Towards the ELEXIS data model: defining a common vocabulary for lexicographic resources

Carole Tiberius[1], Simon Krek[2], Katrien Depuydt[1], Polona Gantar[3], Jelena Kallas[4], Iztok Kosem[2], Michael Rundell[5]

[1] Instituut voor de Nederlandse Taal, Leiden, The Netherlands
[2] Jožef Stefan Institute, Ljubljana, Slovenia
[3]Faculty of Arts, University of Ljubljana, Slovenia
[4]Institute of the Estonian Language, Tallinn, Estonia
[5]Lexical Computing, Brno, Czech Republic
E-mail: {carole.tiberius,katrien.depuydt}@ivdnt.org, {simon.krek,iztok.kosem}@ijs.si, apolonija.gantar@ff.uni-lj.si, jelena.kallas@eki.ee, michael.rundell@gmail.com

Abstract

In this paper we describe ongoing work on the identification and definition of core lexicographic elements to be used in the ELEXIS data model. ELEXIS is a European infrastructure project fostering cooperation and information exchange among lexicographical research communities. One of the main goals of ELEXIS is to make existing lexicographic resources available on a significantly higher level than is currently the case. Therefore, a common data model is being developed which aims to: a) streamline the integration of lexicographic data into the infrastructure (using the ELEXIFIER tool), b) enable reliable linking of the data in the ELEXIS Dictionary Matrix, and c) provide a basic template for the creation of new lexicographic resources, such that they can automatically benefit from the tools and services provided by the ELEXIS infrastructure. Here we focus on the development of a common vocabulary and report on the results of an initial survey that was conducted to collect feedback from experts in lexicography.

Keywords: data model; common vocabulary; lexicographic resource; interoperability

1. Introduction

Reliable and accurate information on word meaning and usage is of crucial importance in today's information society. The most consolidated and refined knowledge on word meanings can traditionally be found in dictionaries – monolingual, bilingual or multilingual. In each and every European country, elaborate efforts are put into the development of lexicographic resources describing the language(s) of the community. Although confronted with similar problems relating to technologies for producing and making these resources available, cooperation on a larger European scale has long been limited. In addition, standardisation efforts have not been particularly successful within the field of lexicography before the digital age, an observation which was confirmed by the results from the ELEXIS[1] survey on lexicographic practices in Europe (Kallas et al., 2019). More specifically, the results from the survey show that:

- most lexicographic projects use structured data, but some projects are still working with a non-structured data and text format;
- proprietary XML and (customised) TEI are the most commonly used XML formats;
- use of existing standard vocabularies for encoding lexicographic data is not yet common practice at the ELEXIS lexicographic partner institutions. IsoCat, GOLD, and TEI were mentioned.

As a consequence, the lexicographic landscape in Europe is still rather heterogeneous. It is characterised by stand-alone lexicographic resources and there is a significant variation in the level of expertise and resources available to lexicographers across Europe. This situation forms a major obstacle to more ambitious, innovative, transnational, data driven approaches to dictionaries, both as tools and objects of research.

The ELEXIS project aims to overcome these obstacles by developing a sustainable infrastructure for lexicography. To allow all different kinds of dictionary data to be included in the infrastructure and ensure that it will be open to a wide range of lexicographers, common protocols have been developed and a common vocabulary is being defined, which is the topic of this paper. Before we turn to the ongoing work on the ELEXIS data model and more specifically the common vocabulary in section 3, we will first introduce the ELEXIS project in more detail in section 2. In section 4 we discuss the results of a pilot survey that was conducted to get feedback from lexicographic experts on the common vocabulary.

2. ELEXIS

ELEXIS (Krek et al., 2018, 2019; Pedersen et al., 2018; Woldrich et al., 2020) is a Horizon 2020 project dedicated to creating a sustainable infrastructure for lexicography. The main objectives of the infrastructure are to:

1. enable efficient access to high quality lexical data/semantic information in the digital age;
2. bridge the gap between more advanced and lesser-resourced scholarly communities working on lexicographic resources;
3. enable the use of new technology and data in industry in the digital single market.

[1] https://elex.is/

Within ELEXIS, strategies, tools and standards are under development for extracting, structuring and linking lexicographic resources to unlock their full potential for Linked Open Data, NLP and the Semantic Web, as well as in the context of digital humanities. In a virtuous cycle of cross-disciplinary exchange of knowledge and data, a higher level of language description and text processing will be achieved. By harmonising and integrating lexicographic data into the Linked Open Data cloud, ELEXIS will make this data available to AI and NLP for semantic processing of unstructured data, considerably enhancing applications such as machine translation, machine reading and intelligent digital assistance thanks to the ability to scale to wide coverage in multiple languages. This, in turn, will enable the development of improved tools for the production of structured proto-lexicographic data in an automated process, using machine learning, data mining and information extraction techniques, where the extracted data can be used as a starting point for further processing either in the traditional lexicographic process or through crowdsourcing platforms.

Lexicographic data is crucial for realising the ELEXIS infrastructure. Within ELEXIS, data comes from a number of different data providers, i.e.:

- Consortium partners
- Observer institutions
- Other open access resources containing lexicographic data available through, amongst others, CLARIN and DARIAH.

To date, 118 different datasets, e.g. general dictionaries, bilingual dictionaries, thesauri, specialised dictionaries (terminology, dialects), and lemma lists have been collected from 32 ELEXIS partner and observer institutions. A sample list of the datasets can be found in the ELEXIS Deliverable 6.3 Intermediate interoperability report.

