

SkELL Corpora as a Part of the Language Portal

Sõnaveeb: Problems and Perspectives

Kristina Koppel¹, Jelena Kallas¹, Maria Khokhlova²,

Vít Suchomel^{3,4}, Vít Baisa^{3,4}, Jan Michelfeit³

¹ Institute of the Estonian Language, Estonia

² St. Petersburg State University, Russia

³ Lexical Computing Ltd., Czech Republic

⁴ Masaryk University, Czech Republic

E-mail: kristina.koppel@eki.ee, jelena.kallas@eki.ee, m.khokhlova@spbu.ru,

vit.suchomel@sketchengine.co.uk, vit.baisa@sketchengine.co.uk,

jan.michelfeit@sketchengine.co.uk

Abstract

The paper provides an analysis of the quality and presentation of authentic corpus sentences from Sketch Engine for Language Learning (SkELL) corpora (Baisa & Suchomel 2014), based on the example of Sõnaveeb (Wordweb), a new language portal being developed in the Institute of the Estonian Language. Currently Sõnaveeb contains a total of 150,000 Estonian headwords; about 70,000 of them have Russian equivalents. Authentic corpus sentences are displayed for both languages. In some cases (e.g. terms, derived forms, compounds and multi-word expressions), corpus sentences are the only source of usage examples that are available on the portal.

We describe the parameters of Good Dictionary Examples (GDEX) (Kilgarriff et al., 2008) configurations for Estonian and for Russian used for the compilation of etSkELL 2018 and ruSkELL 1.6 corpora, give an overview of an evaluation of the GDEX configuration for Estonian, and outline the requirements for the user-friendly presentation of SkELL corpora as a part of the language portal.

Keywords: GDEX; SkELL; learner corpus; Estonian; Russian

1. Introduction

Despite the fact that most modern dictionaries are corpus-based, displaying authentic corpus data in dictionary portals is still quite a new trend in e-lexicography. There are some dictionaries (e.g. the 5th edition of LDOCE¹, Wordnik²) that offer automatically-retrieved corpus sentences alongside manually-selected examples (Cook, 2014) but as it became evident in a survey about lexicographic practices in Europe (Kallas et al., 2019), most dictionary websites do not offer automatically-retrieved corpus sentences nor a

¹ <http://ldoce.longmandictionariesonline.com/main/Home.html> (3 June 2019).

² www.wordnik.com (3 June 2019).

link to corpus data. The survey revealed that if links are offered, they are generally automatic URLs pointing to the Corpus Query System (CQS) for the headword. The user cannot specify which elements they want to retrieve from the corpus (e.g. example sentences with metadata/without metadata). Only after the user has entered the CQS, can they change the query.

One way to optimize this process is to display corpus data not from general corpora but from corpora that consist of pre-filtered examples instead. As an example of such corpora, Sketch Engine for Language Learning (SkELL) corpora can be used. SkELL corpora were initially intended for language learning purposes (e.g. for teachers or students to efficiently find out how a word is used in a language), but they can also be seen as a source of clean and processed examples (which is especially the case when we speak of web data).

The principle of SkELL corpora is to prepare roughly 1 billion tokens of clean sentences from various resources. This is achieved either by compiling several trusted resources (in the case of English) or extensive filtering of web-based corpora (in the case of Estonian and Russian). De-duplication (i.e. the removal of the same or even similar text fragments) is a part of the process. Cleaned data (sentences) are evaluated with the example extraction tool GDEX (Good Dictionary Examples, Kilgarriff et al., 2008) and then sorted by GDEX scores. The scores correspond to sentence values and vary from 0 (the worst) to 1 (the best). Its computation is based on a formula that deals with a variety of formal classifiers, paying attention to various features (see Section 3). The formula itself is described in the GDEX configuration files³. Unlike other corpora, SkELL corpora do not contain whole documents (assuming language learners do not need them) but only sentences with the highest scores (i.e. most suitable as examples in a dictionary according to the heuristic) are taken into the resulting corpus. By treating sentences separately, the intersentential context is lost, but this approach makes it possible to sort the corpus by GDEX score and to have all searches GDEX-sorted by default.

The family of SkELL corpora is led by English SkELL, which was released in 2014. Later, Russian (2016), Czech (2017), Italian (2018), German (2018) and Estonian (2018) were added. The English SkELL is used most extensively (150,000 page views per month)⁴. October, November, March and April are the most active months every year (which synchronizes well with academic year cycles).

SkELL corpora can be searched through a simple user interface⁵, which is a simplified version of the CQS Sketch Engine (Kilgarriff et al., 2004). In SkELL's interface, users

³ <https://www.sketchengine.co.uk/user-guide/user-manual/concordance-introduction/gdex/> (3 June 2019).

⁴ Statistics based on Google Analytics (3 June 2019).

⁵ <https://skell.sketchengine.co.uk> (3 June 2019).

can use Sketch Engine’s most important features: concordances, word sketches and similar words (i.e. the thesaurus). Compared to more advanced CQSs, the output in SkELL’s interface is limited: up to 40 sentences and similar words are shown; in word sketches only simplified grammar relation names are presented. The data accessed via SkELL’s interface give a quick overview of examples, word distribution, collocations and the thesaurus.

2. Corpus sentences in the language portal Sõnaveeb

In Sõnaveeb⁶ (Wordweb) – a new language portal of the Institute of the Estonian Language – SkELL corpus data are displayed directly via API from two different Corpus Query Systems. Estonian sentences are queried from the etSkELL 2018 corpus via the CQS KORP API. Russian sentences are queried from the ruSkELL 1.6 corpus via the CQS Sketch Engine JSON API⁷.

Figure 1. Headword *Patarei vangla* ‘Patarei prison’ in Sõnaveeb.

