Aggregating Dictionaries into the Language Portal Sõnaveeb: Issues With and Without Solutions

Kristina Koppel, Arvi Tavast, Margit Langemets, Jelena Kallas

Institute of the Estonian Language, Estonia
E-mail: kristina.koppel@eki.ee, arvi@tavast.ee, margit.langemets@eki.ee, jelena.kallas@eki.ee

Abstract

In this paper we present Sõnaveeb, a new type of language portal of the Institute of the Estonian Language containing data from a growing number of dictionaries and termbases. Sõnaveeb currently displays a total of 200,000 Estonian headwords, obtained from many databases, with many new types of lexicographic information: collocations, etymology, multi-word expressions, etc.
The paper reports on problems encountered so far: the consistency of information and avoiding duplicates when unifying the dictionaries, turning dictionary-specific information into customizations of the central service, deciding on deliberate ambiguities, parsing data fields containing more than one data element, including textual condensation, moving from annotating form (e.g. italics) to annotating content (e.g. a citation), moving from (near) duplicates to sensible information fragments, deciding between an app and a responsive web page, and possible legal problems regarding the authorship of the new central resource, as it may become difficult to show who authored which part of the published resource.
The development of Sõnaveeb continues in the direction of both the tighter aggregation of existing datasets and the addition of new data from other dictionaries and termbases, as well as compiling new data in the new DWS Ekilex.

Keywords: lexicographic database; data aggregation; unified dictionary; Dictionary Writing System; user needs; Estonian

1. Introduction

Sõnaveeb[1] is the new language portal of the Institute of the Estonian Language containing the linguistic information from a growing number of dictionaries and databases. Sõnaveeb was released in February 2019 and presented with the publishing of two new dictionaries, The Dictionary of Estonian 2019 (DicEst) and Estonian Collocations Dictionary 2019 (ECD). In addition, The Basic Estonian Dictionary 2019 (BED) (1st ed. 2014), intended for beginner and advanced language learners, can be used here, as well as two bilingual dictionaries, the Estonian-Russian Orthographic Dictionary for Students 2019 (1st ed. 2011) and the Estonian-Russian Dictionary 2019 (1st ed. 1997–2009), updated with 10,000 new headwords. Special morphological

[1] https://sonaveeb.ee/ (20 May 2019). Sõnaveeb can be translated into English as Wordweb. It is important to emphasize that it is the language portal, not an ontology.

datasets serve to present morphophonological data for Estonian. The portal contains about 200,000 words and phrases in Estonian and about 70,000 words and phrases in Russian.

The information displayed in Sõnaveeb comes from Ekilex[2] (Tavast et al., 2018), a Dictionary Writing System maintained and developed by the Institute in collaboration with the software company TripleDev. As of May 2019, Ekilex contains over 50 lexical datasets: general as well as specialized dictionaries. Databases are constantly updated and edited, including changes that are made upon receiving feedback from users. Created data is stored in Ekilex's PostgreSQL database. Ekilex is hosted in the Estonian Scientific Computing Infrastructure (ETAIS) cloud. Archive copies of data are also stored in the Center of Estonian Language Resources' repository Entu[3]. The metadata on created resources is available in the META-SHARE[4] repository. Upon creating a metadata entry in META-SHARE, a DOI is assigned to each resource.

A new version of the portal is created and archived once a year. Each version is marked by the year and has the date of its creation, e.g. Sõnaveeb 2019 (14.02.2019).

In the next sections we discuss the list of issues, whether they are already solved, in the process of being solved, or lack a known solution. Undoubtedly, there will be more exciting challenges in the near future as we continue to import new data. Several issues are very much in line with the objectives and outcomes of the Horizon 2020 project ELEXIS (European Lexicographic Infrastructure) [5] developing strategies for extracting, structuring and linking of lexicographic resources.

2. Internet skills and organizing the presentation of data

The Sõnaveeb user interface has two different modes of information display for different types of users: advanced and simple. Robert Lew (2013) has stated that web users tend to resort to very simple strategies for internet-based information retrieval, and that users' general tendency is to gravitate towards natural-language queries. The bad news is that "end-users tend not to change the default settings of an information retrieval system" (Markey, 2007: 1077, cited by Lew, 2013). Online dictionaries should somehow cope with unsophisticated strategies of general web use. We agree with Lew (2013: 29):

> This is a conclusion that many lexicographers find hard to accept, and an argument can be made that a minority of expert users (such as language professionals) are worth catering for as well. Ideally, an online dictionary interface

[2] https://ekilex.eki.ee/ (20 May 2019).

[3] https://entu.keeleressursid.ee/ (20 May 2019).

[4] http://www.meta-share.org/ (20 May 2019).

[5] https://elex.is/objectives/ (20 May 2019).

> will combine simplicity (for those who cannot be bothered) with sophistication (for those who can). A reasonable way to achieve this is to offer a simple default interface with an optional advanced alternative.

In Sõnaveeb, we try to combine simplicity with academic sophistication and trustworthiness. As the system has mostly been developed in cooperation with lexicographers, not laymen, we tend to prefer lexicographers' cultivated taste. However, we have conducted some user interviews on particular topics, e.g. synonyms, parts of speech and web sentences, and we are willing to use this information to present our data in a better, i.e. more flexible way.

2.1 Advanced mode vs. simple mode

Modes are used to filter data. The user can currently choose between two modes of information display: advanced or simple. The advanced mode is intended primarily for native speakers. It displays all the information on a word that comes from different sources. The advanced mode is a sophisticated view that might require more options for further filtering. At present we are working on the inclusion of prescriptive data (from the prescriptive Dictionary of Standard Estonian (ÕS 2018), in order to present both descriptive and prescriptive data. This is a challenge, as there have been quite a number of data conflicts from the user's perspective in parallel separate online dictionaries (the descriptive DicEst vs. prescriptive ÕS).

