Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX

Arvi Tavast, Margit Langemets, Jelena Kallas, Kristina Koppel
Institute of the Estonian Language, Tallinn
E-mail: arvi@tavast.ee, margit.langemets@eki.ee, jelena.kallas@eki.ee, kristina.koppel@eki.ee

Abstract

The Institute of the Estonian Language is developing EKILEX, a new dictionary writing system for both semasiological dictionaries and onomasiological termbases. While the long-term vision is to have a single data source that provides consistent information about Estonian, the system also needs to cope with the multitude of existing datasets. In this paper, we present work in progress on modelling the data and importing an initial sample of legacy dictionaries. The data model is based on an m:n relation between words and meanings, which are both unified across dictionaries, even while there still are separate dictionaries in the system. What is dictionary-specific is only the mapping between word and meaning. The importing of dictionaries has revealed various issues with data quality: ambiguities, underspecification, inconsistencies and conflicts. These need to be dealt with, if the long-term vision is to be achieved. We also outline the next steps of human- and machine-readable publishing, corpus connection and quantification (frequency, salience measures, etc.).

Keywords: data modelling, dictionary portal, interoperability, linked data, Estonian

1 Introduction

The Institute of the Estonian Language has been publishing dictionaries and termbases for decades, providing a comprehensive description of Estonian from a variety of perspectives. At the same time, the state of the art in lexicography has evolved from paper to electronic, introspective to empirical, manual writing to corpus-based generation, normative to descriptive, binary to quantitative, and human only to machine-readable. Software and storage formats have changed too, and one of the reasons for this is increased mutual awareness between linguistics and information technology.

Departments and working groups of the Institute have had a high degree of autonomy in compiling dictionaries, thus leading to the possibility of publishing inconsistent information, duplicating each other's work, and storing data in a way that is only semi-structured. Three separate dictionary writing systems are used for historical reasons, the dictionary data models are far removed from each other, and the whole dictionary system has gradually moved away from current thinking in lexicography.

The Institute has thus reached a point where changes are inevitable, first in the working methods, but consequently also in the tools used. In August 2017, development work was started for EKILEX, the Institute's new dictionary writing system, with the aim of addressing the most pressing issues and supporting the necessary changes in working methods.

In this paper, we report on the work in progress from the point of view of data modelling, including the importing of representative legacy dictionaries as a stress test for the new model.

First we describe what the existing datasets look like, and where the problems are that have caused the Institute to initiate the development of yet another dictionary writing system. We continue by referencing currently existing standards for lexical data representation. This is followed by three sections about

the work in progress itself: data modelling, including comparison to the referenced standards, data import and data harmonization. In the Discussion section, we explain the rationale behind some of the more difficult or controversial design choices, as well as a number of lessons already learned during the project. We conclude by outlining some directions for future work: electronic publishing for humans, connecting dictionaries to corpora, machine-readable publishing, and quantification of lexical data.

2 The Current Situation

The Institute is currently using three separate dictionary writing systems for its dictionaries and termbases:

- EELex[1] (Langemets, Loopmann & Viks 2010; Jürviste et al. 2011) was developed in-house from 2003 to 2015 and currently holds more than 70 dictionary databases of different types. It started out as an XML database, but for performance reasons was later transferred to a mixed model storing chunks of XML in a relational database. EELex predominantly uses semasiological data models and is highly customizable. At a late stage in its development support for onomasiological data structures was added, but it has been rarely used. For electronic publishing, a separate web interface is developed for each dataset, with automatic nightly data transfers.
- Termeki[2] was originally developed from 2007 to 2015 by Werkdata Ltd, and is still available commercially as termbases.eu.[3] Since 2012, a contract with Werkdata has allowed the Institute to provide it for free to Estonian terminologists, and it has mainly been used outside of the Institute. It is a relational database system with a partially customizable onomasiological data model, and has been used for about 40 termbases and one bilingual general language dictionary. Electronic publishing is implemented by allowing anonymous users restricted access to the same database.
- Multiterm,[4] a commercial product using XML technology and a partially customizable onomasiological data model, is used for two major termbases by the terminology department at the Institute. Electronic publishing in our current setup requires manual data transfers, which are only undertaken once a month.

Disparate data models have been used, especially in EELex, where a new data model has been custom developed for each new dictionary. Such flexibility was originally been designed to accommodate the heterogenous wishes of dictionary authors, and has fulfilled this objective well: each author has obtained a data model of their choice. However, the results are not necessarily in line with current thinking in lexicography, the datasets are disconnected, information is duplicated and inconsistent across datasets, and the same information may be located differently in the model depending on the dataset.