This is the first time in Estonian lexicography that users get direct access to automatically-retrieved authentic sentences. The main motivation was to provide usage examples for headwords that do not have in their entries any example sentences

⁶ <https://sonaveeb.ee/> (3 June 2019).

⁷ <https://www.sketchengine.eu/documentation/json-api-query/?highlight=API> (3 June 2019).

compiled by lexicographers, as is the case with many terms, derived forms, compounds and multi-word expressions (MWEs). Figure 1 shows the MWE *Patarei vangla* ('Patarei prison') in Sõnaveeb, with its definition, and the *hea teada* 'good to know' comment. SkELL sentences are the only usage examples of the word displayed in the bottom right corner of the page.

Figure 2 shows the Russian headword *планета* 'planet' in Sõnaveeb with its domain label *ASTRONOOMIA* 'Astronomy' and definition. SkELL sentences are the only usage examples in Sõnaveeb for the Russian headwords.

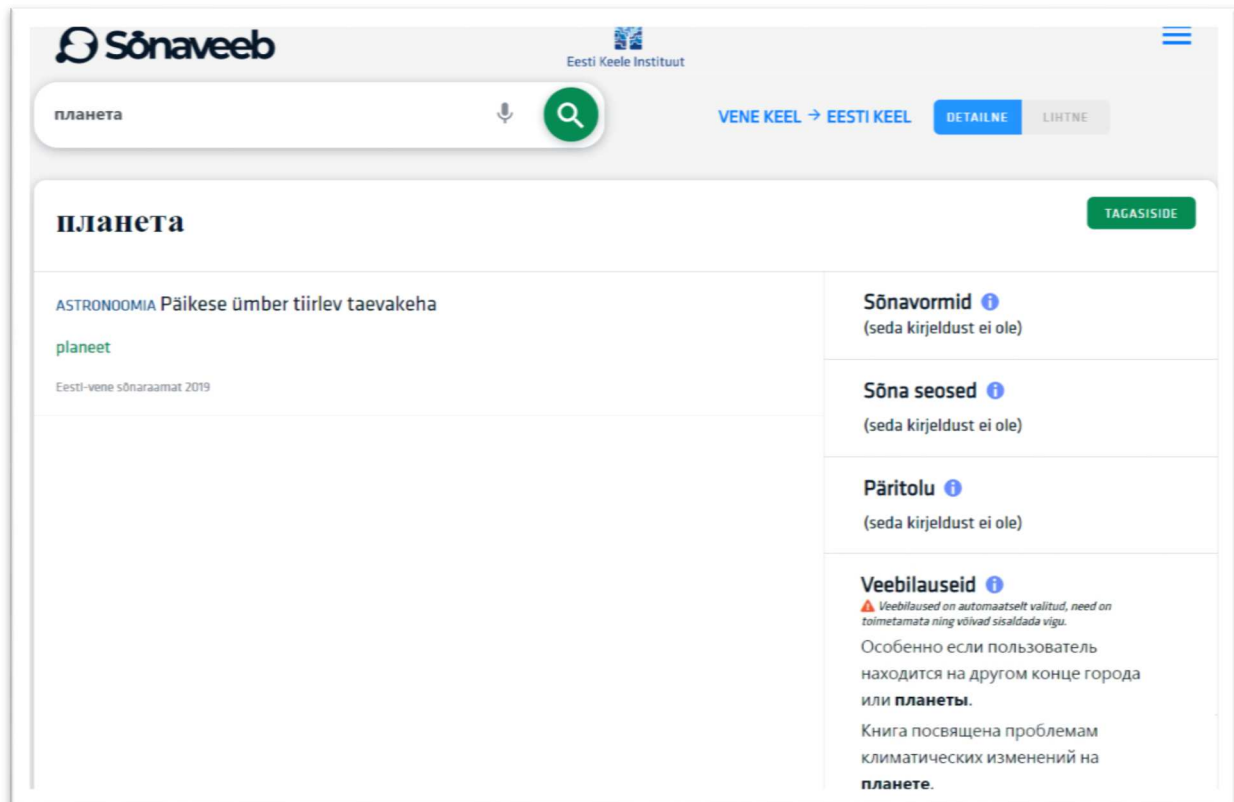


Figure 2. Headword *планета* 'planet' in Sõnaveeb.

Up to 26 sentences are displayed for both languages. Only the first two sentences are displayed by default. The rest of the sentences can be opened by clicking the *Näita rohkem* 'Show more' option. For Russian, sentences are displayed according to GDEX scores: the highest scored sentences are shown first. This was also the case for Estonian sentences at first, but it soon became evident that in some cases the first two SkELL sentences can include errors, e.g. in lemmatization and POS-tagging (see Section 5 for more information). The problem of displaying the same (two) inaccurate sentence(s) for the same headword constantly was solved by displaying random sentences instead. This approach ensures that the presentation of corpus sentences is as dynamic as searches on the web, where the user will get a different set of web sentences each time consulting the same word. Obviously, this approach does not guarantee that the sentences will not contain errors. It still depends a lot on the quality of lemmatization

and morphological analysis.

In the next section, we describe the GDEX configurations for Estonian and Russian used for the compilation of etSkELL 2018 and ruSkELL 1.6 corpora. In Section 4 we present and analyse the results of the evaluation of the Estonian GDEX configuration. In Section 5 we address the main problems with displaying authentic corpus sentences and offer solutions.

3. Good Dictionary Examples (GDEX) and SkELL corpora

Good Dictionary Examples (GDEX) (Kilgarriff et al., 2008) is a function in the Sketch Engine (Kilgarriff et al., 2004) that ranks corpus sentences according to predefined criteria, assigning a numerical score (GDEX score) to each sentence, which separates good candidates from bad ones. This mechanism can be seen as a kind of filter, as it helps lexicographers to work with more relevant citations even though they have not been manually annotated. GDEX scores measure the lexical and syntactical features of the sentence and sort concordances according to how perfectly they meet all the relevant criteria. As a result, GDEX offers a list of example sentences with the best candidates presented first (at the top of the list).

