EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-11 September 2021
Ramada Plaza Thraki
Alexandroupolis, Greece

www.euralex2020.gr

Proceedings Book
Volume 1

Edited by Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-1-5

Edited by: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2020 Edition

Towards the Superdictionary: Layers, Tools and Unidirectional Meaning Relations

Tavast A., Koppel K., Langemets M., Kallas J.

Institute of the Estonian Language, Tallinn, Estonia

Abstract

We report on the ongoing project of developing the Ekilex dictionary writing system and joining existing dictionaries into the EKI Combined Dictionary. To facilitate the joining, several tools have been developed to solve data quality issues and turn textual data into structured entities. The resulting superdictionary thus contains various sets of information, which we call layers, either transformed from existing dictionaries or authored already in Ekilex. Our current focus is on the layers for synonyms and equivalents, which we describe in terms of their data model, lexicographic processes and lexicographer feedback from the first six months of Ekilex in production. As it turns out, the layer system may need expanding to accommodate an ever-growing list of requirements. The unidirectional data model for synonyms fully conforms to its design specification and received favourable first impressions, but extended use has started to cast doubt on the optimality of the model. We describe the pros and cons of this model and possible alternatives.

Keywords: synonyms; equivalents; data modelling; unified dictionary

1 Introduction

The goal of the Ekilex project (Koppel et al. 2019; Tavast et al. 2018) is to join dictionaries into a single superdictionary, the EKI Combined Dictionary (EKI ühendsõnastik, CombiDic), as opposed to linking between dictionaries or aggregated search across dictionaries (Boelhouwer, Dykstra & Sijens 2017). The underlying assumption is that users look for information about words, not about dictionaries, which means that the current system of multiple dictionaries with duplicated and conflicting information is not desirable.

Timing of the project also coincides with the rise of automated, corpus-based processes to replace introspective lexicography (Gantar, Kosem & Krek 2016; Kallas et al. 2019) as well as training lexicographers to pay more attention to the modelling of lexicographic data. Continued development of the superdictionary is an integral part of the project, so the goals are to: 1) join existing dictionaries, 2) create technical and administrative incentives for authors to cooperate, 3) improve the superdictionary to provide a radically better lexical resource for the user.

Despite a consensus about user benefits, these ideological and process-related changes are difficult for lexicographers, due to four interconnected reasons:

- Bringing a legacy dictionary into a structured database exposes its internal conflicts, previously hidden in disconnected entries. Doing the same with a number of dictionaries additionally exposes duplication and conflicts across the dictionaries. The result looks hideous, especially in a traditionally compiled bilingual dictionary trying to fulfil the needs of all conceivable users, where the target language equivalents have been a long list of (partial) translational equivalents fitting many different specific translation contexts. Gathering such occurrences to form a *word* entity mercilessly displays them side by side, which is not a pleasant sight for the authors.
- While specialised tools (described below) can be developed to assist in resolving these data quality issues, it is still largely manual lexicographic work. Given the decades that have gone into compiling the original dictionaries, the volume of this work looks daunting if not unrealistic. Lexicographers also rightly feel that their previous work is not sufficiently respected, and they are forced to start over from scratch.
- Especially in combination with the descriptivism of corpus-based lexicography, this necessitates a shift in thinking. Even if one would ideologically still prefer the old system, it is simply not feasible due to the workload involved. Responsibility gets transferred from the lexicographer, announcing the truth, to the reader, making sense of messy empirical data. While agreeing theoretically that it is better to be broadly right than precisely wrong, in practice authors feel uncomfortable with allowing uncertainty in a dictionary and trusting readers to draw their own conclusions.
- Previously autonomous dictionary working groups, now united into a large group working on layers of a single central dictionary, trust each other to varying degrees. There may also be differences in the lexicographic principles followed by each group. Unifying those principles and achieving trust is an administrative challenge.

Ekilex aims to make the shift easier by delivering tangible benefits for lexicographers, moving processes towards more automation on the continuum between manual authoring and fully automated generation of dictionaries. Development is ongoing and iterative, meaning that tasks are continually adjusted to lessons learned and insights discovered. At the time of writing, we are able to report on two relatively straightforward batch processes (word joiner and meaning joiner), but the main focus of this paper is a specialised tool for synonyms and translational equivalents. After listing some prerequisites in section 2, the specialised tool will be described in the rest of the sections.

2 Prerequisites

2.1 The Ekilex Data Model

Let us first briefly describe the Ekilex data model (Tavast et al. 2018). The central idea is that *word* and *meaning* are connected through *lexeme*, to express a many-to-many relation between words and meanings, see Figure 1. In the text, we use monospaced font to refer to database entities.