All three have data models with a 1:n relation between form and meaning. One word has several meanings in the semasiological case, and one concept has several terms in the onomasiological case. A shortcoming of both is non-normalized data: information on the n-side of the 1:n relation is duplicated, causing inconsistencies due to human error (see Figure 2 for examples).

Especially in the two XML-based systems, the elements on the n-side of the 1:n relation are mostly plain text values, rather than entity references, making them ambiguous. In some newer datasets, homonym and meaning numbers may be included, but mostly the reference only consists of the target headword as a character string. This is understandable, considering that the only use case the authors

1 https://eelex.eki.ee [18.5.2018]

2 https://term.eki.ee/ [18.5.2018]

3 https://www.termbases.eu/ [18.5.2018]

4 https://www.sdl.com/software-and-services/translation-software/terminology-management/sdl-multiterm/ [18.5.2018]

originally had in mind was a human reader, who has no difficulty navigating various meanings of a word. In addition to problems with machine-readable publishing, data reuse and linking of dictionaries, this solution also makes it impossible to automatically enforce internal consistency.

There are violations of atomicity and other abuses of each data model, for instance a definition and its source(s), or multiple definitions, in a single definition field; duplicated classifier codes with one of them containing a typo; domain labels entered in the pronunciation field due to excessive difficulty of using the domain classifier, and so on.

Regarding electronic publishing, a major issue is that each dataset has a separate public interface, in the worst case requiring the user to perform 130+ searches in separate dictionaries with the same search term. There is no machine-readable publishing, apart from custom exports performed at the request of prospective users of the data. For ESTERM[5] and MILITERM,[6] the two termbases compiled in Multiterm, their monthly publishing interval is not nearly enough to serve current needs. There are also performance and usability issues due to architectural choices made years ago, including limited browser compatibility.

Recognizing these issues, the Institute has started developing EKILEX, a dictionary writing system to replace all three current systems for both semasiological and onomasiological data, and importing existing datasets into the new system.

3 Prior Work

Modern lexicography has shifted its focus from compiling stand-alone dictionaries to making lexicographic data findable, accessible, interoperable and reusable (FAIR[7] data). Lexicographers thus need to pay more attention not only to the quality of lexicographic data but also to the data modelling of lexicographic databases.

There are several frameworks that can be used as a starting point for the database model. The most common are Lexical Markup Framework (LMF; ISO 24613:2008)[8] and Text Encoding Initiative (TEI XML)[9] for lexical resources, and TEI-Lex0 Initiative (Bański, Bowers & Erjavec 2017) for encoding of retro-digitized dictionaries. The goals of LMF are to provide a common model for the creation and use of lexical resources, to manage the exchange of data between and among these resources, and to enable the merging of a large number of individual electronic resources to form extensive global electronic resources (Francopoulo et al. 2006). The Text Encoding Initiative (TEI) is aimed at equipping scholars with markup suitable for describing the majority of textual forms (concerning lexicography, especially for printed dictionaries) and analytic approaches, and providing extension capabilities to encompass new or infrequently found phenomena. TEI-Lex0 aims at 'formulating guidelines for the encoding of retro-digitized dictionaries by streamlining and simplifying the recommendations of the "Print Dictionaries" chapter of the TEI Guidelines' (Bański et al. 2017: 485). LMF is widely used for building lexical resources, see e.g. Borin et al. 2012.

There are also models that use ontologies and are geared towards the conversion of lexical resources to linked data. These are the LExicon Model for ONtologies (*lemon*)[10] (McCrae et al. 2012) and its

5 http://termin.eki.ee/esterm/ [18.5.2018]

6 http://termin.eki.ee/militerm/ [18.5.2018]

7 https://www.force11.org/group/fairgroup/fairprinciples [18.5.2018]

8 http://www.lexicalmarkupframework.org/ [18.5.2018]

9 http://www.tei-c.org/index.xml [18.5.2018]

10 http://lemon-model.net/ [18.5.2018]

recently developed OntoLex-Lemon model[11] (McCrae et al. 2017). *lemon* is a model for modelling lexicon and machine-readable dictionaries and is linked to the Semantic Web and the Linked Data cloud. Bosque-Gil et al. (2016) claim that *lemon* is a de-facto standard for representing lexical information in the Web of Data. This model was tested in several lexicographic projects and has proved its success. Tiberius and Declerck (2017) explored reusing, improving and optimizing a dictionary of contemporary standard Dutch (ANW) by porting some of its elements into modules of *lemon*. They claim that encoding information in *lemon* has a number of advantages, including better modularization of the data, linking to other (lexical) data as well as providing improved access to data. *lemon* has been chosen as the backbone of BabelNet's[12] lexical knowledge linked data representation.[13] McCrae et al. (2017: 590) state that dictionaries represented with *lemon* or OntoLex-Lemon can be easily integrated with other resources previously converted to the Resource Description Framework (RDF)[14] without any remodeling efforts.