In order to get a list of good example candidates, one needs to define a GDEX configuration that takes into account various criteria, e.g. sentence length and word frequencies (Kosem et al., 2019). The configuration can be seen as a formula that uses a number of parameters⁸:

formula: >

```
(50 * is_whole_sentence() * blacklist(words, illegal_chars))
+ 50 * optimal_interval(length, 10, 14)
* greylist(words, rare_chars, 0.1)
) / 100
```

variables:

```
illegal_chars: ([<|\]\[\>/\^\@]{*\#\#=>«"~_}())
rare_chars: ([A-Za-z0-9A-Я'.,!?)(;:-])
```

The classifier *is_whole_sentence* gives the highest value of 1 to ‘true’ sentences, e.g. ones that begin with a capital letter and end with a punctuation mark (a full stop,

⁸ The given example does not list all of the parameters; the whole description can be found in the manual <https://www.sketchengine.eu/syntax-of-gdex-configuration-files/>.

question mark or exclamation mark). The formula assigns values from 0 to 1. The most appropriate length of a sentence can be described as an argument in the *optimal_interval* classifier. In the above example, it varies from 10 to 14 and assigns the value 1 to such sentences. Along with a formula (which is a mandatory part of the configuration), a user can define the variables. *Illegal_chars* represents a list of characters that a sentence should not have, otherwise it will receive a low score. In the example we put restrictions on meaningless combinations of punctuation marks. *Rare_chars* represent a set of symbols that when they appear in a sentence will receive a penalty. For example, a Russian text in Cyrillic with many Latin characters is not seen as a good example.

GDEX configurations have been developed for several languages (see e.g. Kosem et al., 2019) and can be optimised using a special interface called the GDEX editor⁹ (Figure 3). The GDEX editor is meant for the evaluation of candidate sentences selected according to a GDEX configuration. This system evaluates sentences using two versions of the configuration and assigns two scores and ranks; based on this information, the configuration developers can mark apt sentences and thus assess which set of parameters is more suitable for the task. Writing these formal rules can be seen as an iterative process.

Old rank	Rank	Sentence	Old score	Score	Flag
1	1	На каком этапе брать деньги и сколько – на ваше усмотрение.	0.98	0.98	?
2	2	Помимо платы за рейс по обязательной таксе извозчик имел право брать чаевые до рубля включительно.	0.97	0.97	?
3	4	Поехали самостоятельно, у гида ничего брать не стали.	0.97	0.97	?
4	3	Парамонов . Ну так что, будем брать пример с наших уважаемых отечественных йогов?	0.97	0.97	?
5	6	Можно вместо всего этого одним молоком брать 1 раз в неделю ходить.	0.96	0.96	?
6	8	Нужно ли журналисту брать согласие субъекта на распространение его персональных данных?	0.96	0.96	?
7	9	Вашему партнеру страшно брать на себя ответственность за вас.	0.96	0.96	?
8	7	Для небольших производств, можно не брать, а использовать специальный инструмент для выполнения этой операции.	0.96	0.96	?
9	5	Застройщиков обяжут брать деньги на строительство не у частных инвесторов, а в банках и продавать готовое жилье.	0.95	0.97	?
10	10	Постараюсь брать с Вас пример и поехать знакомиться.	0.94	0.94	?

Figure 3. GDEX editor interface for evaluating candidate sentences for the Russian headword брать ‘to take’.

⁹ <https://gdexed.sketchengine.eu/> (3 June 2019).

Once the GDEX configuration is developed for a particular language, it can be used for the development of a SkELL corpus.

3.1 Parameters of good dictionary examples for Estonian and etSkELL

2018

The first version of the GDEX configuration for Estonian was developed in 2014. It was used for extracting example sentences into the Estonian Collocations Dictionary (ECD) database (Kallas et al., 2015). The ECD is aimed at learners of Estonian as a foreign or a second language at the upper intermediate and advanced levels (CEFR levels B2-C1).

The latest version of the GDEX for Estonian (GDEX 1.4) (Koppel, 2017) was used to compile the Estonian Corpus for Learners 2018 (etSkELL)¹⁰, which is used in both the etSkELL interface¹¹ and in the language portal Sõnaveeb. The process of corpus compilation was two-part: all sentences of the Estonian National Corpus 2017¹² (1.1 billion words) were first filtered using hard classifiers of GDEX 1.4, which resulted in filtering out about 83% of the sentences. The remaining 17% of the sentences were then scored using soft classifiers and compiled into the etSkELL 2018 corpus. The corpus consists of sentences from various media texts, fiction and scientific texts, Estonian Wikipedia and Estonian textbooks.

Table 1 gives an example of the volume of the etSkELL 2018 corpus. All occurrences of a token are accounted for in the structure size, while the lexicon size consists of a count of unique items.

etSkELL structure sizes		etSkELL lexicon sizes	
sentences	24,811,421	lower-case words	1,853,989
words	248,203,200	lower-case lemmas	813,498

Table 1. etSkELL 2018 corpus structure and lexicon sizes.

The parameters of GDEX 1.4 (Koppel, 2017) were fine-tuned based on the analysis of two datasets: one containing selected sentences from the examples of ECD offered by the original GDEX configuration, and one containing rejected or non-selected sentences

¹⁰ DOI: 10.15155/3-00-0000-0000-0000-07335L

¹¹ <https://etskell.sketchengine.co.uk/> (3 June 2019).

¹² DOI: 10.15155/3-00-0000-0000-0000-071E7L

of ECD. The most important parameters are described below.