Most of these datasets have been compiled within national and regional projects, and as noted they are typically encoded in their own custom data format, i.e. proprietary XML, (customised) TEI, HTML, JSON-LD or are stored in a relational database. A growing number also have API access. To be able to integrate these diverse datasets in the ELEXIS infrastructure a set of common protocols have been developed (McCrae et al., 2019) and different access routes are distinguished into the infrastructure. Data can be contributed either as TEI Lex-0 or Ontolex-Lemon, which are the two data formats supported by ELEXIS. It is also possible to deliver data as proprietary XML or in another format. Proprietary XML data can take advantage of the ELEXIFIER tool which converts custom XML or PDF into TEI Lex-0 (see Section 2.2). Those contributing data in another format can create an implementation of the REST interface according to the specifications provided by ELEXIS (ELEXIS Deliverable 2.2 Interoperable interface for Lemon and TEI resources; McCrae et al., 2019).

Having a set of common protocols ensures what Ide and Pustejovsky (2010) call syntactic interoperability, which "relies on specified data formats, communication protocols, and the like to ensure communication and data exchange. It means that the systems involved can process the exchanged information, but there is no guarantee that the interpretation is the same". This means that an element labelled 'example' in dataset X is not necessarily the same as an element labelled 'example' in Y. If we want to be able to link, edit, enrich and publish data from various sources reliably (as envisaged in ELEXIS, see Figure 1), we also need semantic interoperability.

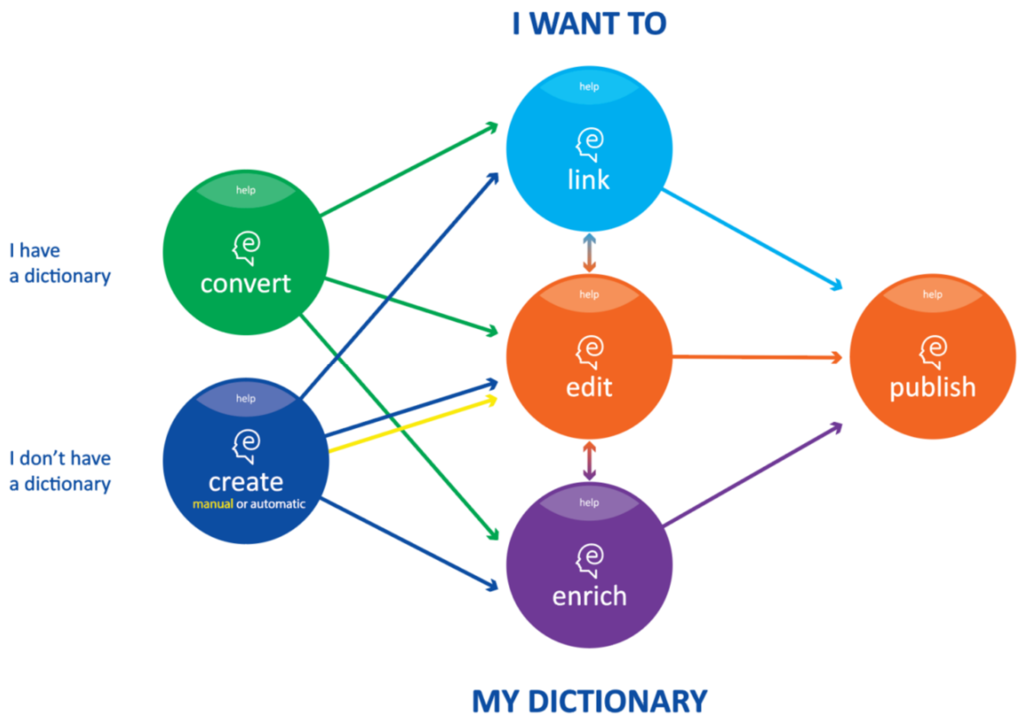


Figure 1: Graphic guide to the ELEXIS Dictionary Tools

According to Ide and Pustejovsky (2010) "semantic interoperability exists when two systems have the ability to automatically interpret exchanged information meaningfully and accurately in order to produce useful results via deference to a common information exchange reference model". The first step towards such a model is the definition of a common vocabulary (see section 3), which is needed among others in the ELEXIFIER tool and the ELEXIS Dictionary Matrix.

2.1 ELEXIFIER

ELEXIFIER[2] (Repar et al., 2020) is a cloud-based dictionary conversion service for converting legacy dictionaries into a shared data format so that it can be integrated in the ELEXIS infrastructure. It can take lexicographic data in two distinct formats as input: (1) custom XML and (2) PDF. In the custom XML scenario, XPath formalisms

[2] https://elexifier.elex.is/

are used for identifying the core elements in the original dictionary data and transforming these to a TEI Lex-0 compliant format. All information contained in the original dictionary is kept, and only the core elements are transformed to the shared format. The supported elements are the same as those defined in the common vocabulary.

In the PDF scenario a more complex process is needed. The PDF is first transformed in a flat structure using a pdf2xml conversion script (based on https://github.com/kermitt2/pdf2xml). Then, a chunk of the resulting XML file is sent to Lexonomy[3] (Měchura 2017), an online dictionary editing tool for manual annotation. Approximately four pages need to be annotated. The annotated text is then used as the training material for machine learning algorithms that produce the entire dictionary converted to TEI Lex-0 compliant format. Dictionaries that have been transformed using ELEXIFIER, can be edited further in Lexonomy.

2.2 ELEXIS Dictionary Matrix

One of the main results of ELEXIS will be the ELEXIS Dictionary Matrix: a universal repository of linked senses, meaning descriptions, collocations, phraseology, translation equivalents, examples of usage and other types of lexical information found in existing lexicographic resources, monolingual, multilingual, modern, historical etc., available through a RESTful web service developed as part of LEX1 infrastructure. LEX1 is the part of the ELEXIS infrastructure which consists of a set of services and tools dedicated to the automatic segmentation, structuring, alignment and conversion of lexicographic resources to a uniform data format. The existence of common data models and standards that are produced bottom-up from within the lexicographic community fostered by ELEXIS is a necessary condition for successful development of this segment of the infrastructure.