4 Data Modelling for EKILEX

Development of EKILEX was started in August 2017 in cooperation with the software house TripleDev Ltd, and the first project stage with currently committed funding will last until the end of 2018.

Initial requirements for the data model are the following:

- Describe language, as opposed to describing dictionaries: combine legacy dictionaries into a single data source about the language, and treat both words and meanings as existing independently of whether any dictionary includes them or not.
- Represent both semasiological and onomasiological data.
- Accommodate all existing dictionaries and termbases.
- Enforce best practices in both lexicography and terminology.
- Support the authors in maintaining data integrity.
- Comply with any current or future standard of data exchange.

The long-term vision is to have a single data source that provides consistent and comprehensive information about Estonian words, combining the research done at all departments and working groups of the Institute. In that ideal situation, each author or working group would be enriching the database with, for example, collocations, Chinese translations, normative recommendations or other data according to their expertise, instead of working in isolation on a collocations dictionary, Estonian-Chinese dictionary or normative dictionary.

Realistically, however, we also need to cope with the current transition stage of still having a multitude of dictionaries, each with their own ideology, working methods, legacy data and (administrative or financial) publishing requirements. The authors are aware of the problems described above and unification is their long-term goal.

Considering this gap between vision and reality, the process agreed for the project is the following:

- Make sense of existing dictionaries with their peculiarities, and import them as they are. Only correct errors (duplicated or non-structured data) that can be corrected automatically, or that the

11 https://www.w3.org/community/ontolex/ [18.5.2018]

12 http://babelnet.org/ [18.5.2018]

13 http://www.w3.org/2015/09/bpmlod-reports/multilingual-dictionaries/ [18.5.2018]

14 https://www.w3.org/RDF/ [18.5.2018]

authors are able and willing to manually correct within the project schedule. The results must be publishable as separate dictionaries.

- While the imported material still contains inconsistencies, provide a clear path towards unification, so that authors will be able to use the new system to gradually improve data quality, by reconciling conflicts found within their own or between datasets. Already at this point the results must also be publishable as a single dictionary.

The development of EKILEX uses an agile methodology (Scrum[15]) and is driven by priorities set by the stakeholders as expressed in the initial discussions and during biweekly sprint planning meetings. For data modelling, we did not start from a ready-made standard, but analyzed customer requirements instead, and optimized our solution to the particular situation of the Institute. In the following sections we describe the main design choices, comparing them to LMF and OntoLex-Lemon where applicable.

4.1 Word and Meaning

Considering the inherent data duplication issues of 1:n models for both semasiology and onomasiology, we use instead an m:n relation between word and meaning: one word can have several meanings, and one meaning (concept) can be referred to by several words (see Figure 1). This could also be described as reuse of word and meaning data. In relational database terms, this is implemented using a link table between Word and Meaning tables.

While this linking entity, called Lexeme in our model and representing 'this word in this meaning', started out as a purely technical link table, it turned out to be the central point of our data model in terms of relating to other entities: the majority of data items are parameters of the Lexeme. In OntoLex-Lemon, our Lexeme corresponds to Lexical Sense and works in the same way, "mapping from a word to a concept" (McCrae et al. 2017).