- **Sentence length.** The average length of an example sentence in ECD is 9 to 10 tokens. Whereas three-word sentences are frequently used in Estonian and are very common in Estonian learners' dictionaries (based on the analysis of example sentences in the Basic Estonian Dictionary (BED) (2014)), the allowed sentence length is set at 4–20 tokens. The optimal interval is set at 6–12 tokens.
- **Word length.** The average word length of the sentences in ECD is six characters. As Estonian has a rich word formation system and some compounds can be quite long, e.g. *kiiruisutamismeistrivõistlused* 'speed skating championships' (30 characters), the maximum word length is set at 20 characters.
- **Low frequency words.** Two different classifiers in the Estonian configuration refer to low frequency words. Firstly, no word forms with a frequency of less than five are allowed in the examples. Following the example of Slovene configuration (Kosem et al., 2013), a classifier penalizing lemmas with a frequency of less than 1,000 was added. This classifier also helps to reduce the probability of complex compounds and rarer proper names occurring in the top ranked examples, which often happened with the previous configuration.
- **Number of elements in the sentence.** The average number of occurrences of certain elements (commas, numerals, proper names, adverbs, verbs, pronouns and conjunctions) in the sentences was determined. Each of the listed elements are grouped together in a classifier in GDEX 1.4, and they share the same weight due to shared characteristics. As a result, if a sentence includes more than one adverb, one pronoun, one proper name, one numeral, one conjunction, one comma, or two verbs, it gets penalized.
- **Sentence initial tags.** 54% of the selected sentences in ECD start with a substantive, 12% with an adjective, 11% with a pronoun and 8% with a verb. None of the selected sentences start with an interjection, abbreviation, genitive attribute or punctuation mark; hence sentences starting with the previously listed tags get heavily penalized.
- **Sentence initial words and word sequences.** Certain words and two-word sequences are not allowed to occur at the beginning of sentences. These are mostly anaphoric words and word sequences that refer to previous sentence(s) and are therefore context dependent, e.g. *pigem* 'rather', *teisisõnu* 'in other words', *seda enam* 'even more' and *teisest küljest* 'on the other hand'.
- **Non-finite constructions.** In order to avoid syntactically complex sentences, certain non-finite constructions are penalized. These constructions occur often, for example, in bureaucratic jargon and formal style, which can be difficult to understand for language learners.

- **Weights.** In GDEX 1.4, weights are assigned to soft classifiers. Optimal interval and word frequency have turned out to be the most distinguishing features of good examples, followed by penalizing anaphors (including certain pro-adverbs and demonstrative pronouns), so they are assigned the highest weight.

The results of the evaluation of the GDEX 1.4. are presented in Subsection 4.1.

3.2 Parameters of good dictionary examples for Russian and ruSkELL 1.6

The first version of Russian GDEX configuration GDEX 1.1. (Apresjan et al., 2016) was used for the compilation of ruSkELL 1.5¹³. However, the preliminary evaluation revealed that ruSkELL 1.5 still contained quite long sentences (up to 150 words). Some sentences did not begin with capital letters; there were also one-word sentences, and sentences containing obscene lexis. It was decided to develop the next version of GDEX configuration 1.2., which would partially solve these problems. GDEX 1.2 for Russian was used for the compilation of the ruSkELL 1.6 corpus, which has been implemented for querying sentences in the language portal Sõnaveeb. It was made on the basis of ruSkELL 1.5, and just re-sorted with the new GDEX 1.2 configuration, favouring average-length sentences with mid-frequency words which are more suitable for learners. Only the top 68 million sentences were used, providing the corpus with 975 million words, or 1,224 million tokens (see Table 2 for details).

ruSkELL 1.6 structure sizes		ruSkELL 1.6 lexicon sizes	
sentences	68,224,440	lower-case words	7,810,025
words	975,584,449	lower-case lemmas	7,403,227

Table 2. ruSkELL 1.6 corpus structure and lexicon sizes.

When writing GDEX configuration rules for Russian, we used several restrictions in order to get more precise results (i.e. more readable sentences).

The most important parameters are described below.

- **Sentence length.** When it comes to the selection of good dictionary examples,

¹³ <https://www.sketchengine.eu/ruskell-examples-and-collocations-for-learners-of-russian/> (3 June 2019).

readability should also be taken into account. According to the Russian Frequency Dictionary (Sharoff¹⁴), the average sentence length is 10.38 words. An analysis of dictionary entries in MAS (Jevgen'jeva, 1981-1984) showed that citations are longer and consist of 13 words. We came to the conclusion that the optimal sentence length would vary from 7 to 16 tokens and thus the allowed length was set at 6–20 tokens.

- **Blacklists.** We defined a number of variables that impose restrictions on sentence content. We filtered out emoticons and other combinations with punctuation marks (e.g. slashes, parentheses and quotes); they should not be used in “good” sentences. But not only characters can lower GDEX scores. An obscene lexicon should not be used in corpora for learners, and thus one of the blacklists includes such words.
- **Greylists.** Unlike the previous lists, greylists describe the elements whose presence in a sentence leads to lower scores. For Russian texts, we had to limit the usage of Latin characters and of words written in capitals. Such sequences may include trademarks, company or other proper names that would not make much sense to language learners. Also, the presence of digits can be seen as a drawback in a sentence.
- **Sentence initial words and word sequences.** Following the Estonian configuration (Koppel 2017), we also prepared a list of words and word sequences that are not allowed to appear at the beginning of sentences. On the one hand, as was stated above, such elements have mostly an anaphoric nature and refer to previous sentences. On the other hand, they can be a trace of a formal language that we prefer to avoid in the corpus, e.g. *vo-pervykh* ‘firstly’, *dalee* ‘then’, *todga* ‘hence’, *sleduet otmetit'* ‘it should be noted’ and *kak sledstvie* ‘as a result’.