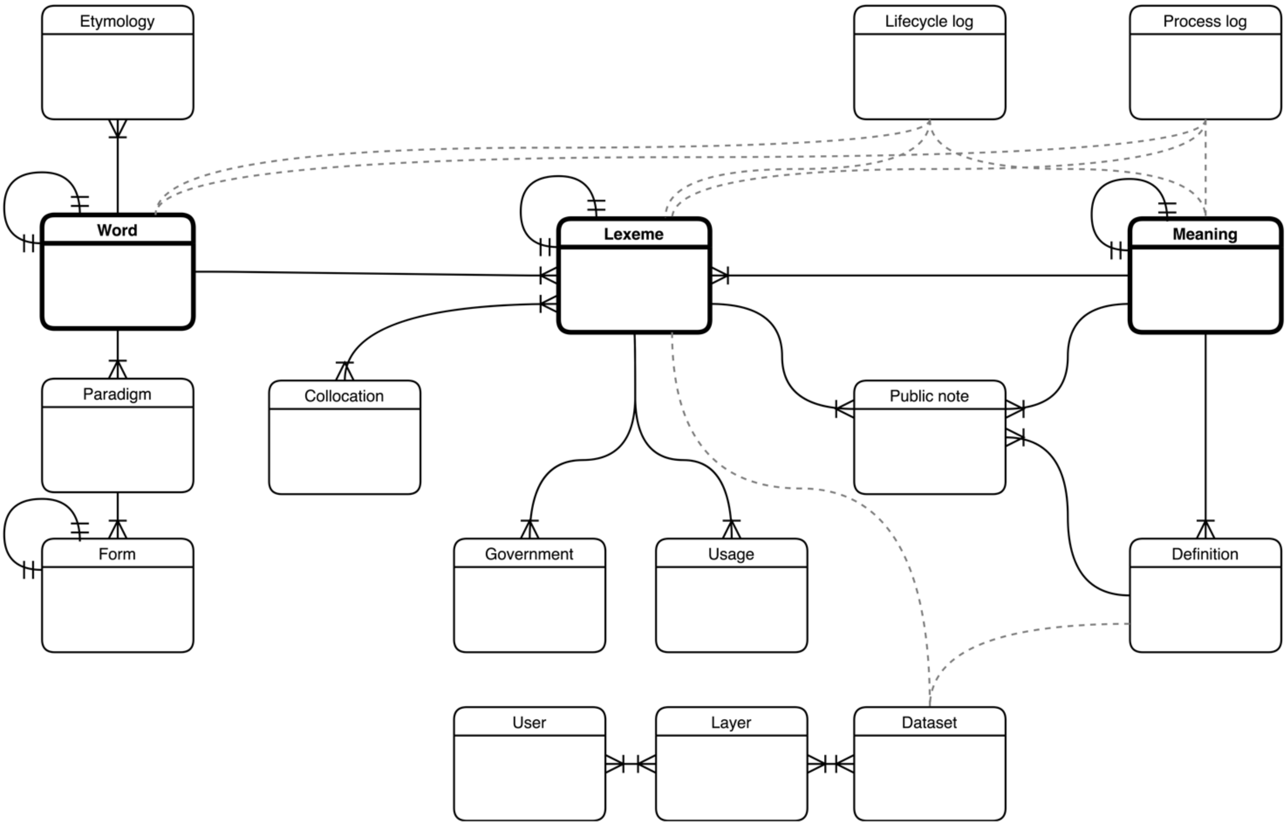

Figure 1: Simplified logical data model of Ekilex. Central entities are highlighted in bold.

- The *word* is an object in language, characterised mainly by its character composition, morphology and etymology. We make a distinction between homonyms (separate *words* with identical character composition) and polysemes (a single word with multiple meanings).
- The *meaning* is an object in cognition, characterised in the database by its domain(s), definition(s) and any related notes.
- The *lexeme* is an object in the dictionary, expressing the connection between *word* and *meaning*. It could be defined as "this word in this meaning as described in this dictionary". It contains information peculiar to the word-meaning combination like part of speech and example sentences, as well as dictionary-specific information like administrative status of the entry.

To refer to the user perspective as opposed to the data model, we also use traditional lexicographic terminology like *entry* (a record in a semasiologically organised dictionary, for us a word entity plus all related entities) and *headword* (a word that has such an entry). Similarly, since Ekilex is also used for terminology work, terminologists have their own *entries* (a record in an onomasiologically organised termbase, for us a meaning entity plus all related entities), *concepts* (a meaning having such a record) and *terms* (a word in a termbase). Viewpoints of the database and the user are distinguished elsewhere too: a *synset* for a user is a set of synonyms (words), which in database terms is described as a meaning that all these words are connected to. A headword may have several *senses*, each represented by a combination of lexeme and meaning in the database. A *lexical resource*, which may be a dictionary or a termbase for the user, is a dataset entity in the database.

2.2 Tools for Data Quality Improvement

Ekilex first obtained its data from importing dictionaries from previous dictionary writing systems, each with its own words and meanings, resulting in massive duplication of both. We used batch joiners to help mitigate this data quality issue.

Words were mostly character strings in the imported dictionaries, so the importer had no way of distinguishing between yet another occurrence of a previously found word and a new legitimate homonym. Only the Dictionary of Estonian 2019

(*DicEst* 2019; Langemets et al. 2018) had treated homonyms systematically. We assumed that a large general language dictionary like DicEst would have found all homonyms. Or conversely, if a word is not homonymous in DicEst, we could safely assume that all such character strings are occurrences of a single word and can be combined into a single *word* entity. Based on this assumption, the word joiner took care of 87,013 duplicated but really non-homonymous Estonian word types imported from multiple dictionaries.