The simple mode is intended primarily for learners at the A2–B1 proficiency levels. It shows 5,000 basic Estonian words (headword list of the Basic Estonian Dictionary (BED); see Kallas et al., 2014) and information is presented in a simpler way: the definitions are shorter, knowledge is organized using controlled vocabulary, there is explicit information about the most frequent morphological forms, etc.

2.2 Choosing languages

As of May 2019, lexical data is available for two languages: Estonian and Russian, each as both source and target language. The list of languages is planned to be increased as there are more bilingual databases available at our Institute.

2.3 Mobile app or responsive web page?

Sõnaveeb.ee is a responsive web page with the same information content for both mobile and desktop resolutions. Around 73% of traffic is desktop, while 25% is mobile and 2% is tablet usage. There are around 22,000 monthly and 2,000 daily active users. 56% are new and 44% returning visitors (Google Analytics, 30 May 2019).

The most frequent question since opening the Sõnaveeb website in February 2019 has been: Will there also be an app? No, for the following reasons:

- The web is better for reaching a wider audience, especially if dictionary use is as sporadic as shown by the high ratio of new visitors. Users cannot be expected to install an app that they will only use once.
- As apps are platform-specific, their development and maintenance are currently beyond our financial means.
- Dictionary content is visually simple enough to be presented using web technologies.
- Lexical resources in the form of a website are more easily indexable by search engines. Although we haven't achieved it yet, it is possible to show up in search results for individual words.

3. Aggregation issues

3.1 The Ekilex data model and the unification of dictionaries

The data model of Ekilex has been described in Tavast et al. (2018). For the purposes of the current paper, it is sufficient to note that we have a many-to-many (i.e. n:m) relations between words and meanings. The link table between these two entities is called a lexeme and is defined as "this word in this meaning, as described in this dataset". Words and meanings are dataset-agnostic, allowing a gradual transition from the initial condition of several independent datasets to the end goal of a single Ekilex resource containing all lexical information known about the Estonian language.

The initial import of the separate datasets resulted in massive duplication of both words and meanings (see Figure 1). Each word had at least as many homonyms in the Ekilex resource as there were imported datasets.

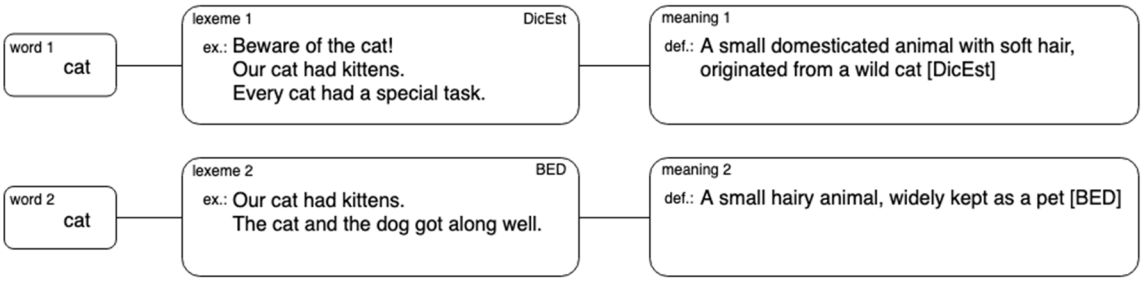

Figure 1. Initial condition: separate dictionaries with duplicate words and meanings

The first step in the transition was the unification of homonyms. Lexicographers manually decided which homonyms were legitimate, and the rest were unified automatically. The result was that there were no longer too many homonyms, but now each word had at least as many senses as there were imported datasets (see Figure 2). The manual effort of unifying the words was relatively small, as there are only about 1,500 legitimate homonyms in Estonian.

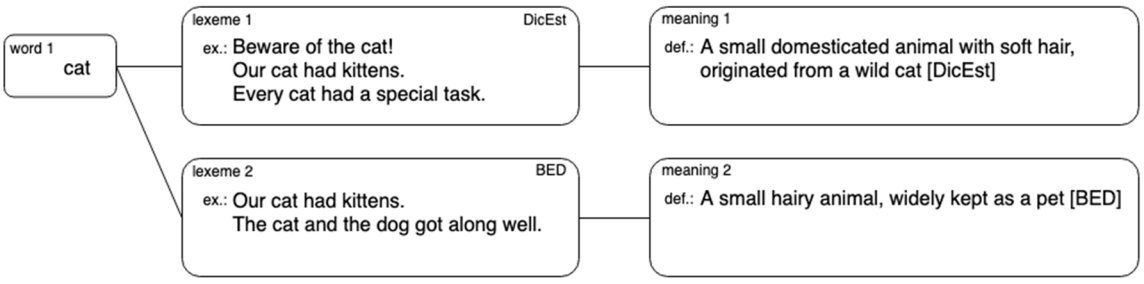

Figure 2. First step in unification: words are unified, but lexemes and meanings are still duplicated

The next step was the manual unification of meanings. The difficulty here is that datasets differ in their sense divisions, often deliberately, depending on the target audience and purpose of the dictionary, so there are no direct correspondences between meanings across datasets. As of May 2019, this work is still ongoing, even for clear cases, and there is no known solution for the unclear ones, unless the solution is to alter the sense divisions of the original datasets. The result for the successfully unified meanings is that there are two lexemes between a word and a meaning, or two statements about the same word-meaning correspondence, see Figure 3.