The evaluation of the GDEX 1.2 configuration for Russian has not been carried out yet, but will occur soon.

4. Users’ attitudes towards authentic corpus sentences

4.1 Evaluation of the GDEX 1.4. configuration for Estonian

In 2019 an evaluation (Koppel, 2019b) of the GDEX 1.4 output was completed by students of Tallinn University and the University of Tartu who speak Estonian at the B2–C1 proficiency levels, and by lexicographers working at the Institute of the Estonian Language. The purpose of the evaluation was to determine whether, according to the

¹⁴ <http://www.artint.ru/projects/frqlist.php> (3 June 2019).

two types of evaluators, authentic and unedited corpus sentences would be suitable example sentences in the language portal.

The GDEX 1.4 output evaluation consisted of two tasks. The first assessment task involved using the open source platform Pybossa¹⁵, which is used to carry out simple crowdsourcing projects and analyse the data collected. The aim of the first assessment task was to rate the suitability of sentences in general. The follow-up assessment task was performed in the Google Forms environment, and its purpose was to identify the reasons why the evaluators considered certain sentences not suitable for the dictionary.

For evaluation, we selected 40 random headwords from the ECD, the dictionary aimed at B2-C1 level learners: ten for each part of speech (substantive, verb, adjective and adverb). Then we took a random selection of sample sentences for each headword which included:

- one corpus sentence that meets the criteria of GDEX 1.4;
- one corpus sentence that does not meet the criteria of GDEX 1.4;
- one unfiltered corpus sentence;
- one example sentence compiled by a lexicographer.

All corpus sentences were taken from the Estonian National Corpus 2017; the dictionary example was taken from the Dictionary of Estonian 2019 (DicEst).

In the first assignment, there were 160 sentences in total, which were divided into four smaller tasks, in which each assignment contained all four types of sentences. The assessment task in Pybossa was set up so that each sentence had to be rated by five different lexicographers and five different language learners. The task was sent to seven lexicographers and 31 students, of whom five lexicographers and nine language learners responded (when one sentence had been rated by five different lexicographers and five different language learners, it was no longer displayed for the next evaluator).

Language learners were asked to assess the sentences based on their Estonian language skills; lexicographers were asked to assess if the sentences were suitable for a dictionary aimed at learners at the B2-C1 language proficiency levels.

One sentence was displayed to the evaluators at a time, preceded by the question “Is this sentence suitable as an example of the word X?” The response options were “yes”, “no” and “I don't know”. Neither the definition nor the source of the sentence was displayed to the evaluator (Figure 4).

¹⁵ <https://pybossa.com/> (3 June 2019).

Kas see lause sobib sõna **inimtühi** näitelauseks?

Inimtühjal tänaval võib keegi sulle sama nähtamatult, nagu on helkurvestita politseinik, joosta sebrale.

Jah Ei Ei oska hinnata

Lahendad praegu ülesannet number **1**. Oled lahendanud **0** ülesannet **160** -st.
 Sa peaksid lahendama **40** ülesannet.
 Kui sul tekib mingeid kommentaare, siis täida tagasiside [küsimustik](#).

Figure 4. Sentence assessment task in Pybossa

While the purpose of the first assessment task was to establish quantitatively whether different types of sentences were in the lexicographers' and language learners' opinions suitable example sentences in a learner's dictionary, the purpose of the follow-up survey was to identify why evaluators considered some good corpus sentences not suitable and some bad corpus sentences suitable. For this reason the evaluators were asked to re-assess three types of sentences in the follow-up survey:

- Corpus sentences that met all the criteria of GDEX 1.4 but most evaluators did not think were suitable (or they did not know).
- Corpus sentences that did not meet the criteria of GDEX 1.4 but most evaluators thought were suitable (or they did not know).
- Dictionary examples that most evaluators did not think were suitable (or they did not know).

The request to participate in the follow-up survey was sent to the same evaluators who had participated in the first assessment task (five lexicographers and nine language learners), and we received replies from five lexicographers and five learners. In the follow-up survey, lexicographers were asked to re-evaluate 18 of the previously mentioned three types of sentences, and language learners were asked to re-evaluate 20 sentences, of which 11 sentences overlapped.

The final results of the two assessment tasks showed that, according to most lexicographers and language learners, as many as 96% of the dictionary examples and 85% of corpus sentences chosen as good examples by GDEX 1.4. were considered to be suitable example sentences. Only 6% of the sentences that were discarded by GDEX 1.4 were considered suitable, meaning that 94% of the bad candidates had been filtered out successfully. As for unfiltered corpus sentences, 60% of those were considered

unsuitable. When evaluators were asked their reasons for considering a sentence unsuitable, the most common arguments were that the sentences included anaphora and hence needed more context, or that the sentences were colloquial, too long or too short.

The results of the evaluation show that even more attention should be paid to anaphora, either by raising the penalty or by adding more words to the greylist. It also makes sense to invest more effort into figuring out the ideal range of sentence length, as short sentences tend to lack context and long sentences were mostly considered unsuitable.

4.2 User feedback on the presentation of corpus sentences in Sõnaveeb

Dictionary users are accustomed to the fact that all data presented in a dictionary are controlled and edited by a lexicographer, and are hence correct. In contrast, corpus sentences in Sõnaveeb are authentic, unedited and may include errors. Since Sõnaveeb's launch in February 2019, the lexicographers working in the Institute of the Estonian Language have received feedback from users in which they have said that they find some of the corpus sentences are inappropriate or incorrect. At the beginning, no clear warning of the authenticity of the sentences was displayed by default. The user could only read the information about the source of the sentences (etSkELL 2018 corpus/ ruSkELL 1.6 corpus) by moving the cursor over the information button. After receiving user feedback, the editors of Sõnaveeb decided to use the same strategy as in Merriam-Webster's¹⁶ and Collins'¹⁷ dictionary portals, and added an explicit note saying that the sentences were chosen automatically, were unedited and might contain errors. The user feedback also indicated that users, especially language professionals, want to see the metadata of each sentence, e.g. author, title, and year. This information would indicate whether the word is archaic, colloquial, to which genre it belongs to, etc.