Legitimate homonyms needed manual disambiguation, most of which was done before importing using specialised tools built by Indrek Hein, a senior developer at the Institute of the Estonian Language.[1] Manual joining of homonyms is also possible in Ekilex, and this is being done as part of normal dictionary compilation or editing. A total of 1,080 homonymous word types have been disambiguated manually, and it has taken about 15 person-days.

A similar approach was used for mapping meanings across component dictionaries of the CombiDic (see Koppel et al. 2019 for details). If a word was monosemous in DicEst, its meaning was joined with meanings of the same word from other component dictionaries where it was also monosemous. We found 57,461 such monosemes in DicEst and connected to them 76,845 meanings from other component dictionaries.

Meanings have also been joined manually, both before import using the same specialised tools, and already in Ekilex as part of the normal lexicographic workflow. Unlike homonyms with clear (even if theory-dependent) criteria for deciding whether two words are the same or not, meanings are completely open to human judgement, therefore also disagreements between authors of different dictionaries. This process is ongoing, much more time-consuming than the 15 person-days of homonyms, and can cause fundamental problems as we will show below.

The last batch tool so far, also with the smallest effect, joined homonyms for other languages. We don't have a similar gold standard for homonymy in other languages as DicEst is in Estonian, therefore we can only guess based on various hints. One such hint has been used, namely that if foreign words with the same form have the same Estonian equivalent, then they are most probably one and the same word, and can be combined. This took care of 12,882 foreign word types.

3 Layers

Uniting previously separate dictionaries into a single CombiDic does not entail, at least not initially, a lack of distinction between types of lexicographic information originating from the component datasets, or even a consolidation of the groups of authors. Lexicographers are still working on separate or at best partially overlapping tasks in their respective projects and entering their own data elements (e.g. synonyms, equivalents, usage examples, normative recommendations). The difference is that since September 2019 all this information now ends up in the same headword entry of CombiDic (i.e. data connected to a single word in the data model, see Figure 1), together with other data types imported from existing dictionaries like etymology, morphology and collocations.

To manage this agglomeration, we use the concept of layers. A layer belongs to one or more datasets, provides a coherent set of data elements, is accessible to and authored by a specific team of lexicographers, and has its own process status to track the team's progress. The idea is to allow multiple teams to contribute their expertise to entries of CombiDic, seeing each other's work, but not being overly disturbed by changes made by other teams.

The following layers are being actively authored in 2020:

- Partial synonyms. We will describe their data sources, data model and authoring process in section 4.
- Russian equivalents. While the synonym process is also applicable to equivalents, it is not used in the particular case of Russian. The reason is that rich information (meaning divisions of the equivalents) is already available from component datasets, which makes it easier to simply join existing meanings across the components in Ekilex.
- Normative recommendations, which will allow a specifically filtered view of CombiDic to replace the revered normative Dictionary of Standard Estonian ÕS (*Eesti õigekeelsussõnaraamat* 2018).[2]

At the time of writing, the current challenge with layers is that both their nomenclature and expectations towards them are growing rapidly. As the latest development, it has become evident that new ad hoc layers need to be created on the fly. The reason is that the lexicographic process is usually not random but organised by distinct tasks even within one team of authors. Each task starts from some kind of search result or list of entries that the lexicographer needs to check. As work progresses, items on the list need to be checked off one by one. The problem is that there may be any number of such lists, and both the lists and their progress status need to be managed somehow. It is also not known in advance which teams want to see the status of which (sub)layers, as the work of a neighbouring team may or may not be relevant for the task at hand. A practical example of when it does become relevant: when the CombiDic core team changes the meaning distribution of a word, then this should reopen the headword for several other teams to update their information accordingly. This is still work in progress without an agreed solution so far.

4 (Partial) Identity of Meaning

Words and meanings, the two central data elements of a lexical resource, differ in how well established their representation in lexical resources is. Words are straightforward to write down as a character sequence, and there is very

[1] http://www.eki.ee/dict/selgroog/ [30/05/2020]

[2] The centre of the Estonian dictionary publishing tradition has been formed by two large, competing, partially duplicating and partially conflicting general dictionaries, the descriptive Dictionary of Estonian 2019 (DicEst) and the normative Dictionary of Standard Estonian ÕS 2018. In an effort enabled by and parallel to the Ekilex project, both will be merged into the CombiDic. Special treatment of the normative layer is needed due to the legal status of the ÕS in normative situations like exam grading.

little room for disagreement about how to do that in most languages. The common practice to organise dictionaries alphabetically also provides a widely accepted (even if arbitrary) similarity metric: words sharing initial characters are treated as belonging together. Meanings, on the other hand, lack a physical form that would simultaneously be human-readable,[3] sufficiently debate-proof to be usable in practice, and capture which other meanings this meaning is similar to.

