Invisible meaning relations for representing near equivalents

Arvi Tavast[1], Kristina Koppel[1], Margit Langemets[1], Silver Vapper[1], Madis Jürviste[1]

[1] Institute of the Estonian language, Roosikrantsi 6, Tallinn, Estonia
E-mail: first.last@eki.ee

Abstract

One of the key design principles of the Ekilex dictionary writing system is its symmetrical many-to-many relationship between word and meaning. Ekilex is currently being used for creating the EKI Combined Dictionary (CombiDic), with a primary goal of increasing coverage of languages beyond Estonian. This paper discusses the pilot project of integrating English, which began with generating a list of candidate equivalents for post-editing. The primary focus of the paper is on how near equivalents (narrower, wider, approximate) are represented in the symmetrical data model. Since meanings are language-independent entities in such a model, and equivalence is essentially about similarity of meanings, the near equivalents are represented using relations between meanings. To the dictionary user, the relations remain invisible and are only queried to retrieve target-language words for display. Transitioning from the traditional flowing text in the target language field to this more structured approach significantly affects the work process. We examine the advantages and disadvantages of this change in the paper.

Keywords: dictionary writing system; data model; multilingual dictionary; near equivalents

1. Introduction

The task of describing the lexical aspect of language has traditionally been assigned to lexical resources such as general and specialised dictionaries, termbases, lexicons, encyclopedias, and so on. The creation and utilization of these resources is a well-established field, boasting traditions dating back thousands of years. Lexicographers can trace their lineage back to the Sumerian-Akkadian bilingual word lists (Boulanger, 2003, p. 76), while terminologists can trace theirs to the Onomasticon of Amenemope (Boulanger, 2003, p. 111).

The format, structure and data model of lexical resources have remained largely unchanged for millennia, due to the restrictions of the publishing medium, which up until very recently has been a flat, two-dimensional, hierarchical, sequentially-accessed format like paper. The enduring influence of the paper mindset also manifests itself in the data models and creation principles of early electronic lexical resources and data exchange standards (e.g. Budin et al., 2012), with the possible exception of Ontolex-Lemon (McCrae et al., 2017). Three aspects of this heritage are now ripe for reevaluation.

1.1 Hierarchical data model

The lexicographic tradition of listing words (alphabetically or otherwise) and providing each with whatever information the lexicographer deems necessary is particularly ingrained among both lexicographers and readers (Atkins & Rundell, 2008; Flinz, 2011)

Indeed, since paper is not searchable, it needed an access structure built into the very organisation of the lexical resource. Two contrasting solutions were employed: onomasiology (concept-orientation, documenting concepts and their designations, mainly used in terminology) and semasiology (word-orientation, documenting words and their meanings, mainly used in lexicography). This is also the reason for the strict distinction between dictionaries and termbases, based on their method of compilation.

Both of these orientations result in a hierarchy rather than a network. In database terms, they are based on a one-to-many (1:n) relationship, either relating one concept to many terms or one word to many meanings. If there is repeated information in the latter side of the relationship (e.g. the same meaning for synonyms), there is no natural way to express that in the model. Such information can be simply repeated, or addressed with a cross-reference. It is worth noting that serial data exchange formats based on XML or JSON are also inherently hierarchical in this regard.

Maintaining consistency, i.e. guaranteeing that all repetitions are handled purposefully without unnecessary duplication or internal conflicts, has been a challenge for even the most diligent lexicographers. Readers are routinely provided with conflicting information within a single lexical resource. (Tavast, 2008; Tavast & Taukar, 2013)

This simplistic type of relationship, and therefore the opposition between the orientations, was natural and necessary on paper. However, with more expressive formats now available, there is no need to uphold it.

