FIN-EST dictionary did not start from scratch: the base for the source language (Finnish) came from another bilingual dictionary, Finnish–Swedish dictionary (1997, Helsinki). The Estonian lexicographers worked with the XML-like ‘electronic’ manuscript, filling in the slots for translation equivalents as well as translating the usage examples for the FIN-EST dictionary.

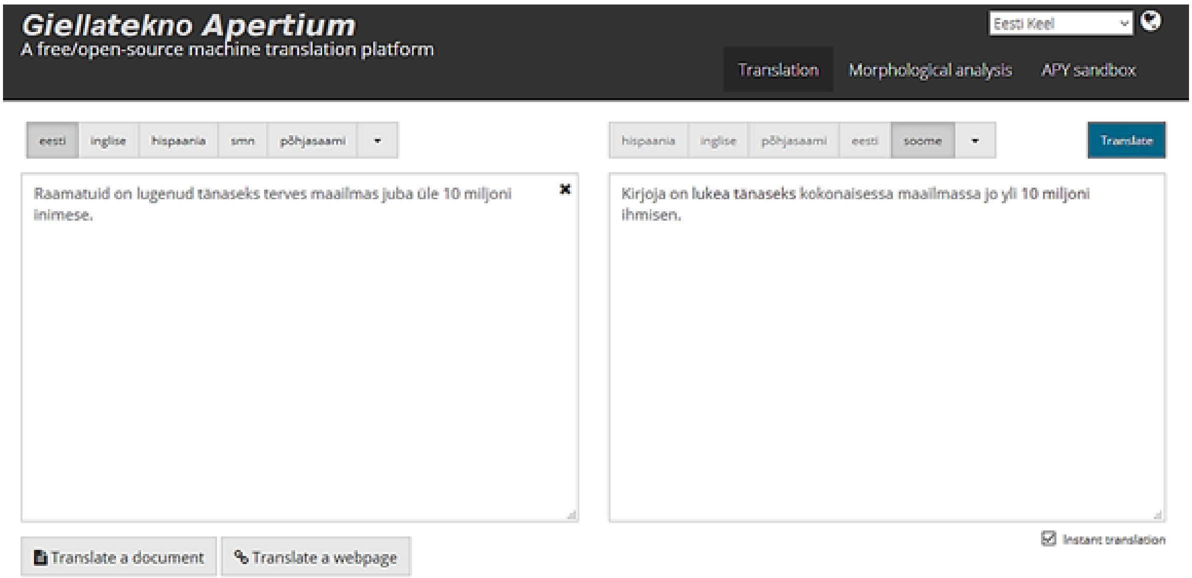
We refined the initial 80,000-lemma list by acquiring frequency information from a corpus (5 frequency groups for approx. 50,000 top frequent lemmas).

The best unused candidates from the FIN-EST database (ca 40,000 words in Estonian), which so far constitute the complementary EST-FIN dictionary volume with skeleton articles filled in, will be used to grow the main dictionary in the future.

We will extract L2 example sentences—daydreaming of getting hold of good examples for both languages, Estonian and Finnish. The first GDEX configuration for Estonian was developed in the Institute of the Estonian Language in 2014 in collaboration with Lexical Computing Ltd. Since then it has been under continuous development. The latest version was set up in 2016. The GDEX configuration for Finnish is based on the Estonian one; it still needs developing.

Subsequently, likely after the finalization of both projects (EST-FIN and ESTDic) we will link the collocations database to these dictionaries to fulfil the productive needs of advanced learners of Estonian as well as the needs of L2 users. The overall development plan at the Institute of the Estonian Language concerning dictionaries and (terminological) databases is to change over to the standardized (Unified) Data Model for better presentation and linking of the lexicographic information congregated at our Institute.

Finally, we have thought of using a free/open-source Machine Translation platform Giellatekno Apertium to provide translations for usage examples (Kaalep et al. 2017). Figure 7 displays the translation of the first sentence in the Estonian GDEX output (Figure 6). Those understanding Finnish might get the feeling that the translation system does not work well at all (Figure 7). However, we still have a well-grounded hope, as the rule-based translation system will be complemented by the real 90,000-lemma Finnish–Estonian dictionary (2003, 2 vols) instead of the small 15,000-lemma dictionary compiled automatically via other pairs of languages.



Raamatuid on lugenud tänaseks terves maailmas juba üle 10 miljoni inimese.

Kirjoja on lukea tänaseks kokonaisessa maailmassa jo yli 10 miljoni ihmisen.

Figure 7: Machine translation platform Giellatekno Apertium: from Estonian to Finnish

7. Acknowledgements

We would like to thank Jelena Kallas for her constructive comments and suggestions. The research was partly funded by ISCH COST action COST-STSM-IS1305-27016 European Network of e-Lexicography.

8. References

Adamska-Salaciak, A. (2006). Translation of Dictionary Examples – Notoriously Unreliable? In *Proceedings of the 12th EURALEX International Congress*. Torino, Italy, 493–501.

Adamska-Salaciak, A. & Kernerman, I. (2016). Introduction: Towards better dictionaries for learners. In *International Journal of Lexicography*, 29(3), 271–278. doi: 10.1093/ijl/ecw033

Atkins, S. & Varantola, K. (1997). Monitoring dictionary use. In *International Journal of Lexicography*, 10(1), 1–45.