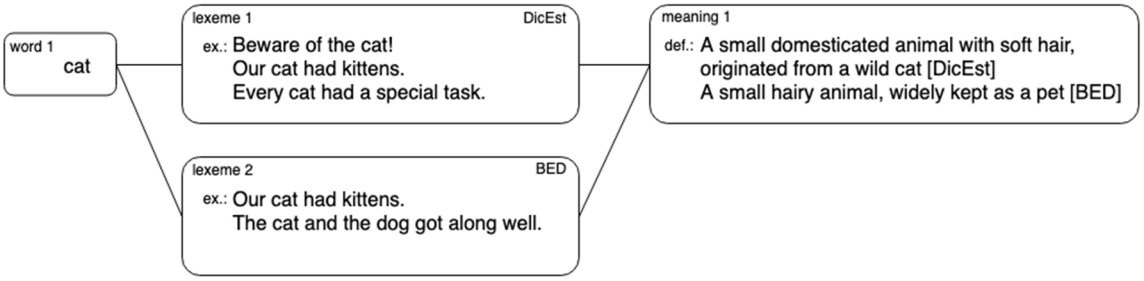

Figure 3. Second step in unification: Words and meanings are unified, each dataset still has its own lexeme between them

Since the Ekilex data model is flattened for display in Sõnaveeb by aggregating lexemes and meanings (this aggregation corresponds to the traditional understanding of word sense), this stage of unification resulted in a very unclear display of information in Sõnaveeb. There were still as many "senses" as imported datasets, but meanings (mainly represented by definitions) were first added together and then repeated under every sense. This was so counter-intuitive for readers that we temporarily disabled version updates of Sõnaveeb, displaying the previous stage instead.

The final stage of unification is still in development. To dispose of the duplicate lexemes they will be added up, with the sum lexeme containing a union of all data elements in lexemes between the same word and the same meaning (see Figure 4).

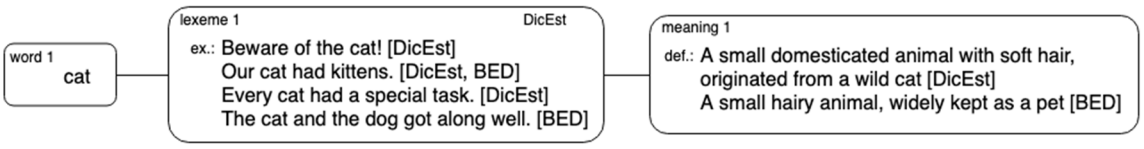

Figure 4. Unified datasets: there are no duplicates in any of the three major entities of the data model

Clear cases of duplication among the first imported datasets will be solved at this step. Work with less clear cases, and especially with more specialized datasets, will continue for a long time. It is also not known yet to what extent such unification is even possible.

3.2 From duplicates to sensible information fragments

Separate datasets have brought into the Ekilex resource several duplicates (or near duplicates) that require special attention. We have to decide whether it is useful or possible to adapt them into information fragments for reuse in other contexts, or if they should just be avoided.

Weitzman (2014) stressed that content management systems must support sensible fragments of information that can be presented in different contexts, e.g. in Ekilex we have "duplicate" information from different datasets for the same meaning (definitions and domain indicators). The task is to be able to describe these different user situations, in which each has its own requirements on the information. These fragments cannot be automatically derived; instead they have to be carefully designed. As the separation of content and presentation has been implemented in Ekilex, we try to reuse the information in the most sensible way, e.g. information from BED has been presented in the simple mode intended primarily for learners at the A2–B1 proficiency levels (see Chapter 2.1).

Sense division in the source datasets (DicEst, ECD and BED) has been manually disambiguated using a specially developed tool. After unifying the senses, we get both long definitions (from DicEst) and short definitions (from BED) that might be presented in different modes: long definitions in the advanced mode, and short/simple definitions in the simple mode. The ongoing migration of senses from separate datasets to a single resource creates several questions about how to merge the pieces of information (e.g. definitions) and to what extent is data provenance important for the users. Answers might depend on different perspectives: lexicographers are protective of their wordings, while user preferences are yet to be seen.

Collocations were found in the same three sources (DicEst, ECD and BED). In Ekilex we faced the problem of overlapping: some multi-word units (MWU) from DicEst were collocations in ECD, e.g. *punane vein* 'red wine' and *kollane kaart* 'yellow card', and some collocations in ECD were usage examples in DicEst, e.g. *kodune aadress* 'home address', *isiklik elu* 'personal life' and *ebaväärikas käitumine* 'undignified behaviour'. To avoid duplicates, the authors of the ECD deleted the collocations that were MWUs in DicEst prior to import into Ekilex to avoid duplicates in both Ekilex and Sõnaveeb. This might have been solved differently by building a connection from the collocation to be presented both ways: as collocation and an MWU.

Concerning usage examples, the authors of DicEst (Langemets et al., 2018) stated that they have added all kinds of usage examples: full sentences, collocations and phrases. However, the research conducted by Kristina Koppel (2019, forthcoming) showed that neither language learners nor lexicographers themselves considered collocational phrases (e.g. *kangesti palav ilm* 'very hot weather', *väljapaistva arhitektuuriga ehitis* 'a building with extraordinary architecture') to be suitable examples. This is an issue to be solved in the future: it might be reasonable to move towards presenting phrases as MWEs or collocations, rather than usage examples.

Senses and collocations have occasionally been presented also in other datasets, e.g. in the prescriptive dictionary ÕS. One of the lessons learned so far is: do not import the dictionary as a whole. Extract valuable pieces of information instead. In this case, there is no need to analyse the dictionary database once again to fully understand for what purpose any fragment of (duplicate) information has been included in the dictionary. It is sufficient to import the pieces of information that undoubtedly add value.