When designing Ekilex, the objective was to be able to represent both full and partial identity of meaning, both within a language and across languages. Full identity, a notoriously debatable concept, is here defined as a function of the particular dataset and lexicographic judgement: two meanings are identical (i.e. they are really one meaning) if the lexicographer decides not to distinguish between different shades of meanings, but to enter their words as full synonyms in one language or exact equivalents across languages.[4] This judgement can change in time and vary across datasets of different sizes or objectives, but within the process of authoring a particular headword entry it can be treated as constant. The design objective also included the ability to represent partial identity or similarity of meanings, which is needed for partial synonyms and non-exact equivalents.

In the following, we first describe requirements of representing meaning similarity from the lexicographer's standpoint, then discuss conceivable solutions in a lexical database, and finally the approach(es) taken in Ekilex.

4.1 Requirements

As the defining characteristic of full synonyms and exact equivalents is that of having the same meaning, the proper way to represent them is to connect those *word* entities to the same meaning entity. As far as we have the data from existing dictionaries, this has already been done, and can further be done in the Ekilex user interface. Being connected to the same meaning, such synonyms and equivalents are direction-agnostic from the lexicographer's viewpoint: if $a = b$, then inevitably $b = a$.

For several reasons however, practical lexicography has a strong tradition of directionality. Lexicographers want to express that $a = b$ without necessarily taking a stand on whether $b = a$ or not. The whole concept of reversing bilingual dictionaries (Krek, Šorli & Kocjančič 2008) is based on the premise that equivalence is directional. Collocation dictionaries (e.g. Kallas et al. 2015) are another directional example, listing collocations according to their frequency or salience relative to the headword, not to the collocate. Earlier dictionary projects may even have been planned to remain unidirectional. E.g. if the objective was to present synonyms for 10,000 most frequent words but the synonyms were not restricted to the same frequency class, then there would have been tens of thousands of words in the synonym dictionary without synonyms of their own. Removing such restrictions has only been made possible by including synonyms as a layer in CombiDic and semi-automatic compilation.

The preference for directionality is amplified by an aspect of using empirical data from corpora, namely quantification. Word-level alignment of parallel corpora (for equivalents) and distributional semantics (for synonyms and equivalents) yield quantitative measures of how close the meanings are. Exact matches occur rarely, if at all, which complicates the picture to a level where the feasibility of ideal directionlessness is no longer beyond doubt. Especially as dictionaries become more comprehensive (and CombiDic is an attempt to maximise comprehensiveness), it is increasingly the norm that meaning identity is not exact, but some subtle differences need to be explicated.

Regarding synonymy and equivalence between polysemes, there is also a pragmatic workflow consideration. When starting to design the Ekilex module for synonyms, lexicographers expressed a strong preference to avoid the rabbit hole of chained relations, and instead complete the compiling of one headword before moving on to the next. Given the sense distribution of the current headword, they wanted to be able to add synonyms and equivalents to each of its meanings, without (yet) taking a stand on the sense distributions of the words added. For example,[5] when finding synonyms for the (sub)senses of the headword *board*, the lexicographer wants to connect the word *plank* to the 'piece of timber' meaning and the word *management* to the 'governing body' meaning, but not to select the correct sense for *plank* or *management*, even if these have other senses totally unrelated to *board*.

This preference entails the need to visit each similarity relation twice, entering *plank* as a synonym for *board*, and then separately entering *board* as a synonym for *plank*. To generalise, the number of required visits to a set of synonyms (a synset, to use the Wordnet term) equals the number of members in the set. Suppose we decide to consider *board*, *committee*, *management* and *directorate* synonymous in our dictionary, this synset needs to be visited four times, each time adding three synonyms. In the design phase, lexicographers were confident that this is an acceptable trade-off for keeping their habitual headword-based process, as opposed to meaning- or synset-based (like in Wordnets or termbases).

4.2 Data Sources

Compilation of the synonym layer of CombiDic follows the semi-automatic method where lexicographers post-edit automatically generated lists of synonym candidates. The candidates were extracted from existing dictionaries, including component datasets of CombiDic itself, taking advantage of the tradition to include synonyms in the definition and other fields, as well as semantic mirroring (Dyvik 1998, 2004). Distributional similarity (Turney & Pantel 2010) has so far only

[3] As opposed to machine-readable. Since Ekilex is an information system, all of its contents, including any representations of meanings, are machine-readable by definition.

[4] Other lexicographically relevant aspects of synonymy and equivalence, like style, register or frequency (see e.g. Yong & Peng 2007: 129–131), are properties of the `lexeme` entity in Ekilex, which makes them a separate discussion. Here we concentrate on identity or similarity of `meanings` as characterised by their definitions and domains.

[5] In the following, we use simplified examples in English to improve readability. CombiDic, including its synonym layer, starts from Estonian. English is not yet among the languages of CombiDic, but adding it is near the top of the wish list.

been calculated from the multilingual FastText model (Grave et al. 2018), with the new Estonian National Corpus 2019 (Kallas & Koppel 2019) being in the queue.