1.2 Directionality

The concept of directionality is deeply entrenched in general language lexicography (e.g. Adamska-Sałaciak, 2014). In this model, one language is designated as the source language, with any others considered target languages. This is based on the understanding that exact equivalence between languages is unattainable in a dictionary of any practical size. Consequently, the target language side can't simply consist of a single word. Instead, it must convey the full richness of the source language using a variety of means: typically more than one equivalent, words that merely suggest the meaning of the source word, rarely used words in the target language, words with domain or register qualifiers, extended explanations, and so on.

Directional compilation raises the issue of dictionary reversal (Sierra, 2000). The experience with the Estonian-Russian dictionary (EVS) reveals that high-frequency

target language words were frequently used as equivalents for non-synonymous headwords. This created a misleading impression that these words had as many meanings as they were used as equivalents for. As this dictionary was compiled semasiologically and directionally, importing it into the non-directional Ekilex system revealed a different perspective on equivalences, which the authors found unsatisfactory.

1.3 Authored work

The third tradition we aim to question is the perception of a dictionary as an authored work, reflecting the views of the author(s) rather than serving as a source of objective information about language. Dictionaries are even granted a certain level of copyright protection (see Langemets & Voll, 2008 for a case study of our own experience).

This perspective has been both convenient and beneficial, acting as a shield: given the impossibility of complete objectivity in language description, the author has full discretion over the dictionary's content. Two authors describing the same language will invariably produce different outputs. The majority of content disagreements can be dismissed by citing the inherent subjectivity of each description.

While the utility of this view on subjectivity is perfectly understandable from the author's perspective, it may not align with the reader's expectations. Although the personal insights of authors can be intriguing, it's reasonable to assume that at least some readers are seeking information about language instead.

2. Background of Ekilex

Since 2017, the Institute of the Estonian Language (EKI) has been developing Ekilex, an in-house dictionary writing system (Tavast et al., 2018, 2021). One of its central design principles is the symmetry of its data model: the many-to-many relationship between word and meaning simultaneously accommodates semasiological and onomasiological resources. It is currently being used for compiling the general dictionary of Estonian – EKI Combined Dictionary (CombiDic) – as well as over 120 termbases. Lexicographers and terminologists are working on the same data, but from opposing viewpoints. Completed resources are accessible to readers via the language portal Sõnaveeb (Koppel et al., 2019).

At the heart of the Ekilex data model is a many-to-many relationship between word and meaning: a word can have multiple meanings and a meaning (or concept) can be designated by multiple words (terms) in several languages.

A word in this model is a language-specific but meaning-agnostic character sequence, containing data elements that do not depend on the meaning, such as language, gender, aspect, morphology, pronunciation, etymology. Conversely, a meaning is a language-agnostic unit of knowledge containing data elements that do not depend on how (or if

at all) this meaning is expressed in any languages, such as domain, semantic type, definition.

To implement a many-to-many relationship between these two main entities in relational database terms, we use a link table. In its purest form, a link table only contains pairs of word IDs and meaning IDs, indicating which word is associated with which meaning. However, during the initial design of the data model, we quickly realised that a substantial proportion of all data categories – ranging from part of speech to example sentences – belong to this link table, rather than to the word or meaning themselves. We defined the link entity as "this word in this meaning as described in this dictionary," and called it a lexeme. A lexeme contains information that depends on the combination of word and meaning, such as part of speech, usage example, collocation, register, and proficiency level.

The number of possible meanings greatly exceeds the number of words in any language. 'The human brain contains eighty-six billion neurons, each with about ten thousand synaptic contacts whose strength can vary. The space of mental representations that opens up is practically infinite.' (Dehaene, 2020, p. 10) A fundamental challenge for creating lexical resources is therefore the need to simplify the continuous reality of language into the discrete representation of a dictionary.

There is an important consequence for the dictionary data model, especially one (like Ekilex) where meanings have their own database entities rather than being represented by free-form text. While the database entities for words correspond non-controversially to words in language and are able to represent their relevant properties (orthography, morphology, etymology, etc.) exhaustively, meanings are more difficult first to individuate and then to describe. Decisions regarding how fine sense distinctions should be and what exactly the senses are, depends on various factors including the volume of the dictionary, purpose, target group and even available funding.