3.3 From dictionaries to information layers

Some of the source datasets are focused on specialized information, such as morphology, word formation, collocations, etymology, language planning or language proficiency levels. They have been authored as separate dictionaries with varying degrees of autonomy from each other. In moving towards a single database, these datasets are turned into information layers and applied to the central "backbone" of headwords already present in the database, removing the need to specify variations of the same information again in separate dictionaries.

Morphology is a case in point. Declination patterns of Estonian words are well established and rarely debated among lexicographers, and morphological information has been centralized into The Estonian Morphological Database of the Institute of the Estonian Language 2019. This database is considered as a central service for all datasets.

Figure 5 shows aggregated information in Sõnaveeb for *diskussioon* 'discussion' from

different datasets: definition (from DicEst), collocations (from ECD), inflected forms (from the morphological database), etymology (from DicEst) and web sentences (external data from etSkELL via the Corpus Query System KORP[6] API).

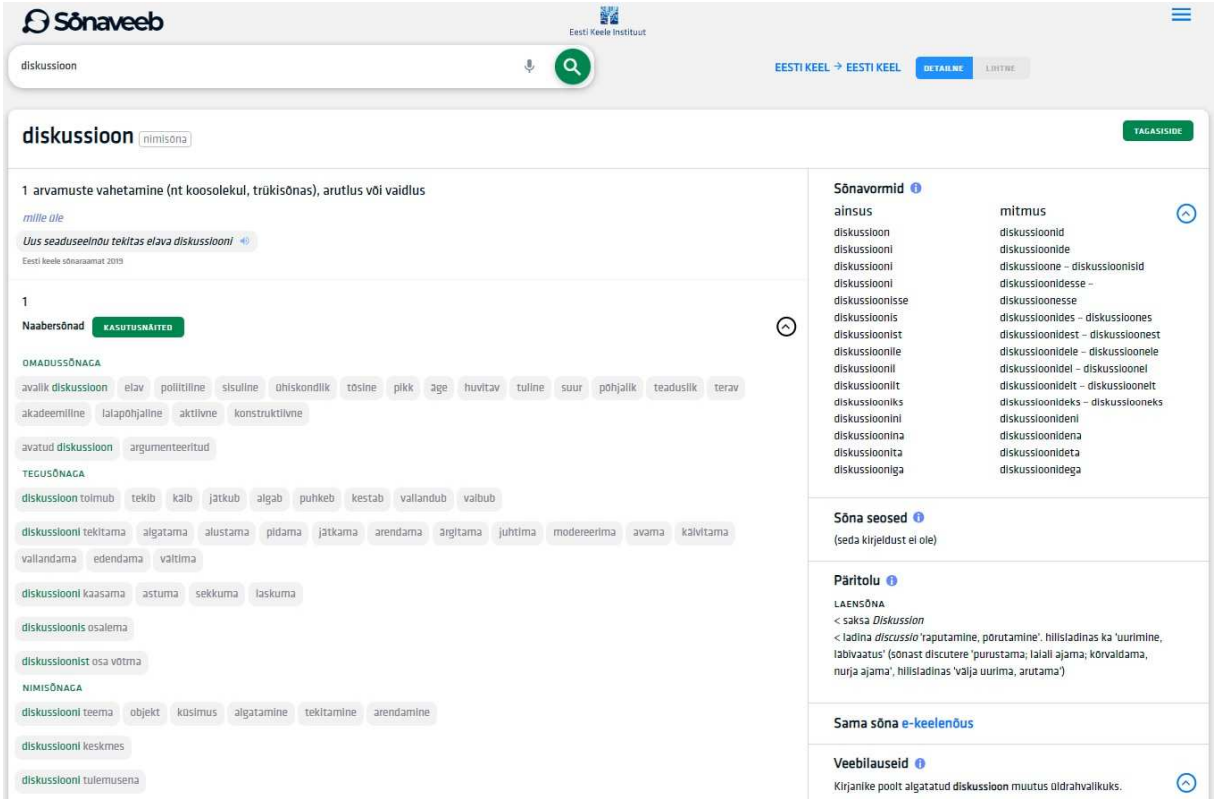

Figure 5. Aggregated information for diskussioon 'discussion' in Sõnaveeb (advanced mode)

Since all lexicographers trust the morphological database, it was agreed that morphology would only come from there, and any morphological information manually added to other dictionaries would be ignored during import. However, not all differences between dictionaries were inconsistencies. Rather than all possible forms from the database, we have chosen to present a subset: only most frequent forms in the simple mode, only approved forms for prescriptive language advice, only corpus-attested forms in advanced mode, and only forms that distinguish homonyms in most other dictionaries.

It would be ideal if inflected forms were labelled accordingly in the morphological database. The problem is that they are not. All target groups see either the full theoretically possible paradigms or trivially filtered subsets (e.g. learners only see the first of alternative forms). For lexicographers, this is a step in the wrong direction. They feel they already had the correct manually selected forms in their dictionary,

[6] https://korp.keeleressursid.ee/ (20 May 2019).

which are now gone. Tagging is planned and can be partially automated based on these same datasets: if a form is listed in a learner's dictionary, it can be labelled as suitable for learners, in addition to attaching corpus frequencies to forms.

The situation is similar with collocations. BED and ECD were compiled as separate dictionaries, and BED was the first dictionary where collocations were presented explicitly. The manually selected learner-level collocations from BED were not imported to Ekilex. Instead, all collocations were imported from ECD and then filtered. The simple mode in Sõnaveeb only shows collocations consisting entirely of words included as headwords in BED. As a result, there are many more collocations for a headword in the simple mode than there were originally in BED, including collocations where the collocate as a word is included in BED but the sense is not, for example there is a collocation *liblika nukk* 'butterfly pupa' under headword *liblikas* 'butterfly', although *nukk* 'doll' is only defined as a toy in BED. Again, the solution would be semi-automatic labelling of collocations for the language level, which is planned but has not been started.