5. Problems and possible solutions

Several problems have arisen with displaying authentic corpus sentences, and it is difficult to eliminate them with the help of a tool operating solely on a rule-based method. Some of these problems are language independent, and some are language specific. The most typical problems are described below.

1. Polysemous words. When choosing the sentences, the polysemy of the headword is not taken into account. For example, the query for the Estonian polysemous headword *leht* ('newspaper', 'leaf', 'webpage') provides sentences in which the word occurs in different meanings.

¹⁶ <https://www.merriam-webster.com/> (3 June 2019).

¹⁷ <https://www.collinsdictionary.com/> (3 June 2019).

2. Lexical and POS-homonymy. When choosing the sentences, the homonymy of the headword is not taken into account. For example, the query for Estonian homonymous headword *tamm* ('aok'; 'dam'; 'king') provides sentences in which the word occurs in different meanings.
3. Lemmatization and POS-tagging errors. These arise particularly in the case of grammatical homonymy. For example, the query for the Estonian grammatical homonym *joon* (*joon-n* 'stripe-Substantive', *joon-v* '(I) drink-Verb', provides sentences in which this word occurs as a noun in nominative case, as well as the first person singular of the verb *jooma* 'to drink' in present indicative. The Russian grammatical homonym *дома* (*дом-n* 'house- Substantive', *дома-d* 'at home- Adverb') can be either the genitive singular or the nominative/accusative plural of the noun *дом* 'house' or the adverb *дома* 'at home'. The query gives examples for both lemmata without distinguishing between them.
4. Machine-translated sentences. Machine-translated sentences get crawled from bilingual web pages that match the predefined parameters of GDEX but may sometimes be ungrammatical.
5. Absence of information for low frequency words. It is difficult to find example sentences for low frequency words. For example, the noun *kalla* 'arum lily' does not appear in the etSkELL 2018 corpus.
6. Multiword expressions. The selection includes sentences where the headword is actually part of an MWE, e.g. in the output for the keyword *tulema* 'to come', sentences with the MWE *toime tulema* 'to manage' might appear.
7. A certain type of problem comes from the source texts (either mistakes, typos or errors of recognition), e.g. the Russian *па* instead of the preposition *на* 'on' (here we should note that in Cyrillic they have similar graphic forms: *па* vs *на*). Hence we can try to filter out such examples, defining a separate blacklist for typical errors.
8. Along with Russian, there are other Slavic languages (Ukrainian and Belorussian) that use Cyrillic. Although the corpus was cleaned up with the right encoding, it still has irrelevant examples in other languages. One of the possible solutions is to prepare a list of frequent words in Cyrillic that are not Russian, in order to filter out such sentences.

Finding suitable example sentences for different meanings of polysemous words could have been facilitated if the corpus had been semantically annotated and queries could be based on using the same semantic types as used in the dictionary. The semantic types developed by Margit Langemets (2010) that have been used in the compilation of BED and DicEst could be applied for the Estonian language.

An additional way of solving the problem of polysemy and lexical homonymy is to consider the collocations of the headword, so that the example sentences with the most frequent collocations appear in the output. For Estonian, the database of the Estonian Collocations Dictionary could be incorporated. For example, if the headword *tamm* ('oak'; 'dam'; 'king') has three homonyms and the user chooses the meaning of 'dam', the sentences with the collocations *tamm puruneb* 'the dam collapses' and *tammi ehitama* 'to build a dam' would appear first.

One other possible way of solving the problem of POS-homonymy (e.g. *noor-a* 'young-Adjective' and *noor-s* 'young person-Substantive'), is to query sentences via API using lempos instead of lemma. This is already done in Sõnaveeb. It helps in cases of POS homonyms, but becomes an obstacle in the case of errors in lemmatization and POS-tagging. It is a very frequent problem, especially in the case of grammaticalization and lexicalization, when a morphological analyser defines lemma and POS on the basis of an outdated lexicon. For example, the headword *tasuta-d* 'for free-Adjective' used to be analysed as the substantive *tasu* 'fee' in abessive case in dictionaries, and is still analysed as a substantive by POS-taggers. But all dictionaries published in Sõnaveeb consider it to be an adjective. Since we query sentences via API using lempos, the system does not find such a lempos *tasuta-a* 'for free-Adjective' and the query shows completely erroneous results.

Detection of machine-translated texts seems to be a topic in its own right (see e.g. Aharoni et al., 2014; Nguyen-Son et al., 2019 for more). In order to avoid automatically translated sentences occurring in the output, machine generated texts should be automatically detected and rejected at the stage of corpus crawling. The best way to do that is to combine multiple approaches. Firstly, there is a need to identify problematic sites and remove them completely from the corpus. If such sites are already known from previous corpus crawling, the crawler can be instructed to avoid them altogether in crawling the new corpus. Secondly, the crawling should start from trustworthy sources. It is also important to keep track of the distance of a site from these trustworthy sources¹⁸. According to our experience, sites having too long names could be avoided¹⁹. Documents coming from web sites not available one month after crawling from the corpus should be removed²⁰. It would also help to build a classifier that recognizes computer-generated texts. For languages where syntactic analysis is possible, it can be used to reveal suspicious sentences. But even in this case some problems remain unsolved: 1) a large part of human-produced text uses unorthodox

¹⁸ For example, www.eki.ee is a seed with distance 0; a site referenced by www.eki.ee, has a distance of 1; a site referenced by sites with distance n but not sites closer to seeds than n has a distance of n + 1.