Language	**Words in CombiDic**	**Words in termbases**	**Total words**
Estonian	159,891	141,982	301,873
Russian	172,393	57,531	229,924
English	2,945	89,521	92,466
Latin	4,051	21,766	25,817
German	2,116	17,264	19,380
French	7,983	9,184	17,167
Norwegian	19	14,097	14,116

Ukrainian	11,270	2	11,272
Finnish	1,518	7,943	9,461
Spanish	127	2,257	2,384

Table 1: Languages with at least 1000 words in the datasets of Ekilex (as of 11 April 2023).

Currently, all datasets in Ekilex contain Estonian as one of their languages. Termbases are mostly concept-oriented and consequently directionless, but in CombiDic, where directionality is still pertinent, Estonian has thus far maintained a special status as the pivot language. For lexicographers, this means originating dictionary entries from the Estonian side and adding equivalents in other languages. For readers, this implies that the opposite direction (e.g. English-Estonian) and other combinations (e.g. German-French) are accessible if searched for, but might not have been thoroughly reviewed by a lexicographer. Table 1 lists the most widely covered languages in the datasets of Ekilex. The seemingly random variations are due to external factors, including the availability of existing material (Russian, Norwegian), special status of a language (Latin in life sciences) and recent world affairs (Ukrainian).

One of our purposes has been to increase foreign language coverage in CombiDic. We started a new project in 2021 to semi-automatically add English equivalents. The project had a dual goal: to add the foreign language most widely spoken in Estonia, and to design and test the whole process for adding other languages in the future. The remainder of this paper addresses two challenges:

- Generating a list of candidate English equivalents for the Estonian headwords for manual post-editing by lexicographers.
- Integrating multiple bilingual dictionaries into the Ekilex data model and systematically managing their interrelations within the model.

3. Generating candidate equivalents

To add English equivalents to the Estonian headwords in CombiDic using a process of post-editing lexicography (Jakubíček et al., 2018, 2021), a dataset of possible candidates was automatically generated. We used two existing English-Estonian dictionaries: the English-Estonian Machine Translation Dictionary compiled by Indrek Hein of the Institute of the Estonian Language, and the Password Estonian-English Glossary compiled by K Dictionaries in cooperation with the Institute and the publishing house TEA (Langemets et al., 1999; Kernerman, 2015). To ensure wider vocabulary coverage, we gathered possible equivalents from parallel corpora.

Equivalents were gathered by processing sentence pairs and doing word alignments using the ArgMax matching method (Sabet et al., 2021), which were then gathered in frequency lists. We chose this specific algorithm based on a sample 'gold standard' parallel data set for Estonian–English word alignments.

Two types of sources based on presumed translation quality were used: a proprietary corpus based on professional translation memories, and publicly available corpora. Detecting potential candidates from publicly available corpora led to a lot of noise in the data, e.g. candidates including numbers, symbols, foreign alphabets and punctuation marks that were all automatically deleted before importing into Ekilex. While frequency lists based on translation memories included less noise than those from public corpora, they still required substantial reductions and filtering.

Most of the additional filtering was based on statistical relevance and heuristics taken from random samples of data. For instance, we removed candidates with a frequency of 1 or 2 from headwords with more than five different equivalent candidates, as these low-frequency matches were almost always incorrect.

When importing the candidate equivalents, we set the threshold from 5 to 30. When a headword had fewer than five candidates, we imported all of them, even if the frequency was 1. Prior to importing the data into Ekilex, we combined the corpus and dictionary data, assigning weights to candidates based on their origin. These weights determined the visible order of equivalent candidates in Ekilex. We also appended metadata—such as part of speech information, example usages, and definitions—to candidates sourced from dictionaries."

4. Bilingual data in a many-to-many data model

In this section, we discuss how bilingual dictionaries fit in the many-to-many data model of Ekilex. Specifically, we detail three key insights this model provides to bilingual dictionary authors, along with their associated costs and benefits.