Concerning prescriptive data, the preparatory phase of the new normative dictionary (ÕS 2025) started in 2019. It has already been agreed that prescriptive statements will be a layer on top of the otherwise descriptive backbone, rather than a separate dictionary. This will constitute a major change for the prescriptive ÕS, and issues may arise.

3.4 Linking and reuse of data

Ekilex treats all word-like entities as words, including ones that were unstructured character strings in previous systems. The objective is to improve data quality by replacing character strings with entity references. A practical problem is that this inevitably requires manual disambiguation, the additional workload of which comes as an unpleasant surprise to the lexicographer. More importantly, such linking exposes inconsistencies. Some of these may be deliberate, and in any case the lexicographer is understandably not happy about this. Notable examples of this type of issue are synonyms, equivalents, collocations, usage examples and definitions.

The representation of synonyms and equivalents was mixed in the earlier systems that Ekilex imported data from. They were word entities in termbases, but character strings in general lexical datasets. Of the latter, DicEst authors had manually ensured that synonyms were all valid, symmetrical (A=B and B=A) and unambiguous (the homonym number and sense number of the target word were also given), and other datasets contained few synonyms, so these were easy to import.

Russian equivalents, on the other hand, were completely ambiguous character strings. If the same string was given as an equivalent more than once, we had no way of knowing if these were the same meaning, a polysemous word or separate homonym.

The current solution has been to import them all as one polysemous word waiting to be manually disambiguated, resulting in the most frequently used Russian words having over 20 meanings. This result can be seen when searching in the Russian-Estonian direction, and was so unexpected for both users and lexicographers that we had to display a special warning about searching in that language direction.

The same problem was in the collocations dictionary database, where the headword, its collocates and possible context were added as character strings. In preparation for importing into Ekilex, the lexicographers semi-manually disambiguated the collocates so that they were easy to interpret as references to word entities. The contexts remained ambiguous and we applied automatic disambiguation where possible.

The Ekilex data model, and also for end users in Sõnaveeb, represents collocations so that one is always a relation between two or more lexeme entities. It is not necessary to specify one of them as the headword or otherwise superior component. The import did give asymmetrical information about the components, because the collocation's relation with the headword, unlike other components, also contained information about which part of speech group and grammatical relation group that collocation belongs to from the point of view of the headword. The following combinations were present in the dictionary, with the following issues:

- The collocation was listed under only one component. Due to the symmetry of the Ekilex model, it also appeared when viewed from the opposite direction, which was unexpected for the lexicographers, who had deliberately only included it in one direction.

- The collocation was listed under the headword, as well as under other collocates. Symmetry was expected here, but another issue emerged. As the collocation was edited separately in each direction, possibly by different lexicographers, it was possible that the information given was different, for example the same collocation could be in plural under one collocate and in singular under the other. This problem was also evident in example sentences. If the importer found identical examples, it imported them only once. Problematic were the cases when one of the lexicographers had edited the sentence for clarity, so the examples were no longer identical, resulting in the collocation having two very similar examples in Ekilex.

The authors of dictionaries currently imported into Ekilex do not have a common understanding of what a usage example is, as mentioned in Chapter 3.2. The shortest examples are word-like entities, making them candidates for being treated as word entities instead of usage examples. We adopted the practical heuristic that we imported an example as a word entity if it was either one word, or was included in the DicEst as a MWU. This is in addition to the issue of the same phrase being described as a MWU/example/collocation across the imported dictionaries (see Chapter 3.2. on duplicates).

Likewise, definitions in the imported dictionaries were sometimes word-like, or consisted of a comma-separated list of word-like strings. The lexicographers agreed that these were more like synonyms or synonym lists than definitions, but we decided not to attempt parsing them during import. If lexicographers consider it necessary, they can manually change those definitions in Ekilex.

While most commas between word-like strings were indeed separators, there were exceptions, e.g. *tee ruttu, muidu jääd hiljaks* ('hurry up, otherwise you'll be late') where the comma was part of the expression. Especially among Russian equivalents and usage examples, the strings often further contained textual condensations that were too underspecified to expand automatically.

1. Examples resulting in two expansions:

   ET olgu peale(gi) = olgu peale / olgu pealegi 'well and good'

   RU женатый [мужчина] = женатый / женатый мужчина 'married man'

   обыденная ~ разговорная речь = обыденная речь / разговорная речь 'colloquial speech'

2. Examples resulting in more than two expansions:

   RU смесь ~ раствор соединяет ~ связывает строительные камни = смесь соединяет строительные камни / смесь связывает строительные камни / раствор соединяет строительные камни/ раствор связывает строительные камни 'the mixture connects building stones'

   RU подорожник снижает ~ понижает опухлость ~ отёчность = подорожник снижает опухлость / подорожник снижает отёчность / подорожник понижает опухлость / подорожник понижает отёчность 'plaintain reduces puffiness'

3. Examples where the expansion requires linguistic knowledge:

   ET ta on töö peale ~ tööle laisk = ta on töö peale laisk / ta on tööle laisk 'he/she is too lazy to work'

   RU в дальнейшем ~ впредь будь осторожнее = в дальнейшем будь осторожнее / впредь будь осторожнее 'be more careful in the future'

Due to the third group, we decided not to attempt automatic expansion, but to leave the corrections to be done manually in Ekilex.