The ELEXIS Dictionary Matrix will be also available as part of the Linguistic Linked Open Data cloud (LLOD), and it will serve as the source for providing links to (particular headwords, senses, etc. in dictionaries available online, through the European Dictionary Portal[4], and included in the matrix.

3. ELEXIS Data Model

To support the development of the Dictionary Matrix, a common data model is being developed which aims to a) streamline the integration of lexicographic data into the infrastructure (using the ELEXIFIER tool, see section 2.1) b) enable reliable linking of the data in the Dictionary Matrix (see section 2.2.), and c) provide a basic template

[3] https://www.lexonomy.eu/

[4] http://www.dictionaryportal.eu/

for the creation of new lexicographic resources, allowing for a smooth integration of new content into the matrix.

The aim of ELEXIS is not to develop a fully-fledged data model. Neither does the project aim to replace existing models. The main goal is to ensure semantic interoperability between lexicographic resources predominantly using their own custom format, focusing on a set of core elements which are necessary for the development of the Dictionary Matrix.

As a first step towards the development of the ELEXIS data model, efforts have been taken to establish a common vocabulary where the main concepts are unambiguously defined.

3.1 ELEXIS Common Vocabulary

As a starting point, a detailed analysis of sample data (provided by ELEXIS lexicographic partners and observer institutions) was carried out resulting in the following core elements: entry, headword, secondary headword, variant headword, part of speech, sense, sense structure, definition, sense indicator, label, example, translation, cross reference, note and inflected form. Table 1 gives an overview of the elements identified and their definitions. The overall strategy was to keep definitions as simple and as unambiguous as possible.

Element	**Definition**
entry	Part of a lexicographic resource which contains information related to at least one headword.
headword	Organising element of an entry in a lexicographic resource. *Note: In printed dictionaries typically at the top of an entry.*
secondary headword	Headword-like lexical item occurring within an entry in a lexicographic resource, for example derived forms, feminine forms, multiword expressions. Often an organising element of a part of an entry.
variant headword	Lexical item representing one of the alternative forms of the headword, for example a spelling or regional variation.
part of speech	Any of the word classes to which a lexical item may be assigned, e.g. noun, verb, adjective, etc.
sense	Part of an entry which groups together information relating to a meaning of a headword (or secondary headword), for example definitions, examples, and translations.

sense structure	Division and ordering of the senses in an entry.
definition	Statement that describes a meaning and permits its differentiation from other meanings within a sense structure of an entry.
sense indicator	Short statement that gives an indication of a meaning and permits its differentiation from other meanings within a sense structure of an entry.
label	Item from a controlled vocabulary indicating some kind of restriction on the use of the lexical item, for example, time, region, domain, register.
example	Instance of a lexical item's usage in a specific sense.
translation	Equivalent in another language of any element in an entry.
cross reference	Element providing any kind of link or reference to another element within or outside the lexicographic resource.
note	Free text remark that can accompany any element in a lexicographic resource.
inflected form	Form of the inflectional paradigm of the headword.

Table 1. ELEXIS core elements

In addition to the core elements, the following terms have been defined as they are used in the definitions of the core elements or they are potentially relevant in the context of ELEXIS:

Term	**Definition**
lexicographic resource	Needs to be defined; see section 4.1.
lexical item	Any word, abbreviation, partial word, or phrase which is described or mentioned in an entry in a lexicographic resource.
word class	A category of words grouped together based on form, meaning or syntactic characteristics.
meaning	The unique semantic, grammatical and/or pragmatic contribution that a headword in a particular sense makes to the overall understanding of an utterance.
controlled vocabulary	Fixed list of items which are used to reduce ambiguity and ensure consistency.
multiword expression	Sequence of lexical items that has properties that may not be predictable

	from the properties of the individual lexical items or their normal mode of combination. For example, collocations, phrasemes, compounds, idiomatic expressions, lexical combinations, and so forth. A multiword expression can have the status of headword or secondary headword in the lexicographic resource.
source language	The language of a lexical item (that is to be translated in another language). [cf. ISO1951:2007]
target language	The language into which a lexical item is to be translated. [cf. ISO1951:2007]

Table 2: Terms used in the definitions of the ELEXIS core elements

The next steps are to refine and finalise the definitions for these core elements and to express the ELEXIS data model in a formalism like UML. This way the serialisations to the two ELEXIS interoperability formats, i.e. Ontolex-Lemon and TEI Lex-0 can be realised.

Work on the ELEXIS data model is done in collaboration with the Lexicographic Infrastructure Data Model and API (LEXIDMA) Technical Committee within OASIS[5].

3.2 Related work

The ELEXIS data model does not stand on its own. In the past decade, several institutions and organisations have started harmonising the internal workflow trying to arrive at a uniform data model to be used for all lexicographic projects within the institution (e.g. Kernerman 2011, Depuydt et al. 2019; Parvizi et al., 2016; Tavast et al., 2018). Other larger initiatives which are particularly relevant to ELEXIS are TEI Lex-0 with a special focus on retrodigitised dictionaries, Ontolex-Lemon, the *de facto* standard for representing lexical information as RDF, and LMF (Lexical Markup Framework) which is being developed by the ISO Technical Committee (TC) 37 titled 'Language and terminology'.

The ISO 24613 LMF multipart standard is based upon the definition of an implementation-independent metamodel combining a core model with extensions. As such it provides mechanisms that allow the development and integration of a variety of electronic lexical resource types and its scope is therefore much broader than that of the ELEXIS model.