4.1 Model structure

The bilingual dictionaries under discussion here belong to general language lexicography, which has traditionally employed a semasiological data model. The central entity in such a model is the word, with its senses branching out hierarchically (each word has one or more senses). As the Ekilex data model is symmetrical between word and meaning, and meaning has its own set of language-independent properties, we can transcend this simple hierarchy.

The essence of equivalence is a meaning relation: equivalent words share the same meaning. The Ekilex many-to-many data model represents these situations using a single mechanism, connecting the word entities to the same meaning entity. A meaning

has two words of the same language in the case of full synonymy, and of different languages in the case of full equivalence.

An immediate objection is that in language reality, perfect equivalence between languages or absolute synonymy within a language is an extremely rare and possibly non-existent phenomenon (Lyons, 1981; Cruse, 1986; Murphy, 2003; Pym, 2010). The meaning of a lexical item is not even identical across speakers of a single language, and keeps developing during the lifespan of a single individual as exposure to linguistic input accumulates (Ramscar et al., 2013, 2014).

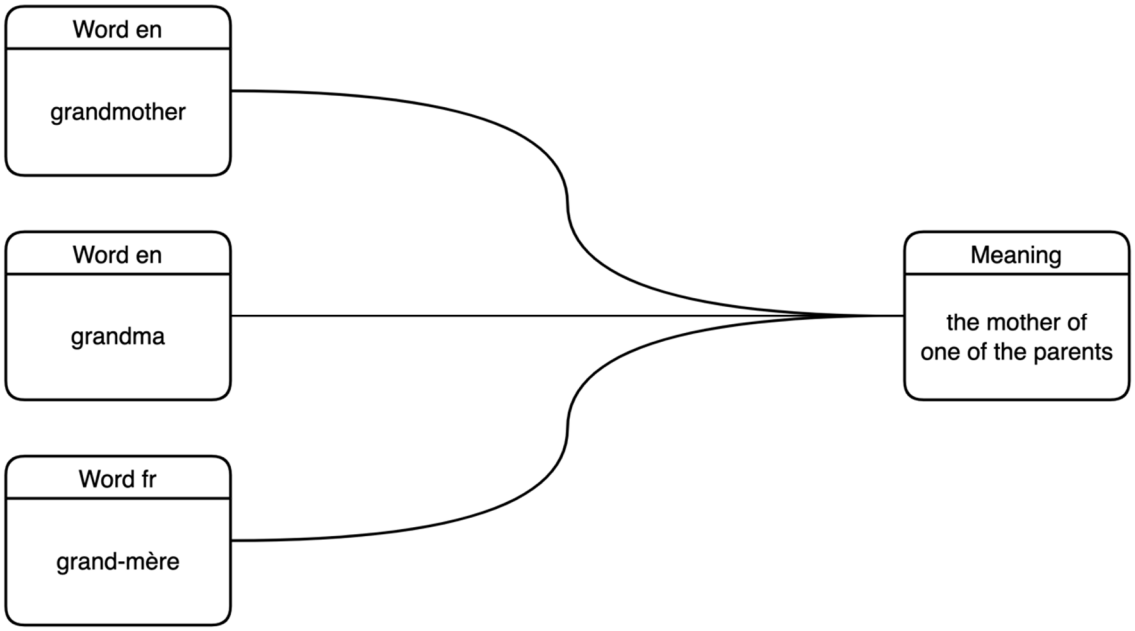

Figure 1: Full synonymy (between *grandmother* and *grandma*) and equivalence (between these and *grand-mère*) represented as all three words having the same meaning.

It is both a lexicographic tradition and an inevitable need to simplify language reality in order to fit it into the finite form of a dictionary. This includes claiming full equivalence or synonymy between lexical items with meanings that the lexicographer considers sufficiently close, as shown on Figure 1. A rule of thumb used in practice is to see full equivalence or synonymy only when the definition is exactly the same. So what we are changing in the case of full equivalents, is only the technical implementation, not the lexicographic principle.