The condensations have been used for conserving space in print dictionaries. In electronic form, space limitations are replaced by the need to search for items. It would of course be possible to create an index that would refer all full forms to the condensed form, but indexing the third group would require exactly the same linguistic

knowledge that expanding them would. We have yet to reach a decision on what to do with such condensations.

Source datasets contained annotations of form (bold, italic, subscript and superscript) using several different markup notations. The use of italic was especially ambiguous. Two frequent meanings of italic script were citations and metalanguage (the "or" between alternatives, for example). We set out to enforce marking up of content, not form, so that the italic would be replaced with a citation or metalanguage as necessary. This was straightforward, thanks to the limited nomenclature of italicized metalanguage items.

Where we ran into a wall, however, was with subscript and superscript. The orthodox way would have been to distinguish between their meanings in mathematics, chemistry, legislation, etc., mark each up with its correct meaning, and then display all of those meanings as subscript or superscript as before. While that would have been the correct way to do it, we decided to take the easier route and leave them marked up as subscript and superscript. After all, it is highly unlikely that mathematics or chemistry would change their notation so that we would have to replace the superscript with some other formatting. So we decided to tolerate an inconsistency in Ekilex that is theoretically messy, but very convenient in practice.

3.5 Authorship of separate dictionaries

Firstly, as mentioned in Chapter 3.1, in the Ekilex data model the words (i.e. headwords) and meanings (i.e. definitions and domain indicators) are dataset-agnostic. Secondly, after having processed, systematized, unified, supplemented, edited, etc. the information across datasets, the Ekilex resource receives the status of a single database containing all lexical information known about the Estonian language, protected by the Copyright Act.

We will make it possible to "(re-)derive" separate datasets from the Ekilex resource if there is a demand for them, e.g. from the owner of the economic rights (the government or a company), or from the authors of previous datasets or government regulations (e.g. from 2006 in Estonia, the literary norm is supposed to be based on the most recent printed (!) prescriptive dictionary ÕS issued by the Institute of the Estonian Language)[7].

Since starting working in Ekilex, the work on separate dictionaries will develop into the work on specific information layers. Again, several questions might arise, for instance the following. Should we show explicitly the origin/authorship of every piece of information after unification of the datasets? Who is the author of a "(re-)derived" dictionary if we use unified information fragments available in Ekilex for free but

[7] https://www.riigiteataja.ee/akt/114062011003 (20 May 2019).

compiled by several other lexicographers? Will the authors develop into content renters rather than owners (Bego, 2018)? These are issues to be solved.

4. External data in Sõnaveeb

4.1 Audio pronunciation, speech synthesis and speech recognition services

In Sõnaveeb, users can listen to the pronunciation of about 5,000 of the most frequent headwords, as well as their most important inflected forms, and of about 7,000 unadapted loan words. The information on pronunciation has been aggregated from different datasets: from BED (headwords and inflected forms) and the dictionary of Foreign Words (VL, unadapted loan words). In the case of unadapted loan words, we used Estonians who speak foreign languages (Italian and Spanish) at high proficiency levels. For the pronunciation of the most frequent words and their inflected forms, we used professional actresses.

Text-to-Speech synthesis[8], developed by the Institute of the Estonian Language, is used for reading out the example sentences chosen by lexicographers. The same application is quite widely used by Estonian newspaper publishers: users can listen to all articles on the internet, as well as on Estonian Public Broadcasting for reading out subtitles[9].

Speech recognition[10], developed by the Department of Cybernetics of the Tallinn Technological University, is used when dictating words. Speech recognition operates in real time. For optimum quality, users have to pronounce the search word clearly and steadily.

4.2 Web sentences

In Sõnaveeb, authentic example sentences from the corpus are displayed. They have been automatically selected and they have not been edited.

The example sentences are queried from the Estonian Corpus for Learners 2018 (etSkELL)[11] (250 million words) via the Corpus Query System KORP API. etSkELL corpus was compiled using the GDEX tool (Kilgarriff et al., 2008; Kosem et al., 2019) in Sketch Engine, and consists of sentences from various media texts, fiction, scientific texts, Estonian Wikipedia and Estonian textbooks. The example sentences for Russian

[8] http://www.eki.ee/heli/ (20 May 2019).

[9] https://heliraamat.eki.ee/ (20 May 2019).

[10] http://bark.phon.ioc.ee/webtrans/ (20 May 2019).

[11] DOI: 10.15155/3-00-0000-0000-0000-07335L

are queried from the ruSkELL 1.6 corpus via Sketch Engine JSON API. In Sõnaveeb, up to 26 web sentences per lemma are shown. In many cases, especially for low-frequency words, these are the only usage examples for a headword (Koppel, 2019, forthcoming).

Although all sentences in the corpus meet the criteria of good dictionary examples (Koppel, 2017), some of them are still incorrect. In many cases, this is due to errors in corpus annotation (lemmatization and part of speech tagging); polysemous words and homonymy also cause problems. (Koppel et al., 2019, forthcoming) Users assume that all information included in Sõnaveeb is compiled or edited by lexicographers, and hence is error-free. Web sentences, on the other hand, are authentic and unedited. After receiving user feedback that some users find some of the web sentences inappropriate, the editors of Sõnaveeb decided to use the same strategy as in Merriam-Webster's[12] and Collins'[13] dictionary portals and added an explicit note saying that the sentences were chosen automatically, they are unedited and they might contain errors. An evaluation of the Estonian GDEX configuration was carried out in 2019. The results show that according to lexicographers and Estonian language learners at the B2-C1 proficiency levels, 85% of the GDEX-selected examples were actually rated as suitable dictionary examples (Koppel, 2019, forthcoming).