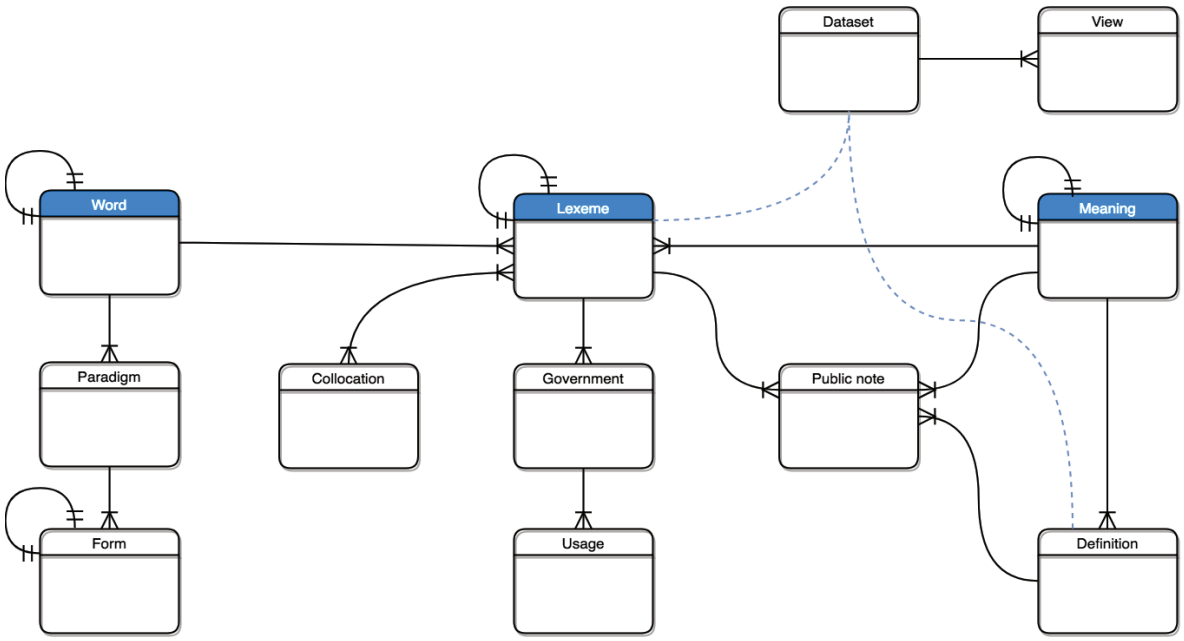

Figure 1: Simplified data model of EKILEX, highlighting the m:n relation between Word and Meaning through the Lexeme link table. Note that only Lexeme and Definition explicitly belong to a Dataset (dictionary or termbase). All other entities are common to all dictionaries in the EKILEX model, possibly being associated with a dictionary through Lexeme or Meaning.

15 https://www.scrum.org/resources/what-is-scrum [18.5.2018]

4.2 Dictionaries

During the current transition phase with multiple dictionaries, the Lexeme also carries dictionary-specific information, making its content effectively 'this word in this meaning as described in this dictionary'.

So we are able to keep both words and meanings independent of dictionaries. Indeed, this makes theoretical sense too: words belong to the language, not any particular dictionary, and the same for meanings. What belongs to dictionaries is only the description of the relationship between words and meanings.

This design also works well with our vision-reality gap. When importing legacy dictionaries and finding two words with the same form, in most cases we have no way of telling whether these are two homonyms, one polyseme or simply data duplication. We can move on by importing them as two homonyms for the time being, with the option of deciding to combine them at a later stage. We apply a similar approach to meanings. During the import process we have no machine-readable information about meanings whatsoever, so whenever finding a meaning in the legacy files, we import it as a new meaning. Since words and meanings are dictionary-agnostic, the process of combining duplicates across dictionaries is exactly the same as within a dictionary.

4.3 Lexical relations

Lexical relations are represented using two distinct methods. The first is used for synonyms and translation equivalents, defined as words with the same meaning within one language or across languages. These are actually connected to the same meaning. There is no explicit synonymy or equivalence relation in addition to the connection through the meaning.

All other lexical relations are expressed using relations between meanings. So there are no explicit antonymy, hyperonymy, and so on relations between words either. Instead of recording that "dog" is a hyponym for "animal", for instance, we record that dog is a type of animal. This is transparent for the user, so dictionary authors may continue thinking in lexical relations if they wish, even if they are stored as conceptual relations in the database.

4.4 Morphology and other dictionary-agnostic linguistic information

As the long-term vision no longer contains separate dictionaries, we are already moving towards centralizing some of the data elements.

One of the data categories that does (or should) not depend on the dictionary is morphology. While there are known differences between dictionary authors in which forms they consider legitimate, it does not make sense to list those differences without explanation. They should be either reconciled or, if the differences continue to be important for the authors, tagged according to their normative, stylistic or other status. In our model, a word has one or more paradigms (inflectional patterns in LMF terms), each containing one or more forms. Forms have written representations and may have various types of phonetic transcriptions and links to sound files.

The word itself does not have any linguistic representation. Instead, one or more of its forms can be marked as canonical, and the word gets its representation(s) from there. Forms can also be quantified, e.g. to not show rare forms to the L2 learner, or explicitly tagged as suitable for some use case.

Other clear examples of dictionary-agnostic data items include collocations and word-formation. The Dictionary of Estonian (DicEst), due to be published in 2018, is the first dictionary where these two will not be written from scratch, but reused from the Collocations Dictionary (COLL,

Kallas et al. 2015) and the Dictionary of Word-Families (Vare 2012), respectively. In EKILEX, the relations between linguistic items (e.g. derivational relations, compounds and their subterms, collocations, etc.) are represented using database relations between the corresponding entities.