Since equivalents are like synonyms, only in another language, this synonym process can be extended to bilingual dictionaries with practically no modifications. We take candidate equivalents from wherever they can be found, including corpora and existing lexical resources, and make them available for the lexicographer to connect to meanings as described above.

4.3 Representation in the Data Model

In the simplest case, if synonyms and equivalents are deemed to have identical meanings, the corresponding words can literally be connected to the same meaning entity. Figure 2 shows the situation where *board* and *management* share the meaning of 'a governing body', while each word has other meanings too. Examples of the other meanings are shown greyed out, and any further synonyms in those meanings are omitted completely.

This is the pervasive approach in termbases, where it is known as concept-based or onomasiological[6] (Wüster 1979; Felber 1984; see also Tavast 2008). For general language, it is used in Wordnets (Fellbaum 1998). The obvious benefit is simplicity, both technically and conceptually: meanings are identical or not, there is no third option or gradation. However, synonymy and equivalence are necessarily directionless in this model, which is contrary to the lexicographic understanding of language. Especially bilingual dictionaries need to represent partial equivalents, which is not possible using this simple model. By forcing lexicographers to take a stand about the meaning distribution of both words simultaneously, it is also in conflict with the design objectives described in section 4.1 above. Finally, it is difficult, although not impossible, to quantify the degree to which each word denotes the meaning, but long-term goals of Ekilex include empirical quantification of as many pieces of information as possible.

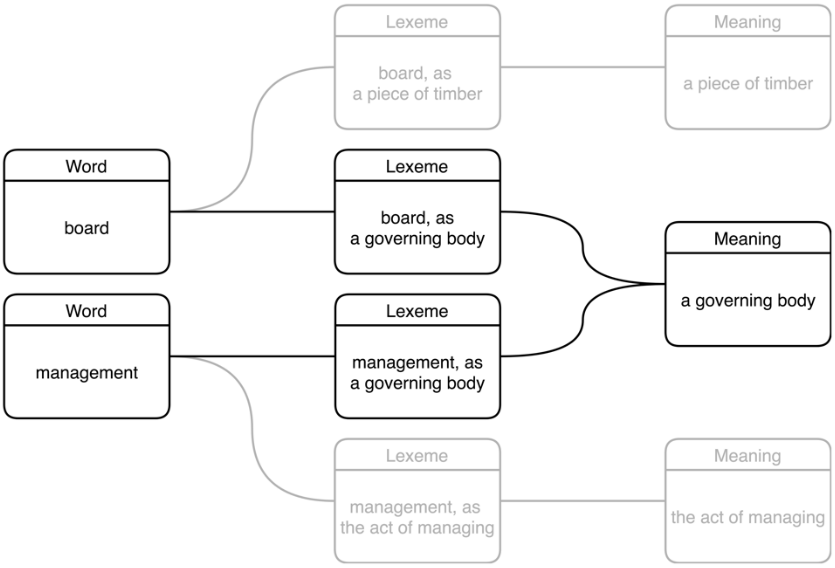

Figure 2: A single meaning: if *board* and *management* are considered full synonyms, they can be connected to the same meaning.

Therefore, this simplest model is only sufficient for full synonyms and exact equivalents, for which it is also used in Ekilex. It alone cannot represent partial synonyms or equivalents; neither does it accommodate lexicographers' preference for a directional, headword-based working process.

The next step is to have a separate meaning for each word, and link those meanings with (possibly weighted)[7] similarity relations (cf. Rudnicka et al. 2019). This is shown on Figure 3, where the 'governing body' meaning has been split in two and then reconnected with a similarity relation with a high similarity value. Taking this approach to the extreme by *never* allowing a meaning to have more than one word would make the lexeme entity redundant and reduce the data model to semasiology, which would be unacceptable for terminological users of Ekilex. This approach does, however, work seamlessly together with the single-meaning model above, so that full synonyms share a meaning (and terms share a concept), while providing the additional capacity to represent partial synonyms as meaning relations.

The similarity relations could further be made directional, which would allow describing situations where the similarity of a to b is different from the similarity of b to a, or one of the directions is absent altogether. This is a step in the right

[6] Pure onomasiology would treat all words as homonyms rather than polysemes, i.e. there would be two *boards* and two *managements* in the figure. This distinction is omitted here for simplicity. Incidentally, since Ekilex is used for both general language dictionaries and termbases, Ekilex users are also shielded from this distinction. The same words can be shown as polysemes to lexicographers and as homonyms to terminologists.

[7] How to obtain the weights is a separate topic. They could be based on the lexicographer's introspection, distributional similarity measures from a corpus, a function of which previous dictionaries have listed the relation, etc., or any combination thereof. The point here is that weighting is possible, should it be desired.

direction but does still not address the main point of the directionality requirement of section 4.1 above, because the correct meaning still needs to be specified on both sides of the relation at the same time.