¹⁹ Too long is ≥ 40 characters (or ≥ 50 characters to reduce false positives) according to our unpublished experience with reviewing the content of random sites with long names.

²⁰ According to our unpublished experience with checking sources of computer-generated text in the corpus, the life of spam sites is short. The reason may be they become useless once blacklisted by search engines.

syntax, so we don't know what "faulty" is, and 2) neural machine translation produces syntactically perfect sentences, and it is difficult to detect them.

Absence of information for low frequency words is a possible bias in the corpus crawling procedure. It makes sense to combine queries from different sources, e.g. when a word is not found in the (smaller) learner corpus, the query will be made on the basis of a (large) general corpus (e.g. Estonian NC).

In addition, the origin of source texts must be taken into account when creating a new learner corpus: this would make it possible to give priority to sentences from Estonian Wikipedia and periodicals rather than sentences from blogs and forum posts.

In addition, it is obvious that one corpus cannot satisfy the needs of all users. One possibility is to apply additional filters (e.g. vocabulary lists of different language proficiency levels). For Estonian, special GDEX configurations aimed at different CEFR (Common European Framework of Reference) levels of Estonian L2 proficiency (Koppel, 2019a) and CEFR vocabulary lists (Kallas & Koppel, 2018a, 2018b, 2018c) have been developed, but have not yet been used to compile SkELL corpora for different CEFR levels.

6. Conclusions and Future Work

In this paper we analysed different issues that are connected with the quality and presentation of authentic corpus sentences as a part of an (academic) language portal. Most recent dictionaries in Estonia are corpus-based but have traditionally not included authentic corpus data in their online versions. Sõnaveeb is the first of its kind where its users can read authentic corpus sentences without leaving the language portal's interface.

The important question is what kind of corpora are more suitable for this purpose. In the paper we argue that one type of corpora that might be used is Sketch Engine for Language Learning, or SkELL corpora. SkELL corpora were initially intended for learning purposes but they can be seen as a source of good dictionary examples, as they contain only sentences which are ranged as 'good' according to the GDEX system. In order to compile such corpora, Good Dictionary Examples (GDEX) (Kilgarriff et al., 2008) configurations for Estonian (GDEX 1.4.) and for Russian (GDEX 1.2) were developed and later used for the compilation of etSkELL 2018 and ruSkELL 1.6 corpora, which are used as sources of authentic corpus sentences in the new language portal Sõnaveeb. Estonian sentences are queried from the etSkELL corpus via the Corpus Query System Korp API. Russian sentences are queried from the ruSkELL 1.6 corpus via the Corpus Query System Sketch Engine JSON API.

The evaluation of the GDEX 1.4 configuration for Estonian showed that as many as 85% of corpus sentences chosen as good examples by GDEX 1.4 were also evaluated as "good" by users (lexicographers and Estonian language learners). Only 6% of the

sentences that were discarded by GDEX 1.4 were considered suitable, meaning that 94% of the bad candidates had been filtered out successfully. The main reasons for considering a corpus sentence unsuitable were anaphora, colloquialisms and sentence length (too long or too short). User feedback has also revealed that some users get confused if they see inappropriate or incorrect sentences as a part of the portal. For this reason it was decided to add a clear note that says that the sentences are chosen automatically and that they may contain errors. Some users also pointed out that they need the description of a source of sentences, e.g. author, title, and year. These parameters help to understand whether a word is archaic, colloquial, which genre it belongs to, etc.

However, there are also problems that are difficult to eliminate solely by using the GDEX system. These are lemmatization and POS-tagging errors in corpus data, homonyms, polysems, low frequency words, sentences with inappropriate content, machine-translated sentences etc. Finding suitable example sentences for different meanings of polysemous words could have been facilitated if the corpus had been semantically annotated and queries could be based on using the same semantic types as used in the dictionary. One way of solving the problem of polysemy and homonymy is to consider the headword's typical collocates, so that the example sentences with the most frequent collocates of the headword appear first in the output. The problem of POS homonymy can be solved by querying sentences via API on the basis of lemmas instead of lemmas.

Some problems with mistakes and inconsistencies in corpora can be solved during the compilation. Various reference databases can be applied, e.g. a database of common spelling mistakes and a database of frequent foreign or dialectal lexis.

For the purpose of customization, there is a need to compile special SkELL corpora for each CEFR (Common European Framework of Reference) level. Special GDEX configurations aimed at different CEFR levels of Estonian L2 proficiency (Koppel, 2019a) have already been developed. This will allow us to show different sets of sentences for users with different Estonian L2 proficiencies.

In order to facilitate the development of GDEX configurations and to decrease the number of incorrect and/or inappropriate sentences shown in Sõnaveeb, we plan to use crowdsourcing methods. Users will be given an opportunity to vote on each sentence via the portal, and after a sentence receives a certain number of downvotes, the system would not show them again, while the upvoted sentences could be displayed first. This approach will help to create two datasets – one with upvoted sentences and the other with downvoted sentences – which then could be used for the development of learning algorithms (as patterns or features). Implementing crowdsourcing would also make the language portal more interactive.

7. Acknowledgements

The creation and development of the portal was funded by the Digital Focus Program of the Ministry of Education and Research (2018–2021) and by EKI-ASTRA program (2016–2022). The creation of the dictionary and terminology database Ekilex was funded by EKI-ASTRA program (2016–2022). Software development has been provided by OÜ TripleDev.

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarín infrastructure LM2015071. This publication was written with the support of the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic.

This work was supported by the grant of the President of Russian Federation for state support of scholarly research by young scholars (Project No. MK-2513.2018.6).