The TEI Lex-0 (Tasovac et al., 2018) initiative aims at establishing a baseline encoding and a target format to facilitate the interoperability of heterogeneously encoded lexical

[5] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=lexidma

resources. As specified in the TEI Lex-0 rationale[6], TEI Lex-0 should be primarily seen as a format which implements a set of constraints on top of those provided by the TEI Guidelines so that existing TEI dictionaries, once univocally transformed, can be queried, visualised, or mined in a uniform way. Furthermore, TEI Lex-0 aims to stay as aligned as possible with the TEI subset developed in conjunction with the revision of the ISO LMF standard (cf. Romary, 2015), ensuring future interoperability and sustainability.

Ontolex-Lemon (Cimiano et al., 2016) was originally developed to act as a model for the representation of lexical information in ontologies and is now the *de facto* standard for representing lexical information as RDF. It is also widely used to present data from lexicographic resources as Linked Data on the web. However, a mapping of traditional dictionary content to Ontolex-Lemon was not feasible without the development of an additional model, to be able to represent aspects of dictionaries like order and hierarchy of senses, or the fact that there is not always a 1:1 match between a dictionary entry and an ontolex:LexicalEntry (which requires it to have only one part of speech). The Lexicog module[7] is aimed to deal with these issues.

Both TEI Lex-0 and Ontolex-Lemon are supported within ELEXIS and serialisations will be provided from and to both TEI Lex-0 and Ontolex-Lemon. In addition, a tei2ontolex[8] conversion stylesheet has been developed.

4. Survey on the ELEXIS core elements and their definitions

A pilot survey was set up in order to collect feedback from experts in lexicography on the ongoing work on the common vocabulary. The survey was conducted in the autumn of 2020. It was sent to the lexicographic experts on the ELEXIS international advisory board and to the lexicographic partners in the project.

As it was a pilot survey, the goal was primarily qualitative rather than quantitative. Therefore, none of the questions in the survey was made obligatory and additional comments could be given for almost all questions. The survey was implemented in the 1ka survey system[9] which has been used for several other surveys within ELEXIS.

Only the following core elements were included in the pilot – entry, headword, secondary headword, sense, sense structure, definition, translation and example. For each of these a separate section was created in the survey where the relevant definitions were given together with a few extracts from existing dictionaries (see Figures 2-12).

[6] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

[7] https://www.w3.org/2019/09/lexicog/

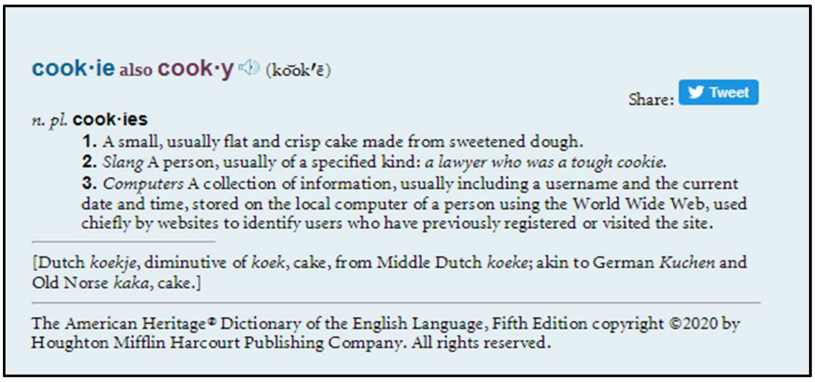
[8] https://github.com/elexis-eu/tei2ontolex

[9] https://www.1ka.si/

The lexicographic experts were then asked to answer questions about the element in relation to these extracts and the definitions provided. In order to get a wide range of examples, extracts were taken from various monolingual, bilingual, general-purpose, and also specialised dictionaries. Average completion time of the survey was 15 minutes, and we received 10 valid responses. Although this is undoubtedly a small number of responses, the results clearly show what the bottlenecks are when trying to define and identify core elements in lexicography. In the remainder of this section we discuss the results from this initial survey.

4.1 Entry

For 'entry', three extracts from three completely different dictionaries (traditional, born-digital, and specialised) were given: one from the American Heritage Dictionary[10] (see Figure 2), one from dictionary.com[11] (see Figure 3), and one from The Right Rhymes[12] (see Figure 4), a dictionary of hip-hop language.



cook·ie also **cook·y** (ko͞ok′ē)

Share: Tweet

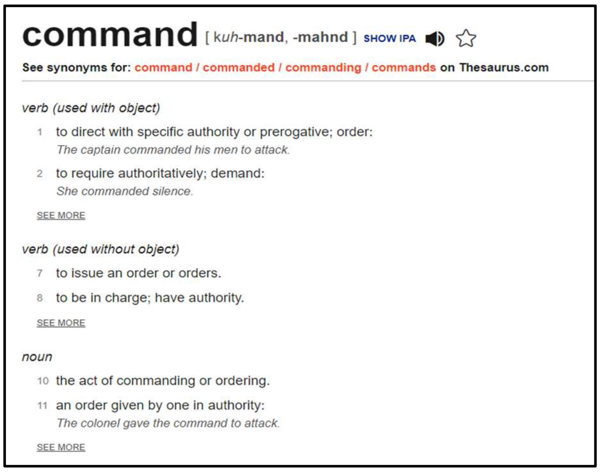
n. pl. **cook·ies**

1. A small, usually flat and crisp cake made from sweetened dough.
2. *Slang* A person, usually of a specified kind: *a lawyer who was a tough cookie.*
3. *Computers* A collection of information, usually including a username and the current date and time, stored on the local computer of a person using the World Wide Web, used chiefly by websites to identify users who have previously registered or visited the site.

[Dutch *koekje*, diminutive of *koek*, cake, from Middle Dutch *koeke*; akin to German *Kuchen* and Old Norse *kaka*, cake.]

The American Heritage® Dictionary of the English Language, Fifth Edition copyright ©2020 by Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Figure 2: Extract from the American Heritage Dictionary



command [*kuh*-mand, -mahnd] SHOW IPA ☆

See synonyms for: command / commanded / commanding / commands on Thesaurus.com

verb (used with object)

1 to direct with specific authority or prerogative; order:
The captain commanded his men to attack.