5 Data import and harmonization

The minimum required selection of datasets to be imported during the first project stage was the following:

- At least one representative dataset from each source (EELex, Termeki, Multiterm), to verify that this source is readable and get an overview of the problems.
- The Dictionary of Estonian as the largest and most modern general dictionary, to serve as a basis for the database backbone.
- The Collocations Dictionary, due to its specific requirements and the fact that it is scheduled to be completed by November 2018 and published using the new system.
- Three resources with the Estonian-Russian language pair, as a requirement from one of the financers.

When talking of 100+ databases the process of harmonization is the key concept. Harmonizing the lexical data involves an iterative process of capturing, defining, analyzing and standardizing the data. Harmonizing definitely improves the quality of the data by eliminating redundancies, inconsistencies and duplications, as well as facilitating the exchange of data and improving automation by ensuring interoperability (see Figure 2). The problem is that to a large extent this work is to be done manually by lexicographers editing the dictionaries, or half-manually, using some simple in-house tools for editing and adjusting the lexical data.

The backbone of the new EKILEX will be the corpus-based comprehensive scholarly Dictionary of Estonian (DicEst), which has been compiled at the Institute of the Estonian Language since 2010, and will be published online in our new dictionary portal Sõnaveeb ('Web of Words') in autumn 2018. The dictionary focuses on written Estonian, being the descriptive, not the normative dictionary. There are ca 110,000 words in the dictionary and when published it will then be constantly updated. Most of its elements have been imported into the new EKILEX model. The morphosyntactic properties of the words (for all dictionaries) will be imported from the Morphological Database that is currently being developed at the Institute.

The core entities of Word and Meaning, as well as more peripheral Morphology, Collocation, Usage Example and Etymology, and the like, are common to all dictionaries in the EKILEX model (note that the Lexeme is dictionary-specific, see Figure 1). This means that during the import process data needs to be unified across the legacy dictionaries – e.g. importing a headword as many times as it has legitimate homonyms, not as many times as it is found in the 100+ dictionaries to be imported. This is a nontrivial task even for the relatively clear case of homonyms. Not only may dictionaries differ in their level of detail for a particular headword, authors may also have various working definitions of what to consider homonyms in the first place. We currently use a combination of morphology-based word sense disambiguation, manual disambiguation and organizational measures (persuading authors to reach an agreement) for unifying the word list. Next in line for unification are example sentences, collocations and etymologies, followed finally by meanings.

The first task lexicographers were involved in with EKILEX was aligning and linking at the lemma (homonym) level. With the help of a special mini-tool we could connect homonyms across many dictionaries. Table 1 shows three homonyms and different forms of 'luup' from seven dictionaries.

In the dictionary portal, when searching 'luup', the user will get three homonyms ('luup1', 'luup2', 'luup3'). Each of these is connected to the content (all data from entries) from several dictionaries. The morphophonetic data (the degree of quantity) in ET-RU (learners) and ORTH is harmonized via the Morphological Database.

Table 1: Three homonyms in seven contemporary dictionaries. DicEst = *Dictionary of Estonian* (to be published in 2018, the backbone of EKILEX), BasicDic = *Basic Estonian Dictionary* (2014), ET-RU (learners) = *Estonian-Russian Dictionary of Standard Estonian for Learners* (2011), ET-RU (general) = *Estonian-Russian I-V* (1997-2009), COLL = *Estonian Collocations Dictionary* (to be published in 2018), ET-FI (general) = *Estonian-Finnish* (to be published in 2018), ORTH = *The Dictionary of Standard Estonian ÕS 2018* (to be published in 2018).

DicEst	BasicDic	ET-RU (learners)	ET-RU (general)	COLL	ET-FI (general)	ORTH
luup1 'loupe'	–	.luup I	luup I	luup	luup1	l`uup 1.
luup2 'sloop'	–	.luup II	luup II	–	luup2	l`uup 2**.**
luup3 NEW! 'looper'	–	–	–	–	luup3	–

Another major challenge for importing existing dictionaries is that the information in them is often ambiguous or underspecified. Collocations, lexical relations and other references to entities in the same dictionary are increasingly expressed using relations, not unstructured text any more, but the target of that relation is still a string of characters, not an object reference. We solve these case by case. When first attempting to import a dataset, such ambiguities are logged for the dictionary owner to review and decide what to do. Some can be resolved using hints found elsewhere in the data, some can be manually disambiguated before the next import attempt, and for some, the dictionary owner may decide to omit them from the import altogether. The rest are usually more labor-intensive to resolve, so we import the ambiguity as it is, and leave data harmonization to be performed at a later time, when it is already in the new system.