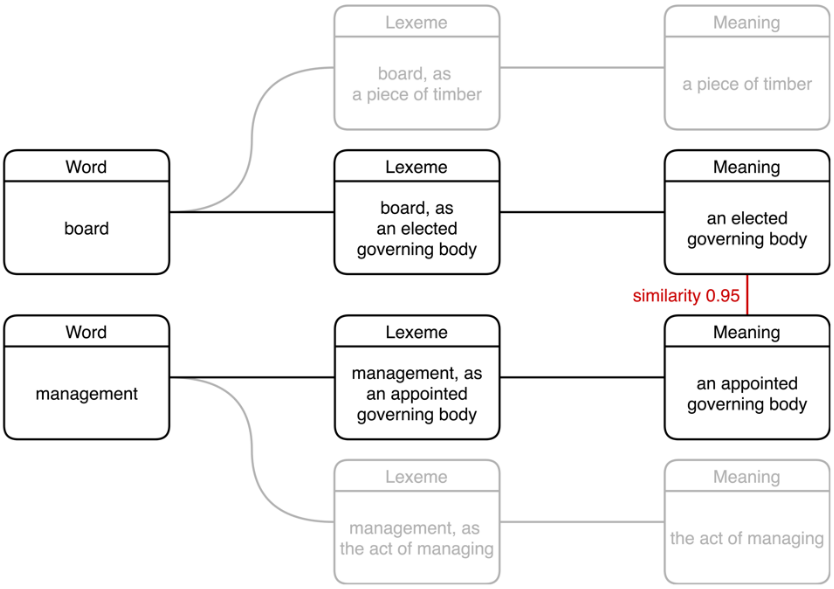
Lexeme
board, as a piece of timber
Meaning
a piece of timber
Word
board
Lexeme
board, as an elected governing body
Meaning
an elected governing body
similarity 0.95
Word
management
Lexeme
management, as an appointed governing body
Meaning
an appointed governing body
Lexeme
management, as the act of managing
Meaning
the act of managing

Figure 3: Related meanings: each word has its own meaning, and the meanings are related, with a similarity measure.

To cater for a completely directional, headword-based authoring process, Ekilex uses a special type of *lexeme*, the *secondary lexeme*, defined as "this meaning can *also* be expressed with that word". Figure 4 shows the result of a lexicographer working on the entry for *management* and adding the word *board* as a partial synonym to its first sense 'an appointed governing body'. The secondary lexemes can be weighted, so that a meaning can have stronger or weaker relations to many words in the same language (partial synonyms) or other languages (partial equivalents), like the weight 0.95 on Figure 4.

Note that *board* does not at this stage get *management* as a synonym. Theoretically it could get a new, third meaning through the secondary lexeme, but this behaviour was quickly ruled out based on feedback from lexicographers. To recap, adding *board* as a synonym in another entry leaves the entry for *board* itself completely unchanged. This is exactly the result that lexicographers requested: they only have to specify the meaning on one side of the relation.

Whether or not a corresponding partial synonymy relation needs to be added in the opposite direction, i.e. if *board* is a synonym for *management*, then whether *management* is also a synonym for *board*, will be decided only when the lexicographer reaches the other headword in the authoring process. This unidirectionality part of the requirements differs markedly from the habitual process of describing synonymy and equivalence used by other teams in EKI and elsewhere, and was intended by the synonyms team as a means of optimising their workflow. However, as discussed in section 5, using this solution for practical work has surfaced negative side effects that may motivate returning to the related meanings model.

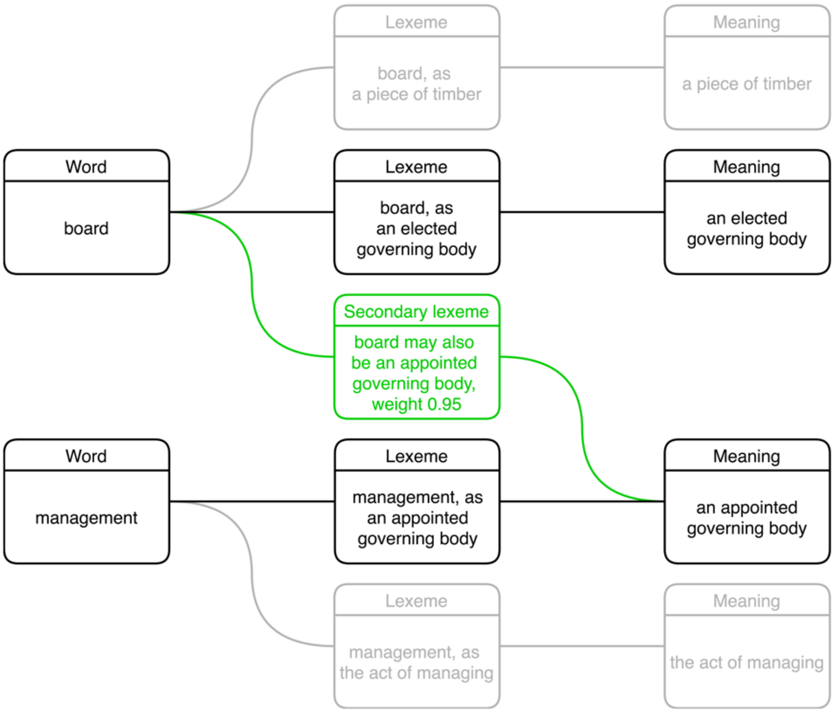
Lexeme
board, as a piece of timber
Meaning
a piece of timber
Word
board
Lexeme
board, as an elected governing body
Meaning
an elected governing body
Secondary lexeme
board may also be an appointed governing body, weight 0.95
Word
management
Lexeme
management, as an appointed governing body
Meaning
an appointed governing body
Lexeme
management, as the act of managing
Meaning
the act of managing

Figure 4: Unidirectional relation: *management* in its first meaning has a partial synonym.