2 to require authoritatively; demand:
She commanded silence.

SEE MORE

verb (used without object)

7 to issue an order or orders.

8 to be in charge; have authority.

SEE MORE

noun

10 the act of commanding or ordering.

11 an order given by one in authority:
The colonel gave the command to attack.

SEE MORE

Figure 3: Extract from dictionary.com

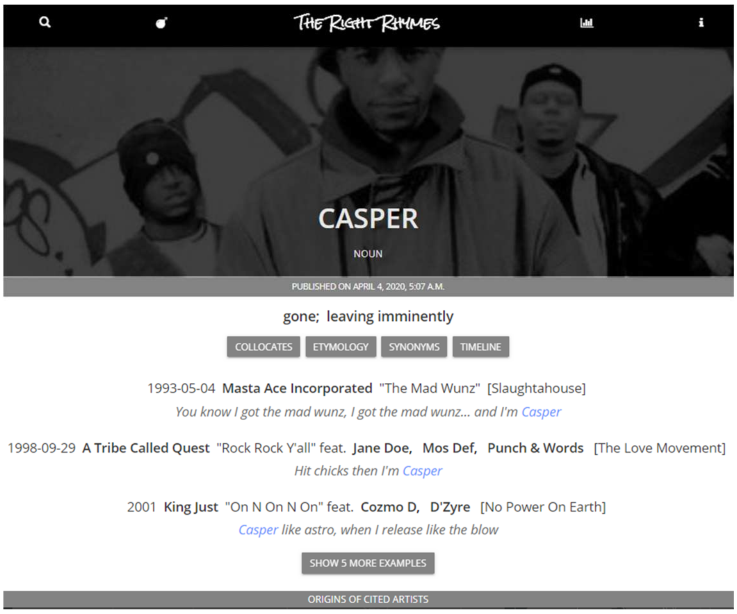
[10] https://www.ahdictionary.com/word/search.html?q=cookie

[11] https://www.dictionary.com/browse/command

[12] https://therightrhymes.com/casper

All experts considered the extract from the American Heritage dictionary in Figure 2 as an ‘entry’ according to the definition provided.

In relation to the extract from dictionary.com, one respondent noted that this should be considered as two entries, one for the verb and one for the noun. Indeed, one of the macrostructural decisions lexicographers need to make relates to what is considered as a homograph and how to treat them.[13]



CASPER

NOUN

PUBLISHED ON APRIL 4, 2020, 5:07 A.M.

gone; leaving imminently

1993-05-04 **Masta Ace Incorporated** "The Mad Wunz" [Slaughtahouse]

You know I got the mad wunz, I got the mad wunz... and I'm Casper

1998-09-29 **A Tribe Called Quest** "Rock Rock Y'all" feat. **Jane Doe, Mos Def, Punch & Words** [The Love Movement]

Hit chicks then I'm Casper

2001 **King Just** "On N On N On" feat. **Cozmo D, D'Zyre** [No Power On Earth]

Casper like astro, when I release like the blow

Figure 4: Extract from The Right Rhymes

There was more disagreement on the extract from The Right Rhymes. One expert felt that it did not fulfil the definition of ‘entry’ because it does not seem to be part of a lexicographic resource and only contains headword and part of speech information. Another respondent also found it difficult to consider this an entry. This shows that there are different views on what counts as a lexicographic resource,[14] and this term also needs to be defined. Some lexicographers/linguists may not consider a dictionary such as The Right Rhymes a lexicographic resource.

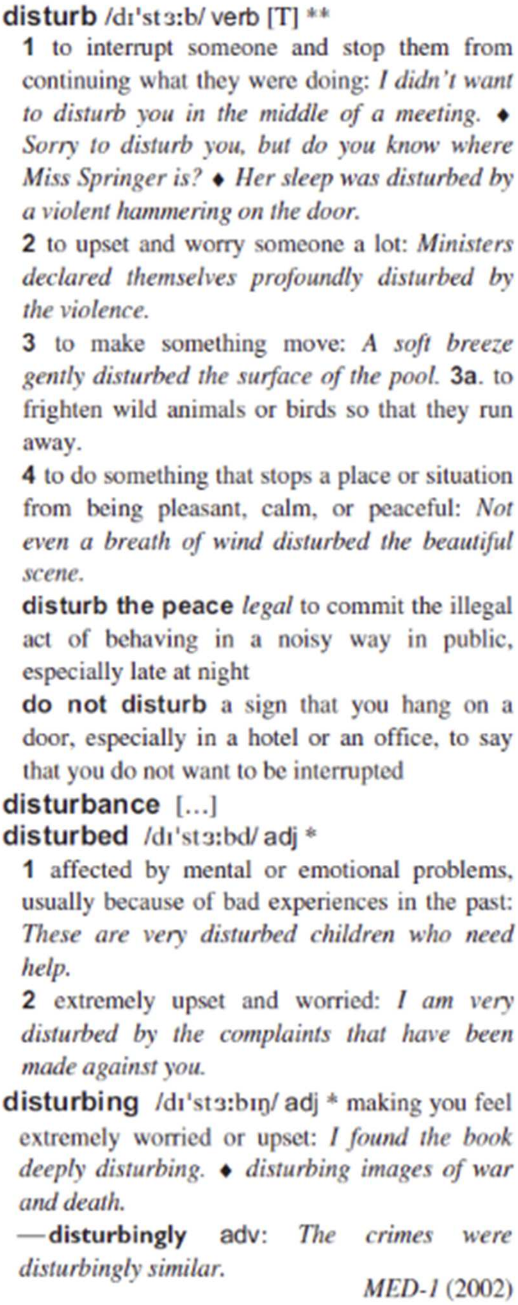
4.2 Headword and secondary headword

The questions on ‘headword’ and ‘secondary headword’ were combined. Again, three extracts from different dictionaries were given: the verb entry for *disturb* from the Macmillan English Dictionary (2002) (see Figure 5), the noun entry for *Katze* ‘cat’ from

[13] See e.g. Atkins and Rundell (2008: 192-193) for criteria that are used in lexicography in relation to homographs to decide whether there should be one entry or more and the discussion in Svensén (2009: 94-102) on the establishment of lemmas.