5. Issues for the future

The future challenges involve compiling new data in the Ekilex, as well as the addition of new data from other dictionaries and termbases to be presented in Sõnaveeb.

1) **Prescriptive and descriptive data**. Concerning prescriptive data, the preparatory phase of the new normative dictionary (ÕS 2025) started in 2019. It has already been agreed that prescriptive statements will be a layer on top of the otherwise descriptive backbone, rather than a separate dictionary. This will constitute a major change for the present prescriptive dictionary (ÕS 2018), and issues may arise. Langemets et al. (2020, forthcoming) mention upcoming controversial cases where data from a descriptive dictionary (e.g. DicEst 2019) is opposed to data from a prescriptive dictionary (e.g. ÕS 2018).

2) **Synonyms**. At the moment only synonyms from DicEst are displayed in Sõnaveeb. We initiated the project for a synonyms database in 2019. Synonym candidates will be automatically extracted from different resources for importing into Ekilex, using word embeddings and semantic mirroring methods.

3) **Etymological data**. Dealing with etymology is an especially complicated area

[12] https://www.merriam-webster.com/ (20 May 2019).

[13] https://www.collinsdictionary.com/ (3 June 2019).

in the data model. Etymological data is an information layer for all dictionaries, currently only consisting of the etymological information contained in DicEst. For importing, etymologies were structured by creating and linking word entities for all the source languages: automatically where possible, but leaving several types of corrections to be done manually. We also plan to import the academic Estonian Etymological Dictionary (ETY), which will add more complexity.

4) **Information on different language levels according to language proficiency**. About 13,000 headwords will have indications of language proficiency level (A1-C1). The data on proficiency levels comes from etLex[14]: a database of vocabulary of different proficiency levels compiled in the Institute.

5) **Frequency information**. We plan to visualize frequency information in Sõnaveeb. The information comes from the Estonian National Corpus (crawled every two years since 2017). Periodic renewals of the corpus will also make it possible to present language change information.

6) **Terminological data**. Ekilex contains and supports both semasiological and onomasiological termbases. Only general dictionaries have been published so far in Sõnaveeb, however. Publishing termbases is planned for 2019 and involves the decision of whether to display their information onomasiologically, as is traditional for such termbases as IATE[15], or semasiologically, to be consistent with the current Sõnaveeb. Terminologists are convinced it should be onomasiological, but evidence suggests that users don't really understand the difference, and proper user research is needed.

7) **Bilingual data**. We plan to continue providing Russian equivalents to Estonian headwords (approx. 10,000 per year). We plan to increase the list of languages as there are more bilingual databases available at our Institute, e.g. Estonian-Latvian/Latvian-Estonian, Estonian-Finnish/Finnish-Estonian, Estonian-Chinese.

6. Conclusions

In this paper, we have described principles of aggregating and presenting of information in Sõnaveeb: a new language portal of the Institute of the Estonian Language, released in February 2019. The user can choose between two modes of information display: advanced or simple. The advanced mode is intended primarily for native speakers. The simple mode is intended primarily for learners of Estonian L2 at the A2–B1 proficiency levels. There are (so far) two language options in Sõnaveeb: it is

[14] http://www.eki.ee/keeletase (20 May 2019).

[15] https://iate.europa.eu (20 May 2019).

possible to choose between Estonian (monolingual) and Russian (bilingual). Users are provided with both the desktop and the responsive mobile design.

The project started in 2017 and so far the main focus has been on the development of a unified data model and on the import of different lexicographic and terminological databases from the earlier used DWSs.[16] The final goal is to develop a single source of lexicographic and terminological data in order to avoid duplication of data, to improve accessibility and to force the reuse of data.

This paper reported on problems encountered so far while aggregating the data into the single source, together with the solutions we have elaborated. When unifying the dictionaries, we have paid special attention to (near) duplicates, considering their possible usefulness for different user perspectives or an empty duplication to be avoided. We have parsed and are still parsing data fields containing more than one data element.

In centralizing data from separate dictionaries and databases, we consider different information layers as specific central services. These are multimedia files (audio services and pictures), morphology, etymology, collocations, synonyms, etc. We also provide access to different kinds of external sources: corpora sentences (through Corpus Query System's API), speech synthesis and speech recognition.

We have started user research on specific information layers to get a better understanding of users' wishes and needs. We are aware that, while developing the user interface to be more and more intuitive, internet skills still need to be improved.

We will make it possible to "(re-)derive" separate datasets from the Ekilex resource if there is a demand for them. We are trying to be very careful about the authorship of different pieces of information after unification of the datasets.

The development of Sõnaveeb continues both towards tighter aggregation of existing datasets and the addition of new data from other dictionaries and terminological databases, as well as compiling new data in Ekilex. In the near future, we foresee the compilation of prescriptive data, synonyms, Estonian L2 data, neologisms, other bilingual data, terminological data, etc.

7. Acknowledgements

The creation and development of the portal was funded by the Digital Focus programme of the Ministry of Education and Research (2018–2021) and by the EKI-ASTRA programme (2016–2022). The creation of the dictionary and terminology

[16] EELex http://eelex.eki.ee/, Termeki https://term.eki.ee/ and Multiterm https://www.sdl.com/software-and-services/translation-software/terminology-management/sdl-multiterm/ (20 May 2019).

database Ekilex was funded by the EKI-ASTRA programme (2016–2022). Software development has been provided by OÜ TripleDev.

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

8. References

BED: *Eesti keele põhisõnavara sõnastik 2019. [The Basic Estonian Dictionary 2019]* Eesti Keele Instituut. Sõnaveeb 2019. Available at: https://sonaveeb.ee (14.2.2019).