DicEst = The Dictionary of Estonian (to appear in 2018/2019). Institute of the Estonian Language.

ECD = The Estonian Collocations Dictionary (to appear in 2018). Institute of the Estonian Language.

EST-FIN = Estonian-Finnish database (to appear in 2018). Institute of the Estonian Language.

ESTMORF = Eesti keele morfoloogiline analüsaator [Morphological Analyzer of Estonian]. Filosoft OÜ. http://www.filosoft.ee/html_morf_et/morfoutinfo.html (3.7.2017).

FIN-EST = Soome-eesti suursõnaraamat. Suomi–viro-suursanakirja (2003). [Finnish-Estonian dictionary.] 2 vols. Tallinn: Eesti Keele Sihtasutus.

Gantar, P. & Kosem, I. & Krek, S. (2016). Discovering automated lexicography: the case of the Slovene database. In *International Journal of Lexicography*, 29 (2), 200–225. doi: 10.1093/ijl/ecw014

Giellatekno Apertium = Giellatekno Apertium: A free/open-source machine translation platform http://gtweb.uit.no/mt-testing/index.est.html?dir=est-fin#translation

Heinonen, T. (2015). Development of Sketch Grammar and GDEX (Good Dictionary Example) for Finnish. Scientific Report of Short Term Scientific Mission, COST STSM.

Hmeljak Sangawa, K. & Erjavec, T. (2012). JaSlo: Integration of a Japanese-Slovene bilingual dictionary with a corpus search system. In *Acta Linguistica Asiatica*, 10 (3). http://revije.ff.uni-lj.si/ala/article/view/223

Kaalep, H.-J. & Tyers F. M. & Trosterud, T. (2017). Soome-eesti reeglipõhise masintõlkesüsteemi DEMO. [Finnish-Estonian rule-based machine translation DEMO.] In Abstracts of EAAL 16 th Annual Conference, April 20-21, 2017, Tallinn, Estonia. Accessed at: https://www.rakenduslingvistika.ee/wp-content/uploads/
2016/04/Teesid-2017-3.pdf (18.5.2017)

Kallas, J. (2013). Syntagmatic relationships of Estonian content words in corpus and pedagogical lexicography. PhD thesis.

Kallas, J. & Koppel, K. & Tuulik, M. (2015). Korpusleksikograafia uued võimalused eesti keele kollokatsioonisõnastiku näitel. [New possibilities in corpus lexicography based on the examples of the Estonian Collocation Dictionary.] In *Eesti Rakenduslingvistika Ühingu aastaraamat*, 11, 75–94. doi: 10.5128/ERYa11.05

Kilgarriff, A. & Rychly, P. & Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds.) *Proceedings of the XI Euralex International Congress.* Lorient: Université de Bretagne Sud, 105–116.

Kilgarriff, A. & Husák, M. & McAdam, K. & Rundell, M. & Rychlý, P. (2008). GDEX:

Automatically finding good dictionary examples in a corpus. In E. Bernal, J. DeCesaris (eds.) *Proceedings of the 13th EURALEX International Congress.* Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, 425–432.

Koppel, K. (2017). Heade näitelausete automaattuvastamine eesti keele õppesõnastike jaoks. [Automatic detection of good dictionary examples in Estonian learner's dictionaries.] In *Eesti Rakenduslingvistika aastaraamat*, 13, 53−71.

Kosem, I. & Husák, M. & McCarthy, D. (2011). GDEX for Slovene. In *Proceedings of eLex 2011*, 151–159. http://elex2011.trojina.si/Vsebine/proceedings/eLex2011-19.pdf

Kosem, I. & Gantar, P. & Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, & M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper.* Proceedings of the eLex 2013 conference, 17–19 October 2013, Tallinn, Estonia, 17–19. http://eki.ee/elex2013/proceedings/eLex2013_03_Kosem+Gantar+Krek.pdf

Kovář, V. & Baisa, V. & Jakubíček, M. (2016). Bilingual Word Sketches. In *International Journal of Lexicography*, 29 (3), 339–352. doi: 10.1093/ijl/ecw029

Langemets, M. (2010). *Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras.* [Systematic polysemy of nouns in Estonian and its lexicographic treatment in Estonian language resources.] Tallinn: Eesti Keele Sihtasutus.

Lew, R. (2015). Dictionaries and their users. In *International Handbook of Modern Lexis and Lexicography.* Berlin-Heidelberg: Springer-Verlag, 1–9. doi: 10.1007/978-3-642-45369-4_11-1

Macmillan 2002, 2007 = *Macmillan English Dictionary for Advanced Learners* (First and Second editions). 2002, 2007. Rundell, ed. Macmillan, London.

Maks, E. (2007). OMBI: the practice of reversing dictionaries. In *International Journal of Lexicography*, 20(3), 259–274. doi: 10.1093/ijl/ecm028

Viks, Ü. (2008). Eesti-X-keele sõnastik ja grammatika. [Estonian-X dictionary and grammar.] In *Eesti Rakenduslingvistika aastaraamat*, 4, 247−261.