[14] A lexicographic resource was not yet defined at the time of the survey and thus not included.

the DWDS dictionary[15] (see Figure 6), and the entry for *ohulaada* 'make smth firm' from the Webonary Lynyole dictionary[16] (see Figure 7). The experts were asked to indicate whether they considered various lexical items from these extracts as 'headword', 'secondary headword' or something else.



disturb /dɪˈstɜːb/ verb [T] **
1 to interrupt someone and stop them from continuing what they were doing: *I didn't want to disturb you in the middle of a meeting.* ♦ *Sorry to disturb you, but do you know where Miss Springer is?* ♦ *Her sleep was disturbed by a violent hammering on the door.*
2 to upset and worry someone a lot: *Ministers declared themselves profoundly disturbed by the violence.*
3 to make something move: *A soft breeze gently disturbed the surface of the pool.* **3a**. to frighten wild animals or birds so that they run away.
4 to do something that stops a place or situation from being pleasant, calm, or peaceful: *Not even a breath of wind disturbed the beautiful scene.*
disturb the peace *legal* to commit the illegal act of behaving in a noisy way in public, especially late at night
do not disturb a sign that you hang on a door, especially in a hotel or an office, to say that you do not want to be interrupted
disturbance [...]
disturbed /dɪˈstɜːbd/ adj *
1 affected by mental or emotional problems, usually because of bad experiences in the past: *These are very disturbed children who need help.*
2 extremely upset and worried: *I am very disturbed by the complaints that have been made against you.*
disturbing /dɪˈstɜːbɪŋ/ adj * making you feel extremely worried or upset: *I found the book deeply disturbing.* ♦ *disturbing images of war and death.*
—**disturbingly** adv: *The crimes were disturbingly similar.*

MED-1 (2002)

Figure 4: Extract from the Macmillan English Dictionary taken from Atkins and Rundell (2008: 36). The experts were asked whether *disturb, disturb the peace, do not disturb, disturbance, disturbed, disturbing* and *disturbingly* are a 'headword', 'secondary headword' or something else.

[15] https://www.dwds.de/

[16] https://www.webonary.org/lunyole/

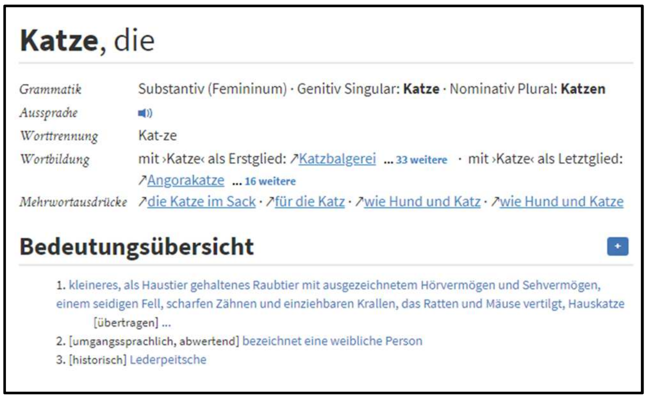
For the extract from the Macmillan English Dictionary (see Table 3) there was complete agreement on *disturb* being a ‘headword’, but the opinions on the status of *disturb the peace*, *do not disturb* varied significantly. Approximately half of the experts considered these as a ‘secondary headword’ whereas the other half considered them as something else.

	headword	**secondary headword**	**something else**[17]
disturb	10		
disturb the peace		5	4
do not disturb		4	6
disturbed	9	1	
disturbing	9	1	
disturbingly		9	1

Table 3. Experts’ decisions on ‘headword’/ ‘secondary headword’/ something else

When the option ‘something else’ was chosen, terms such as phrase, collocation, idiom and derivative forms were given to describe the item. It was also mentioned that structurally these items can be considered as ‘(secondary) headwords’ as in the tagging structure they represent discrete blocks, but that conceptually they should be tagged for what they are, e.g. an idiom block, a phrasal verb block or a run-on. It was also pointed out that this type of structural choice (that has been done for search-engine-friendly reasons) divorces the phrase or idiom from its context, from the environment of its source "word".

In the entry for *Katze* (see Figure 5) the results for the hyperlinked items *Katzbalgerei* und *wie Hund und Katze* were mixed. The reason that was given several times for calling these something else was that they look like cross-references to other entries and that the user thus has to go to another page to view them.



Katze, die

Grammatik Substantiv (Femininum) · Genitiv Singular: **Katze** · Nominativ Plural: **Katzen**
Aussprache
Worttrennung Kat-ze
Wortbildung mit ›Katze‹ als Erstglied: ↗Katzbalgerei ... 33 weitere · mit ›Katze‹ als Letztglied: ↗Angorakatze ... 16 weitere
Mehrwortausdrücke ↗die Katze im Sack · ↗für die Katz · ↗wie Hund und Katz · ↗wie Hund und Katze