Figure 5 shows the result after the lexicographer does decide to add the same synonym in the opposite direction, only with a different weight. There are now two independent and unrelated secondary lexemes, one for each direction, which again conforms exactly to initial requirements specification.

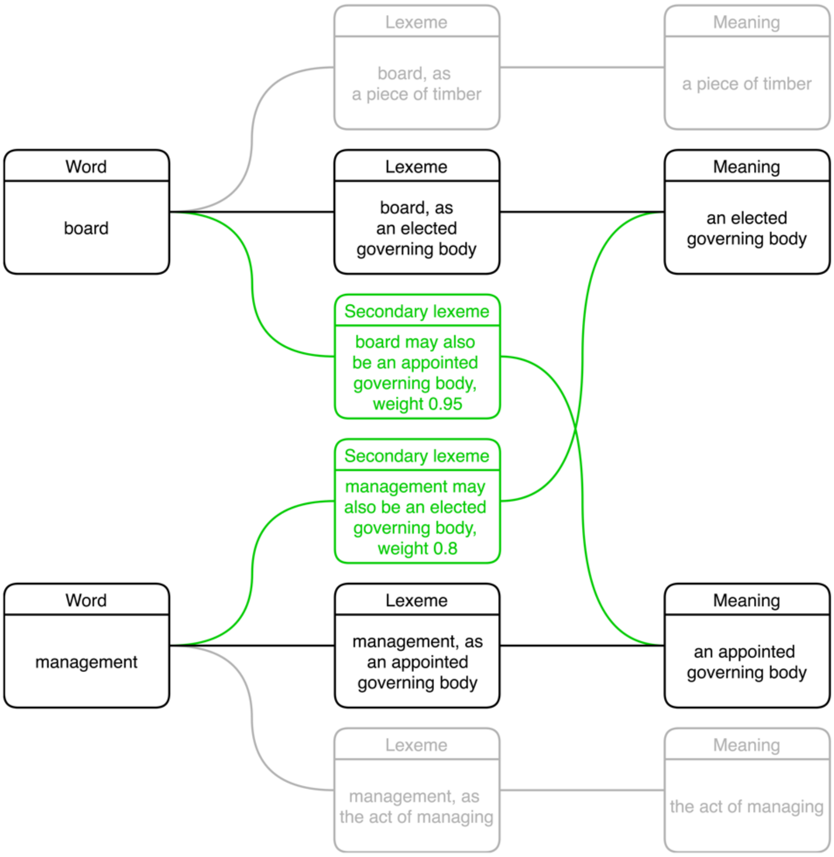
Lexeme
board, as a piece of timber
Meaning
a piece of timber
Word
board
Lexeme
board, as an elected governing body
Meaning
an elected governing body
Secondary lexeme
board may also be an appointed governing body, weight 0.95
Secondary lexeme
management may also be an elected governing body, weight 0.8
Word
management
Lexeme
management, as an appointed governing body
Meaning
an appointed governing body
Lexeme
management, as the act of managing
Meaning
the act of managing

Figure 5: Several unidirectional relations: *management* in its first meaning and *board* in its second meaning each have a partial synonym.

5 Discussion and Conclusions

The experience of the first six months of compiling the synonym layer has shown that providing the lexicographer with an automatically generated list of synonym candidates makes the task of compiling an entry less time-consuming. The candidates are already there in the database, preventing the lexicographer from leaving the dictionary writing system to look up possible synonym candidates from existing dictionaries, thesauri, corpora, etc. What makes the specialised tool especially easy to use is the option of working on a keyboard instead of dragging and dropping the candidates to corresponding senses, as well as tooltips that display the definition of the words when hovering the mouse cursor over the candidates.

On the other hand, feedback from lexicographers over this extended period has provided valuable insights into the design choices, in some cases even casting doubt on the initial requirements.