Besides, lexicographers have been faced with the idea of 'linking at sense level'. This is not an easy process, but EKILEX goes a step further than linking, actually combining equivalent meaning entities from separate dictionaries into a single entity. As a result, that single meaning entity will link things such as mini-definitions or glosses to longer definitions, collocations to senses of their counterparts, translation equivalents (from bilingual dictionaries) to senses in monolingual dictionaries, and so on. When looking closer at *luup**1*** 'loupe' we recognize instances that we have to harmonize (Figure 2):

Five definitions (in Estonian, all defining the same sense 'loupe')**:**
lihtne optikariist, mis annab esemeist suurendatud kujutise [DicEst]
– [BasicDic]
– [ET-RU (learners)]
suurendusklaas [ET-RU (general)]
suurendav optikariist [COLL]
lihtne optikariist, mis annab esemeist suurendatud kujutise [ET-FI (general)]
suurendusklaas [ORTH]

Synonyms or candidates for synonyms (in Estonian)**:**
suurendusklaas [DicEst, explicitly marked as a synonym]
suurendusklaas [ET-RU (general), originally encoded as a definition]
suurendusklaas [ORTH, originally encoded as a definition]

Translation equivalents (Russian, Finnish)**:**
лупа, увеличительное стекло [ET-RU (learners)]
лупа [ET-RU (general)]
luuppi, suurennuslasi [ET-FI (general)]
Usage examples (in Estonian, almost the same combinations recurring in slight variations)**:**
kümnekordse suurendusega luup [DicEst]
vanahärra uuris fotosid läbi luubi [DicEst]
tugev / terav / suurendav luup [COLL]
luupi kasutama / luubiga uurima / luubiga vaatama / luubiga lugema [COLL]
kümnekordse suurendusega luup [ET-FI (general)]
vanahärra uuris fotosid läbi luubi [ET-FI (general)]
uurib luubiga, läbi luubi postmarke [ORTH]
Translations of the usage examples (translations into Russian, Finnish)**:**
kümnekordse suurendusega luup – kymmenkertaisesti suurentava suurennuslasi [ET-FI (general)]
vanahärra uuris fotosid läbi luubi – vanhaherra tutki valokuvia luupilla [ET-FI (general)]

Figure 2: Instances across dictionaries to be harmonized in the case of *luup**1*** 'loupe'.

We can observe how fuzzy the boundary is between (short) definitions and synonyms, e.g., the term *suurendusklaas* 'magnifying glass' appears to serve as both in different original encodings. The duplication of the same material in different dictionaries has been unavoidable when compiling printed dictionaries as well as standalone and strictly separated dictionary databases (as in EELex up to now). In the case of the EKILEX model these are the inconsistencies.

6 Discussion

The EKILEX project has brought up a number of issues to be addressed and decisions to be made. Some of the choices described above have not been straightforward at all, and some have even been revisited and changed as a result of new information.

A major discussion point was whether to make the model recursive, i.e., unify all form-related entities (Word, Collocation, Usage Example, and maybe even Definition) into what is now the Word. We decided otherwise, and to have separate entities for each. The reason was that while these entities are theoretically similar and do share some important properties, notably that of having a meaning, many properties are not shared and the business logics applied to them are still very different. We thus chose a wider and shallower model over deep recursion, to keep queries and program logic simpler. Recursive RDF can still be exported from our model if needed.

The central idea of our model is the m:n relation between form and meaning. We do not know of any dictionary writing system that would implement this idea, at least not as radically as EKILEX (by not having any synonymy or equivalence relations at all). However, the idea itself is not new. In a well-hidden form it already exists in the LMF standard, where the Synset entity can be construed to correspond to what we call the Meaning. So while we do agree with Borin et al. (2012: 3599) that at first glance LMF looks unusably semasiological, it seems that theoretically the Synset entity there could be used to represent onomasiological data too, similar to Wordnet. In OntoLex-Lemon, the idea is more visible in the form of the Lexical Concept, relating m:n to the Word.

Despite prior familiarity with the data models of existing datasets, we underestimated the workload of data import. The initial plan was to complete the first round of imports by the end of 2017, but as of

this writing in late March 2018 it is still not completed. Technical implementation is not the only or perhaps not even the main reason for delays. Attempts to first create a mapping between the models and then to actually load the existing XML files have revealed decisions that authors were able to ignore in the semi-structured model, but that have to be decided now. The importing activity has even sparked discussions on the principles and objectives of some datasets, like who is the target group and what are their needs. Such questions do not have a single correct answer, resulting in lengthy discussions delaying the project.