This paper is concerned with the next step: what if the meanings are so different that they can't possibly be simplified into a claim of full equivalence, but still close enough to qualify as candidates for being represented as some sort of equivalents in a bilingual dictionary? Recurrent examples of this include the following:

- Meanings that are not lexicalised in one of the languages, or where the target-language word is too rare for inclusion in the dictionary. Example: *'grandmother'*

in Swedish, where one needs to specify between *mormor* 'mother's mother' and *farmor* 'father's mother' instead.

- Meanings that are culturally different but are still considered to somehow correspond to each other, at least within the precision limits of the dictionary. Example: French *pain*, English *bread* and German *Brot* may be equivalent in a very broad sense, but they are culturally different enough in their shape, colour, texture and taste to warrant a more detailed treatment in a more advanced dictionary.

Our current solution to this situation is that the data model stays the same, each word still has its own meaning (exact meaning, given the level of simplification chosen for the dictionary), and there is a similarity relation between those meanings. So instead of representing that these *words* are similar in their meaning, we represent that these words have *meanings* that are similar, see Figure 2.

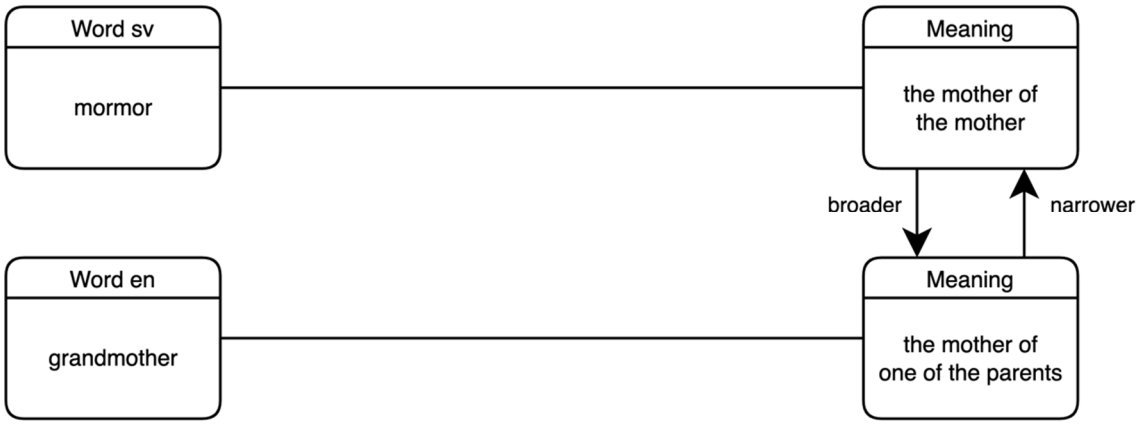

Figure 2: Narrower and wider equivalents represented as a meaning relation.

We have chosen to use three types of meaning relations for representing near equivalents: narrower *(A > B)*, wider *(A < B)* and approximately same *(A ≈ B)*, where *A* and *B* designate meanings.

Representing near equivalents with meaning relations has the counter-intuitive consequence that not all meanings have designations in all languages. For the reader, these meaning relations themselves remain invisible, but are traversed in order to retrieve the corresponding target language words and render a habitual presentation of near equivalents.

4.2 Directionality

Cross-linguistic equivalence is symmetrical by nature. From $A = B$, it inevitably follows that $B = A$, and any claim to the contrary is motivated solely by lexicographic tradition.

Outside the dictionary, languages are not inherently "source" or "target". The same lexical resource can be queried[1] in any direction, and should yield sensible results regardless of the direction, even if the equivalences are not exact. Table 2 lists situations that arise when near equivalence relations are viewed from directions other than the original premise that the lexicographer had in mind.