- In an ideal world, layers as currently conceived would be sufficient to soften the transition from separate dictionaries to CombiDic, allow specialised teams to work on different aspects of the same entry and prevent conflicts. Reality has proven to be different in two ways: current layers are not really independent or even separated clearly enough, and lexicographers need a growing nomenclature of new (sub)layers. This necessitates a reconceptualisation of the layer system.
- The view used for synonym and equivalent layers in Ekilex is narrowly specialised for the simple repetitive task of connecting target words to source meanings. There are or will be other views for other tasks, including the clean-up task described above. Lexicographers, however, prefer to organise their work by headword, not by task, which necessitates either jumping between the specialised views, or adding more and more ad hoc functions to the views, thereby losing the ergonomics benefits of specialisation. We don't have a solution for this at the time of writing.
- While the described unidirectional approach of secondary lexemes exactly conforms to initial requirements and received favourable first impressions from lexicographers, doubts have started to emerge. Especially for large synsets, the need to enter all synonyms again for each member of the set has proven to be a significant drawback. Lexicographers have even submitted bug reports on the grounds that they remember having added a synonym, but the synonym is not there (admittedly, this confusion was amplified by deficiencies of the logging system of Ekilex at the time). Investigation then showed that indeed, the synonym was added, but to another member of the synset. This may mean that the conceptually and procedurally complicated approach of secondary lexemes is not justified after all, and it may be necessary to fall back on the related meanings approach.
- Another indication in the same direction is that for lexicographers, synonymy, antonymy and cohyponymy belong to the same category of semantic relations and should receive similar treatment. The current Ekilex data model differs from this categorisation by treating synonyms in a completely different way. Falling back on the related meanings approach would also even out this difference.
- In bilingual dictionaries, it has been the norm to allow sense distributions of the source headword to be influenced by the target language. In a central dictionary like the CombiDic, this is not sustainable, as there will eventually be many languages. The solution, again, has been agreed to be the related meanings approach described above: each language has its own meanings, and there are links of (partial) equivalence between meanings.
- Since we have limited information about homonyms in other languages, there are massive data quality issues in the target languages. Although some semi-automatic tools can be conceived, achieving a quality level comparable to that of Estonian words (a task that would traditionally be called "reversing" a dictionary) will employ lexicographers for a long period.

6 References

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.

Dyvik, H. (1998). A translational basis for semantics. *Language and Computers*, *24*, 51–86.

Dyvik, H. (2004). Translations as semantic mirrors: from parallel corpus to wordnet. *Language and Computers*,, *49*(1), 311–326.

Eesti keele sõnaraamat 2019 [The Dictionary of Estonain 2019, DicEst]. (2019). Tallinn: Eesti Keele Instituut. Retrieved from https://doi.org/10.15155/3-00-0000-0000-0000-08240L [30.07.2020]

Eesti õigekeelsussõnaraamat 2018 [Dictionary of Standard Estonian 2018, ÕS]. (2018). Tallinn: Eesti Keele Sihtasutus.

EKI ühendsõnastik 2020 [EKI Combined Dictionary 2020, CombiDic]. (2020). Tallinn: Eesti Keele Instituut, Sõnaveeb. Retrieved from https://sonaveeb.ee/ [30.07.2020]

Felber, H. (1984). *Terminology Manual*. Paris: UNESCO.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Gantar, P., Kosem, I., & Krek, S. (2016). Discovering automated lexicography: The case of the Slovene lexical database. *International Journal of Lexicography*, *29*(2), 200–225.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *ArXiv Preprint ArXiv:1802.06893*.

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.

Kallas, J., Koeva, S., Langemets, M., Tiberius, C., & Kosem, I. (2019). Lexicographic Practices in Europe: Results of the ELEXIS Survey on User Needs. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal*. Lexical Computing, pp. 519–536.

Kallas, J., & Koppel, K. (2019). Eesti keele ühendkorpus 2019 [Estonian National Corpus 2019]. Retrieved July 30, 2020, from https://doi.org/10.15155/3-00-0000-0000-0000-08489L [30.07.2020]

Koppel, K., Tavast, A., Langemets, M., & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal*. Lexical Computing, pp. 1–3.

Krek, S., Šorli, M., & Kocjančič, P. (2008). The Funny Mirror of Language: The Process of Reversing the English-Slovenian Dictionary to Build the Framework for Compiling the New Slovenian-English Dictionary. In *Proceedings of the XII EURALEX International Congress. Barcelona: Universitat Pompeu Fabra*, pp. 535–542.

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

Rudnicka, E., Piasecki, M., Bond, F., Grabowski, L., & Piotrowski, T. (2019). Sense Equivalence in plWordNet to Princeton WordNet Mapping. *International Journal of Lexicography*, *32*(3), 296–325.

Tavast, A. (2008). *The Translator is Human Too: A Case for Instrumentalism in Multilingual Specialised Communication*. Tartu: Tartu Ülikooli Kirjastus.

Tavast, A., Langemets, M., Kallas, J., & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data : The Case of EKILEX.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*(1), 141–188.

Wüster, E. (1979). *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. Dordrecht: Springer.

Yong, H., & Peng, J. (2007). *Bilingual Lexicography from a Communicative Perspective*. John Benjamins Publishing.

Acknowledgements

The creation of the dictionary and terminology database Ekilex was funded by EKI-ASTRA program (2016–2022). The creation and development of the portal Sõnaveeb was funded by the Digital Focus Program of the Ministry of Education and Research (2018–2021) and by EKI-ASTRA program (2016–2022). Technical support is provided by OÜ TripleDev. The research received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.