The research received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

8. References

- Aharoni, R., Koppel, M. & Goldberg, Y. (2014). Automatic detection of machine translated text and translation quality estimation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 289-295.
- Apresjan, V., Baisa, V., Buiivolova, O., Kultepina, O. & Maloletnjaja, A. (2016). "RuSkELL: Online Language Learning Tool for Russian Language." *Proceedings of the XVII EURALEX International Congress*, Tbilisi, Georgia, pp. 292-299.
- Baisa, V. & Suchomel, V. (2014). SkELL: Web Interface for English Language Learning. *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Tribun EU, 2014, pp. 63-70.
- BED = *Eesti keele põhisõnavara sõnastik 2019* (1. trükk 2014). [The Basic Estonian Dictionary 2019, BED] Eesti Keele Instituut. Sõnaveeb 2019 [Wordweb 2019]. Available at: <https://sonaveeb.ee> (3 June 2019).
- Cook, P., Rundell, M., Lau, J. H., & Baldwin, T. (2014). Applying a word-sense induction system to the automatic extraction of diverse dictionary examples. *Proceedings of the XVI EURALEX International Congress*, pp. 319-328.
- ECD = *Eesti keele naabersõnad 2019*. [The Estonian Collocations Dictionary 2019] Eesti Keele Instituut. Sõnaveeb 2019 [Wordweb 2019]. Available at: <https://sonaveeb.ee> (3 June 2019).
- DicEst = *Eesti keele sõnaraamat 2019*. [The Dictionary of Estonian 2019] Eesti Keele Instituut. Sõnaveeb 2019 [Wordweb 2019]. Available at: <https://sonaveeb.ee> (3 June 2019).
- etSkELL 2018 = Sketch Engine for Estonian Language Learning 2018. Accessed at:

- <https://etskell.sketchengine.co.uk/> (3 June 2019)
- GDEX editor*: Accessed at <https://gdexed.sketchengine.eu/> (3 June 2019)
- Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M. & Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. In I. Kosem et al. (eds.) *Proceedings of the eLex 2015 conference*, Herstmonceux Castle, United Kingdom, pp. 1-20.
- Kallas, J., Koeva, S., Kosem, I., Langemets, M. & Tiberius, C. (2019). *Lexicographic practices in Europe: a survey of user needs*. Available at: https://elex.is/wp-content/uploads/2019/02/ELEXIS_D1_1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf (3 June 2019).
- Kallas, J. & Koppel, K. (2018a). *Eesti keele B1-taseme sõnavara*. [B1 Estonian Vocabulary List.] Eesti Keele Instituut.
- Kallas, J. & Koppel, K. (2018b). *Eesti keele A2-taseme sõnavara*. [A2 Estonian Vocabulary List.] Eesti Keele Instituut.
- Kallas, J. & Koppel, K. (2018c). *Eesti keele A1-taseme sõnavara*. [A1 Estonian Vocabulary List.] Eesti Keele Instituut.
- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of the 13th EURALEX International Congress*. Barcelona: Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra, pp. 425-432.
- Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX international congress*. Lorient, France: Université de Bretagne Sud, pp. 105-115.
- Koppel, K. (2017). Heade näitelausete automaattuvastamine eesti keele õppesõnastike jaoks [Automatic detection of good dictionary examples in Estonian learner's dictionaries]. *Eesti Rakenduslingvistika Ühingu aastaraamat* [Papers in Applied Linguistics], 13, pp. 53-71. DOI:10.5128/ERYa13.04.
- Koppel, K. (2019a). Eesti keele kui teise keele õpikute lausete analüüs ja selle rakendamine eri keeleoskustasemete sõnastike näitelausete automaatsel valikul. [Parameters of CEFR-graded coursebook sentences and their use for automatic detection of good dictionary examples]. *Eesti Rakenduslingvistika Ühingu aastaraamat* [Papers in Applied Linguistics], 15, pp. 99-119. DOI:10.5128/ERYa15.06.
- Koppel, K. (2019b). Leksikograafide ja keeleõppijate hinnangud automaatselt tuvastatud korpuslausete sobivusele õppesõnastiku näitelauseks [Suitability of automatically selected example sentences for learners' dictionaries as tested on lexicographers and language learners]. *Lähivõrdlusi. Lähivertailuja*, 29, [forthcoming].
- KORP*: Accessed at: <https://korp.keeleressursid.ee/> (3 June 2019).
- Kosem, I., Gantar, P. & Krek, S. (2013). Automation of lexicographic work: An opportunity for both lexicographers and crowd-sourcing. In I. Kosem et al. (eds.) *Proceedings of the eLex 2013*, Tallinn, Estonia, pp. 17-19.
- Kosem, I., Koppel, K., Zingano Kuhn, T., Michelfeit, J. & Tiberius, C. (2019).

Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, 32(2), pp. 119-
<https://doi.org/10.1093/ijl/icy014>.

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.

LDOCE = Longman Dictionary of Contemporary English. Accessed at: <http://ldoce.longmandictionariesonline.com/main/Home.html> (3 June 2019).

Nguyen-Son, H-Q., Thao, T. P., Hidano, S. & Kiyomoto, S. (2019). Detecting Machine-Translated Paragraphs by Matching Similar Words. arXiv preprint arXiv:1904.10641.

ruSkELL1.6: <https://www.sketchengine.eu/ruskell-examples-and-collocations-for-learners-of-russian/> (3 June 2019).

Sketch Engine. Accessed at: <https://www.sketchengine.eu/documentation/api-documentation/> (3 June 2019)

Sõnaveeb = Sõnaveeb 2019 [Wordweb 2019]. Accessed at: <https://sonaveeb.ee> (3 June 2019)

Wordnik: Accessed at: <https://www.wordnik.com/> (3 June 2019).

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>