Premise	**Inference**	**Description**
$A \approx B$	$B \approx A$	Approximate equivalence is symmetrical. If *bread* is almost equivalent to *Brot*, then *Brot* is almost equivalent to *bread*.
$A > B$	$B < A$	Wider and narrower are opposites. If *mormor* is a narrower equivalent for *grandmother*, then *grandmother* is a wider equivalent for *mormor*.
$A > B$ and $B = C$	$A > C$	Adding more languages requires coordination between all languages. If *mormor* is a narrower equivalent for *grandmother*, and *grandmother* is a full equivalent for *grand-mère*, then *mormor* is a narrower equivalent for *grand-mère*.

Table 2: Types of relations between the meanings of words of three languages: A, B and C.

The possibility of being queried in any direction could also be described as automatic and immediate reversal of the bilingual dictionary, which understandably complicates the lexicographer's task. It is no longer sufficient or even possible to use the target language field for any explanatory information that comes to mind (equivalent, approximate equivalent, several equivalents based on meaning nuances or usage patterns, explanation in case there is no equivalent, etc.). The following additional tasks need to be considered:

- Separation of data types. Each data element needs to go to its own field, rather than as flowing text in a single large field. The fields may not even belong to the same database entity, e.g. it is important to distinguish between properties of the target word in this meaning (e.g. register) and the meaning itself (e.g. domain, definition). In our experience, this has proven to be difficult already in a monolingual situation, and the situation will be further complicated with

[1] In our own resources, this was the case already before Ekilex, e.g. in the Estonian-Finnish http://www.eki.ee/dict/efi or the Estonian-Russian https://portaal.eki.ee/dict/evs/.

more languages, as described below.

- In the process of entering an equivalent, the lexicographer needs to immediately consider all properties of the target word, including definitions and example sentences, what other meanings does the target word have, or where else it has been or will be used as an equivalent. Especially the latter potentially creates a rabbit hole for the lexicographer to fall down, in the style of Dyvik's semantic mirrors (1998, 2004). The work process thus needs to accommodate the following of mirrored chains of equivalence, limit their depth somehow, or include a separate step for cleaning up the opposite language direction.

4.3 Authorship

It has traditionally often been the case that bilingual dictionaries (e.g. Estonian-English and Estonian-French) are separate works authored by non-overlapping groups of lexicographers, even if they share one of the languages. This organisation of work is incompatible with the understanding that equivalence is about meanings: full equivalents share the same meaning and partial equivalents have related meanings. Meanings are independent of languages and especially of language pairs. To continue with the example used above, it is difficult to imagine how the assertion that "mother's mother is a type of grandmother" could depend on the language(s) in question, so it should be safe to enter it as a language-independent meaning relation.

Consequently, equivalence information entered by the team working on one language pair has an effect on all other language pairs. Here are some situations from our initial experience where this may become an issue:

- An assertion may or may not correspond to facts of life, or its degree of simplification may be debatable. However, both its truth value and the suitability of the degree of simplification remain language-independent. If some factual claim needs correcting, then it needs correcting for all languages, which in turn requires coordination between the teams of all languages.

- The need to express the meaning relation in the first place does depend on specific languages, in this case Swedish. Without Swedish, full equivalence between *grandmother*, *grand-mère*, *Großmutter* etc. would probably be sufficient for a general dictionary. Once Swedish is added, though, the number of related meanings is increased from one to three, and all language teams need to decide whether the added meanings of mother's mother and father's mother require a word in their language. For Estonian, as an example, they might, as the words *emaema* 'mother's mother' and *isaema* 'father's mother' do exist, even if used much less frequently than *vanaema* 'grandmother'.

- Adding even more languages may introduce more distinctions based on

parameters that were unlexicalised in previously added languages, e.g. whether the grandparent is living or deceased. Two issues may arise here: cooperation will be needed between languages making the same distinction, and intersecting multiple distinctions may result in a network of relationships that is difficult to understand as a whole.