Bedeutungsübersicht

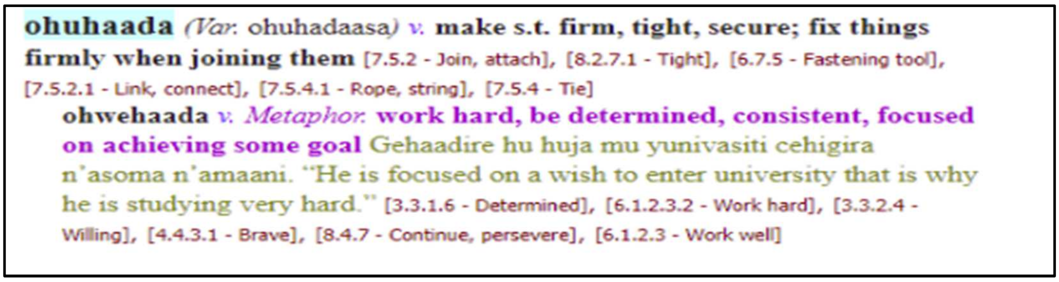
1. kleineres, als Haustier gehaltenes Raubtier mit ausgezeichnetem Hörvermögen und Sehvermögen, einem seidigen Fell, scharfen Zähnen und einziehbaren Krallen, das Ratten und Mäuse vertilgt, Hauskatze
   [übertragen] ...
2. [umgangssprachlich, abwertend] bezeichnet eine weibliche Person
3. [historisch] Lederpeitsche

Figure 5: Entry for *Katze* ‘cat’ in the DWDS dictionary[18]. The experts were asked whether *Katzen* ‘cats’, *Katzbalgerei* ‘scuffle’, *wie Hund und Katze* ‘like dog and cat’ are ‘headword’, ‘secondary headword’ or something else.

[17] As it was not made obligatory to check a box for each item, the numbers do not add up.

[18] https://www.dwds.de/wb/Katze

Experts did agree on the third extract containing the entry for *ohuhaada* from the Webonary Lynyola dictionary (see Figure 6), considering *ohuhadaasa* as a 'headword' and *ohwehadaa* as a 'secondary headword'. To the question as to whether there were any other items that could be considered as a 'secondary headword' in this entry, one respondent mentioned *ohuhadaasa* (the form in between brackets given after the 'headword').



ohuhaada *(Var.* ohuhadaasa*)* *v.* **make s.t. firm, tight, secure; fix things firmly when joining them** [7.5.2 - Join, attach], [8.2.7.1 - Tight], [6.7.5 - Fastening tool], [7.5.2.1 - Link, connect], [7.5.4.1 - Rope, string], [7.5.4 - Tie]

ohwehaada *v. Metaphor.* **work hard, be determined, consistent, focused on achieving some goal** Gehaadire hu huja mu yunivasiti cehigira n'asoma n'amaani. "He is focused on a wish to enter university that is why he is studying very hard." [3.3.1.6 - Determined], [6.1.2.3.2 - Work hard], [3.3.2.4 - Willing], [4.4.3.1 - Brave], [8.4.7 - Continue, persevere], [6.1.2.3 - Work well]

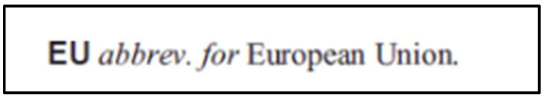
Figure 6: Entry for *ohuhaada* 'make smth firm' in the Lynyole dictionary[19]. The experts were asked whether *ohuhaada* and *ohwehaad*a are 'headword', 'secondary headword' or something else.

These results show that the definition of 'secondary headword' may need to be refined or at least further explained if we want to get a consistent transformation for this element in the ELEXIFIER tool across different datasets.

4.3 Part of Speech

As noted by Svensén (2009: 136), "there is considerable variation between languages, lexicographic traditions and user categories as concerns the occurrence, format and function of part-of-speech indications". This can also be observed in the survey results where experts noted that it is a tricky question as to whether something like *transitive verb* should be considered as a 'part of speech' or as two separate labels. Most respondents noted that strictly speaking *verb* is the 'part of speech' and *transitive* additional information. However, it was also noted that if it is the style of the dictionary to conflate two concepts in a single element, then it is a 'part of speech'. Similar observations were made in relation to *proper noun.*

With part of speech there are clear cases, but there are also some problematic cases, as is illustrated by the extract in Figure 7.



EU *abbrev. for* European Union.

Figure 7: Entry for *EU* in the Collins English Dictionary (2000) (Atkins and Rundell, 2008: 196)

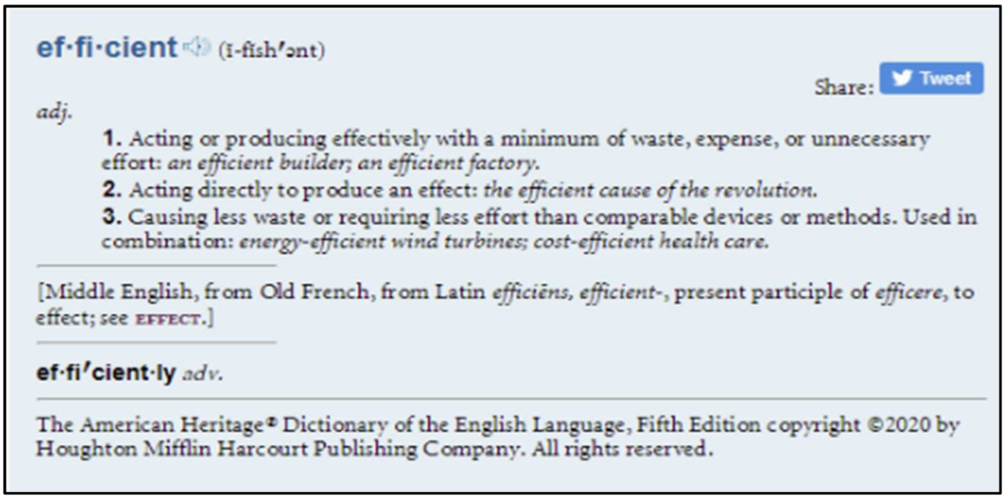
[19] https://www.webonary.org/lunyole?s=ohuhaada

Only three experts considered *abbrev.* as a ‘part of speech’, whereas seven marked it as something else. The reason for this is clearly summarised by one expert:

> “If you want to split hairs, abbreviations, acronyms, etc. aren’t really a separate word class; the underlying part of speech is whatever the thing they’re an abbreviation for is. But in terms of listing this information in the header information of the dictionary, you’ll find that most dictionaries put this kind of indicator inside POS tags.”