Bego, K. (2018). *Ten challenges for the Internet.* Available at: https://www.ngi.eu/news/2018/10/22/ten-challenges-for-the-internet/ (30 May 2019).

Collins Dictionary. Accessed at: https://www.collinsdictionary.com/ (20 May 2019)

DicEst: *Eesti keele sõnaraamat 2019. [The Dictionary of Estonian 2019.]* Eesti Keele Instituut. Sõnaveeb 2019. Available at: https://sonaveeb.ee(14.2.2019).

ECD: *Eesti keele naabersõnad 2019. [The Estonian Collocations Dictionary 2019]* Eesti Keele Instituut. Sõnaveeb 2019. Available at: https://sonaveeb.ee (14.2.2019).

Eesti Keele Instituudi eesti keele morfoloogiline andmebaas 2019. [Morphological database of Estonian 2019] Eesti Keele Instituut. Sõnaveeb 2019. Available at: https://sonaveeb.ee (14.2.2019).

Eesti Keele Instituudi vene keele morfoloogiline andmebaas 2019. [Morphological database of Russian 2019] Eesti Keele Instituut. Sõnaveeb 2019. Available at: https://sonaveeb.ee (14.2.2019).

EELex: Langemets, M., Loopmann, A. & Viks, Ü. (2006). The IEL dictionary management system of Estonian. In G.-M. de Schryver (ed.). *Proceedings of the Fourth International Workshop on Dictionary Writing Systems: Pre-EURALEX workshop: Fourth International Workshop on Dictionary Writing System.* Turin: University of Turin, 2006, pp. 11–16.

Eesti-vene sõnaraamat 2019. [Estonian-Russian dictionary 2019] Eesti Keele Instituut. Sõnaveeb 2019. Available at: https://sonaveeb.ee (14.2.2019).

Eesti-vene õpilase ÕS 2019. [The Standard Estonian Dictionary for Russian School Students 2019] Eesti Keele Instituut. Sõnaveeb 2019. Available at: https://sonaveeb.ee (14.2.2019).

Ekilex. Accessed at: https://ekilex.eki.ee/ (20 May 2019)

etSkELL 2018: *Sketch Engine for Estonian Language Learning 2018*. Accessed at: https://etskell.sketchengine.co.uk/ (15 May 2019)

ETY: *Eesti etümoloogiasõnaraamat (2012) [Estonian etymological dictionary (2012)].* Tallinn: Eesti Keele Sihtasutus.

Kallas, J., Tuulik, M. & Langemets, M. (2014). The Basic Estonian Dictionary: the first Monolingual L2 learner's Dictionary of Estonian. In A. Abel, C. Vettori & Ralli, N. (eds.) *Proceedings of the XVI EURALEX International Congress: The*

User in Focus, Bolzano/Bozen, Italy, pp. 1109−1119.

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*, 13, pp. 53−71. DOI:10.5128/ERYa13.04.

Koppel, K. (2019) (forthcoming). 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.

Koppel, K., Kallas, J., Khokhlova, M., Suchomel, V., Baisa, V. & Michelfeit, J. (2019) (forthcoming). SkELL corpora as a part of the language portal Sõnaveeb: problems and perspectives. *Proceedings of eLex 2019*.

KORP: Accessed at: https://korp.keeleressursid.ee/ (15 May 2019)

Kosem, I., Koppel, K., Kuhn, T. Z., Michelfeit, J. & Tiberius, C. (2018). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, 32(2), pp. 119-137. https://doi.org/10.1093/ijl/ecy014.

Langemets, M., Tiits, M., Udo, U., Valdre, T. & Voll, P. (2018). Eesti keel uues kuues: Eesti keele sõnaraamat 2018. *Keel ja Kirjandus*, 12, pp. 942–958.

Langemets, M., Kallas, J., Norak, K. & Hein, I. (2020) (forthcoming). New Estonian Words and Senses: Detection and Description. *Dictionaries*.

Lew, R. (2013). Online dictionary skills. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.) *Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Available at: http://eki.ee/elex2013/proceedings/eLex2013_02_Lew.pdf.

Merriam-Webster Dictionary. Accessed at: https://www.merriam-webster.com/ (20 May 2019)

ruSkELL1.6: *Sketch Engine for Language Learning (SkELL) for learners of Russian.* Accessed at: https://www.sketchengine.eu/ruskell-examples-and-collocations-for-learners-of-russian/ (15.5.2019).

Sketch Engine. Accessed at: https://www.sketchengine.eu/documentation/api-documentation/ (15.2.2019)

Sõnaveeb: *Sõnaveeb 2019 [Wordweb 2019].* Accessed at: https://sonaveeb.ee (20.5.2019)

Tavast, A., Langemets, M., Kallas, J. & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts.* Ljubljana, Slovenia, pp. 749–761.

VL: *Võõrsõnade leksikon. [The Dictionary of Foreign Words] 8., põhjalikult ümber töötatud trükk.* Eesti Keele Instituut, Kirjastus Valgus, 2012. Available at: http://www.eki.ee/dict/vsl/ (30.5.2019).

Weitzman, L. (2014 [2004]). Meta-design for "sensible" information. *Interactions*, Vol.

11, Issue 2, March, April, pp. 71–73. Updated by author in 2014. DOI: 10.1145/971258.971284.

ÕS 2018: *Eesti õigekeelsussõnaraamat ÕS 2018. [The Dictionary of Standard Estonian ÕS 2018]*. Eesti Keele Instituut. Tallinn: Emakeele Sihtasutus, 2018). Available at: http://www.eki.ee/dict/qs2018/ (30.5.2019).