- One of the methods for starting a new dictionary project is to import material from existing dictionaries. If these are traditional enough, they will probably contain manually written textual solutions for representing near equivalents, e.g. a sentence explaining that Swedish distinguishes between maternal and paternal grandmother. If this sentence is imported for one language *and* the meaning relations are created for another language, then the same information will appear twice in different wordings for the reader. Again, cooperation is required, and rephrasing or even simply removing such duplication may involve significant amounts of work.

- As the number of authors increases with the number of languages, they will more frequently introduce changes that may affect other languages. Staying on top of the flow of changes will require either an alerting system or periodic "sanity check" queries from the database. In both cases it depends on non-trivial organisational decisions about what kind of changes need the attention of other languages. A balance between overwhelming numbers of notifications and the danger of missing an important change needs to be worked out in practice.

So if we continue with the assumption that lexicographers are human (as opposed to artificial intelligence) and therefore limited in their language proficiency in all the language pairs that may need a bilingual dictionary, the only way forward is cooperation.

A recurring request that EKI receives from potential dictionary teams is to use Ekilex for authoring a stand-alone unidirectional bilingual dictionary, often with Estonian as the target language. While granting such requests would be technically possible in the same way that specialised dictionaries are created as stand-alone works, we have chosen not to. We invite them to cooperate with the CombiDic team to add their language(s) to CombiDic instead.

The objective is to eventually have hundreds of languages in CombiDic, with the consequence that the dictionary will have hundreds, if not thousands of authors contributing to various languages, some of them professionally, but many sporadically. The potential challenge of managing such a huge team, both organisationally and regarding intellectual property rights, is acknowledged, but is outside the scope of this paper.

5. Summary

As we have started to add new languages to CombiDic, the symmetrical data model of Ekilex has brought about a number of changes compared to traditional bilingual lexicography.

Our primary objective has been to develop a data structure intended to unify and formalise relationships of both full and near equivalence. Although the meaning relations proposed in this paper will remain invisible, the dictionary users stand to benefit from these in the form of better considered and coordinated equivalents.

We began by detailing the process of generating candidate equivalents for post-editing lexicography and subsequently explored the costs and benefits of the symmetrical data model for integrating a multitude of languages into CombiDic. Given that equivalence fundamentally pertains to meanings, it is represented at the meaning level in the database. Full equivalents relate to the same meaning, while each near equivalent has its own meaning, with these meanings being interrelated. Currently, we employ three types of meaning relations: wider, narrower, and approximate.

The flip side of the benefits of better coordination and uniformity of lexicographic principles is the required change in the work process. Entering more information or more thoroughly considered information is inevitably more labour-intensive than the habitual approach of entering much less information. The only reason for undertaking such change is to eventually provide a superior dictionary for the user.

As we are in the first phases of adding the pilot language (English), there a lot to learn about the data model and the work process, especially how both unfold in practice.

6. Acknowledgements

This work was supported by the Estonian Research Council grant PRG 1978.

7. References

Adamska-Sałaciak, A. (2014). Explaining Meaning in a Bilingual Dictionary. In P. Durkin (Ed.), *The Oxford handbook of lexicography. Oxford: Oxford University Press (forthcoming).*

Atkins, S. B. T., & Rundell, M. (2008). *The Oxford guide to practical lexicography.* Oxford University Press.

Boulanger, J.-C. (2003). *Les Inventeurs de dictionnaires: De l'eduba des scribes mésopotamiens au scriptorium des moines médiévaux.* Les Presses de l'Université d'Ottawa.

Budin, G., Majewski, S., & Mörth, K. (2012). Creating Lexical Resources in TEI P5. *Journal of the Text Encoding Initiative*, *Issue 3*, Article Issue 3.

https://doi.org/10.4000/jtei.522

CombiDic: Langemets, M., Hein, I., Jürviste, M., Kallas, J., Kiisla, O., Koppel, K., Kuusk, K., Leemets, T., Mäearu, S., Paet, T., Päll, P., Raadik, M., Risberg, L., Tammik, H., Tavast, A., Tiits, M., Tsepelina, K., Tuulik, M., Valdre, T., Viks, Ü., Sai, E., Smirnova, A., Tubin, V., (2022). EKI ühendsõnastik 2022 [EKI Combined Dictionary 2022]. Sõnaveeb: Eesti Keele Instituut.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge university press.