There is a constantly nagging gap between the long-term vision and what is currently possible, given the legacy data. Namely, when we find words with the same written representation in several datasets when importing, and there is no information on how to combine them, we import them as homonyms. This is an obvious temporary solution for insiders familiar with the import process. For normal users, it looks like a simple UI issue ("you are displaying this word too many times"), while solving it would actually require a major undertaking of manual sense-level linking of each dataset to the backbone. This linking is still firmly in the plans, but the workload is daunting.

From the linked data perspective, we are not linking the various dictionaries of the Institute. Instead, we combine them into a single dictionary or Lexicon in the OntoLex-Lemon sense. This can then be published as linked data, if needed. The reason is that unlike the global community of linked linguistic data, dictionary authors are (or at least should be) under common management, following the same objectives and working methods, and capable of cooperation. This creates an opportunity to provide the added value of actually making the dictionaries consistent and non-redundant, in addition to making them link to each other.

7 Next steps

We are continuing with the process of data import and should release the first version of the dictionary writing system for lexicographers and terminologists in November 2018. In addition, the following next steps have been planned.

7.1 Dictionary portal

The user interface and the types of access to data depend very much on the data model behind the actual data. Access to data in dictionary portals ranges from searching different dictionaries (via linking) to searching in the data within the entry (Boelhouwer, Dykstra & Sijens 2017: 755). The user is generally expected to be capable of identifying and classifying dictionaries according to type. However, this might not be the best premise, as quite often there are dozens of dictionaries and databases on a website (e.g. the Estonian dictionary website[16], European Dictionary Portal[17]), which makes it difficult for the user to decide which one is the right one.

Our near-future EKILEX-based dictionary portal Sõnaveeb ('Web of Words', to be launched in autumn 2018) is meant to serve human users as an aggregator with items of content collected to one web page and enabling access to data within several dictionaries. These days, when searching for what a word means or how it is translated, people do not necessarily realize that they are searching a dictionary. They are just looking for the answers to their questions. The variety of data now available means that it is possible to meet both learners' productive as well as receptive needs.

16 http://portaal.eki.ee/sonaraamatud.html [18.5.2018]

17 http://dictionaryportal.eu/ [18.5.2018]

The new portal is linked to the new Estonian Corpus for Learners 2018 (etSkELL)[18]. The Corpus was compiled using the Estonian GDEX module (Koppel 2017) to filter the sentences in the Estonian National Corpus 2017, which is the largest and newest Estonian corpus (about 1.1B tokens) in Sketch Engine (Kilgarriff et al. 2004). GDEX (Good Dictionary Example, Kilgarriff et al. 2008; Kosem et al. 2018) scores sentences according to how well they meet predefined conditions. All sentences that met the conditions of the hard classifiers were collected into the Corpus for Learners (others were removed). All sentences were then scored and reordered (with the highest scores at the top) using the soft classifiers of the Estonian GDEX module. The resulting Corpus contains about 248,000 words and about 25M sentences that derive from various media and scientific texts, fiction, Estonian Wikipedia and the Estonian Coursebook Corpus of CEFR-graded sentences.

When searching for a word, the portal Sõnaveeb directs the query to the corpus query system KORP[19] using an API, and a certain number of authentic example sentences is presented.

7.2 Machine-readable publishing

Since EKILEX stores data in a structured and normalized form, there will always be a mapping from our database to any existing or future standard of data exchange, including any that will be developed in the ELEXIS project. The mapping to OntoLex-Lemon is especially straightforward. Over the years, the Institute has received and fulfilled several requests for wordlists, usually in very simple text formats. While there have been no requests to access the Institute's datasets as linked open data, providing such access is technically possible.

7.3 Quantification

A future development that we want to prepare for with this data model is quantification. The model allows any relation to be quantified, from morphological preferences to the relation between a Word and its Meaning. The collocations that we import, for example, already have empirical frequency and salience measures attached, widening the selection of possible display methods for collocations. These measures themselves may pose additional temporary challenges, like undifferentiated frequency counts for homonyms due to lack of semantically tagged corpora, but we believe in empirically based quantification in the long term, and have already left room for this in the data model.

8 Conclusion

In this paper, we presented the data model of EKILEX, a new dictionary writing system for both semasiological dictionaries and onomasiological termbases. We also discussed various issues that arose in the process of importing legacy dictionaries into the new system, such as issues with data quality: ambiguities, underspecification, inconsistencies and conflicts.