Dehaene, S. (2020). *How We Learn: The New Science of Education and the Brain*. Penguin UK.

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.

EVS: Laasi, H., Lagle, T., Leemets, H., Liiv, M., Pärn, H., Simm, L., Viks, Ü., Õim, A., Kallas, J., Kotova, N., Melts, M., Matt, K., Tubin, V., (auth.); Liiv, M., Melts, N., Romet, A., Kallas, J., Riikoja, E., Martoja, I., Smirnov, S., Tetsov, M., Tiits, M., Valdre, T., Veskimägi, E., (ed.) (2019). Eesti-vene sõnaraamat 2019. (1. trükk 1997–2009.) 2., täiendatud ja kohandatud veebiväljaanne [Estonian-Russian dictionary 2019]. Sõnaveeb: Eesti Keele Instituut.

Flinz, C. (2011). The microstructure of Online Linguistics Dictionaries: Obligatory and facultative elements. *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of ELex 2011, 10-12 November 2011, Bled, Slovenia*, 83–88.

Jakubíček, M., Kovář, V., & Rychlý, P. (2021). Million-Click Dictionary: Tools and Methods for Automatic Dictionary Drafting and Post-Editing. *EURALEX XIX*.

Jakubíček, M., Měchura, M., Kovar, V., & Rychly, P. (2018). Practical post-editing lexicography with lexonomy and sketch engine. *The XVIII EURALEX International Congress*, 65.

Kernerman, I. (2015). A multilingual trilogy: Developing three multi-language lexicographic datasets. *Dictionaries: Journal of the Dictionary Society of North America*, *36*(1), 136–149.

Koppel, K., Tavast, A., Langemets, M., & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. *Electronic Lexicography in the 21st Century. Proceedings of the ELex 2019 Conference. 1-3 October 2019, Sintra, Portugal*, 1–3.

Langemets, M., Tiits, M., Uibo, U., Pihlak, A., & Kernerman, L. (Eds.). (1999). *Password: Inglise-eesti seletav sõnaraamat = English dictionary for speakers of Estonian* (6. tr). TEA Kirjastus.

Langemets, M., & Voll, P. (2008). Sõnaraamatu kohtulingvistiline analüüs: Eesti pretsedent [Linguistic forensic analysis of a dictionary: an Estonian precedent]. *Eesti Rakenduslingvistika Ühingu Aastaraamat*, *4*, 67–86.

Lyons, J. (1981). *Language, meaning and context*.

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* (pp. 587–597). Lexical Computing CZ s.r.o. https://elex.link/elex2017/proceedings-download/

Murphy, M. L. (2003). *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge University Press.

Pym, A. (2010). *Exploring Translation Theories*. Routledge.

Ramscar, M., Hendrix, P., Love, B., & Baayen, R. H. (2013). Learning is not decline: The mental lexicon as a window into cognition across the lifespan. *The Mental Lexicon*, *8*(3), 450–481.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, R. H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, *6*(1), 5–42.

Sabet, M. J., Dufter, P., Yvon, F., & Schütze, H. (2021). *SimAlign: High Quality Word Alignments without Parallel Training Data using Static and Contextualized Embeddings* (arXiv:2004.08728). arXiv. https://doi.org/10.48550/arXiv.2004.08728

Sierra, G. (2000). The onomasiological dictionary: A gap in lexicography. *Proceedings of the Ninth Euralex International Congress*, 223–235.

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

Tavast, A., Koppel, K., Langemets, M., & Kallas, J. (2021). Towards the Superdictionary: Layers, Tools and Unidirectional Meaning Relations. *EURALEX XIX*.

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

Tavast, A., & Taukar, M. (2013). *Mitmekeelne oskussuhtlus*. Valgus.