The Institute of the Estonian Language has been using three separate DWSs (EELex, Termeki, Multiterm) for its dictionaries and termbases, which has resulted in disconnected datasets and duplicated, inconsistent information across these. All three DWSs have a 1:n relation between form and meaning – one word has several meanings in the semasiological case, and one concept has several terms in the onomasiological case. The data model of EKILEX on the other hand is based on an m:n relation between words and meanings – one word can have several meanings and one meaning (concept) can

18 https://etskell.sketchengine.co.uk/run.cgi/skell [18.5.2018]. The authors would like to thank Jan Michelfeit for compiling the corpus.

19 https://korp.keeleressursid.ee/ [18.5.2018]

be referred to by several words. Words and Meanings are linked by Lexemes, which carry dictionary-specific information and which in our model represent 'this word in this meaning as described in this dictionary'. We are keeping both Words and Meanings independent of dictionaries, both belonging to the language, not to any particular dictionary. Dictionary data is held by the Lexeme, i.e. the description of the relationship between Words and Meanings.

The long-term vision is to have a single data source (EKILEX) that provides consistent and comprehensive information about Estonian words, combining the research done at all departments and working groups of the Institute. The backbone of the new EKILEX will be the corpus-based comprehensive scholarly *Dictionary of Estonian* (DictEst), and other linguistic items (morphology, compounds, derivational relations, collocations, etymology) will be linked with DictEst.

References

Bański, P., Bowers, J., & Erjavec, T. (2017). TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, V. Baisa (eds.), *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference, Leiden, 2017*. Brno: Lexical Computing CZ s.r.o. Accessed at: https://elex.link/elex2017/proceedings-download/ [18.5.2018].

Boelhouwer, B., Dykstra, A., & Sijens, H. (2017). Dictionary portals. In P. A. Fuertes-Olivera (ed.), *The Routledge handbook of lexicography*. London and New York: Routledge, pp. 754–766.

Borin, L., Forsberg, M., Olsson, L.-J., & Uppström, J. (2012). The open lexical infrastructure of Språkbanken. In *Proceedings of Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pp. 3598–3602. Accessed at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/249_Paper.pdf [18.5.2018].

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., & Aguado-de-Cea, G. (2016). Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case. In *GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop Programme*, pp. 65–73. Accessed at: http://www.citeulike.org/user/jgracia/article/14024090 [18.5.2018].

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., & Soria, C. (2006). Lexical markup framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Accessed at: http://www.lrec-conf.org/proceedings/lrec2006/pdf/577_pdf.pdf [18.5.2018].

Jürviste, M., Kallas, J., Langemets, M., Tuulik, M., & Viks, Ü. (2011). Extending the functions of the EELex dictionary writing system using the example of the Basic Estonian Dictionary. In *Electronic lexicography in the 21st century: New Applications for New Users: Proceedings of eLex 2011, Bled, 10-12 November 2011*, pp. 106–112. Accessed at: https://dialnet.unirioja.es/servlet/articulo?codigo=4567368 [18.5.2018].

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 *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*. Ljubljana; Brighton: Trojina, Institute for Applied Slovene Studies; Lexical Computing Ltd.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal, J. DeCesaris (eds.), *Proceedings of the 13th EURALEX International Congress*. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, pp. 425–432.

Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). Itri-04-08 the sketch engine. *Information Technology*, 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 [Estonian Papers in Applied Linguistics]*, *13*, 53–71. https://doi.org/10.5128/ERYa13.04 [18.5.2018].

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*.

Langemets, M., Loopmann, A., & Viks, Ü. (2010). Dictionary management system for bilingual dictionaries. In S. Granger, M. Paquot (eds.), *eLexicography in the 21st Century: New Challenges, New Applications*. Louvain-la-Neuve: Presses universitaires de Louvain, Cahiers du CENTAL, pp. 425–429.

McCrae, J. P., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., & Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, *46*(4), 701–719.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, V. Baisa (eds.), *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference, Leiden, 2017*. Brno: Lexical Computing CZ s.r.o., pp. 587–597. Accessed at: https://elex.link/elex2017/proceedings-download/ [18.5.2018].

Tiberius, C., & Declerck, T. (2017). A lemon model for the ANW dictionary. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, V. Baisa (eds.), *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference, Leiden, 2017*. Brno: Lexical Computing CZ s.r.o., pp. 237–251. Accessed at: https://elex.link/elex2017/proceedings-download/ [18.5.2018].

Vare, S. (2012). *Eesti keele sõnapered [Dictionary of Estonian Word-Families]*. Accessed at: http://www.eki.ee/dict/sp/ [18.5.2018].

Acknowledgements

This work has been supported by receiving funding from the European Regional Development Fund. Project EKI-ASTRA 2014-2020.4.01.16-0034