4.4 Sense and sense structure

To learn more about the perception of ‘sense’ and ‘sense structure’, we took an extract from the American Heritage Dictionary illustrating the entry for *efficient*[20] (Figure 8).



ef·fi·cient (ĭ-fĭsh′ənt)

adj.

1. Acting or producing effectively with a minimum of waste, expense, or unnecessary effort: *an efficient builder; an efficient factory.*
2. Acting directly to produce an effect: *the efficient cause of the revolution.*
3. Causing less waste or requiring less effort than comparable devices or methods. Used in combination: *energy-efficient wind turbines; cost-efficient health care.*

[Middle English, from Old French, from Latin *efficiēns, efficient-*, present participle of *efficere*, to effect; see EFFECT.]

ef·fi′cient·ly *adv.*

The American Heritage® Dictionary of the English Language, Fifth Edition copyright ©2020 by Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Figure 8: Entry for *efficient* in the American Heritage Dictionary

There was full agreement that the numbers 1., 2. and 3. represent the ‘sense structure’. There was, however, quite some disagreement on what actually constitutes a ‘sense’, as shown in Table 4.

	Sense	Something else
2. Acting directly to produce an effect: *the efficient cause of the revolution.*	5	4
2. Acting directly to produce an effect	6	3
Acting directly to produce an effect	6	4
Acting directly to produce an effect: *the efficient cause of the revolution.*	3	5

Table 4: Experts’ decisions on whether the options provided are a ‘sense’ or something else

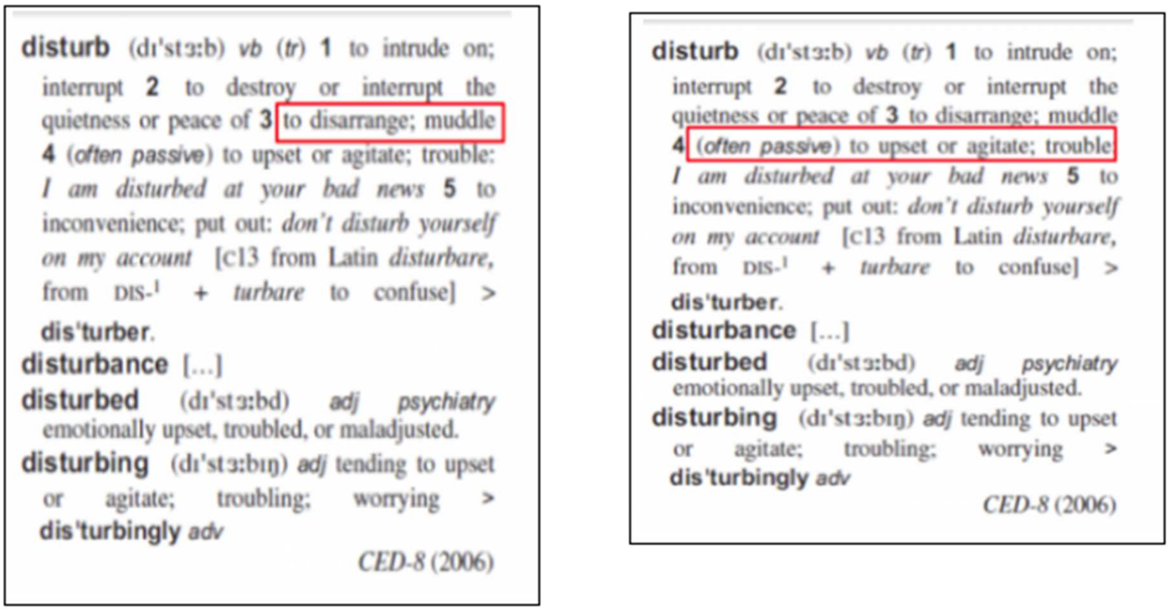
[20] https://www.ahdictionary.com/word/search.html?q=efficient

Four possible variants were provided and there was actually none that all the experts agreed on. Some considered the inclusion of the example necessary for it to be a 'sense' (which is in line with the definition provided), others mentioned the presence of a sense number (unless numbering is automatic) and for some, 'sense' itself is the definition. The latter was motivated by stating that structurally, explanatory examples are part of the sense and tend to be included in the sense block in a tagging structure. They can illustrate the sense, but they are not truly the sense.

These answers suggest that there is an interplay between how elements are commonly marked in dictionary structures and how lexicographers think about them conceptually.

4.5 Definition

In relation to the 'definition' element, we were particularly interested to find out whether information which is sometimes included in brackets is considered as part of the definition or not. Two extracts, both from Atkins and Rundell (2008) were taken, one from the Collins English Dictionary (see Figure 9) one from the Oxford Advanced Learner's Dictionary (see Figure 10).



disturb (dɪ'stɜːb) *vb* (*tr*) **1** to intrude on; interrupt **2** to destroy or interrupt the quietness or peace of **3** to disarrange; muddle **4** (*often passive*) to upset or agitate; trouble: *I am disturbed at your bad news* **5** to inconvenience; put out: *don't disturb yourself on my account* [c13 from Latin *disturbare*, from DIS-[1] + *turbare* to confuse] > **dis'turber**.
disturbance [...]
disturbed (dɪ'stɜːbd) *adj psychiatry* emotionally upset, troubled, or maladjusted.
disturbing (dɪ'stɜːbɪŋ) *adj* tending to upset or agitate; troubling; worrying > **dis'turbingly** *adv*

CED-8 (2006)

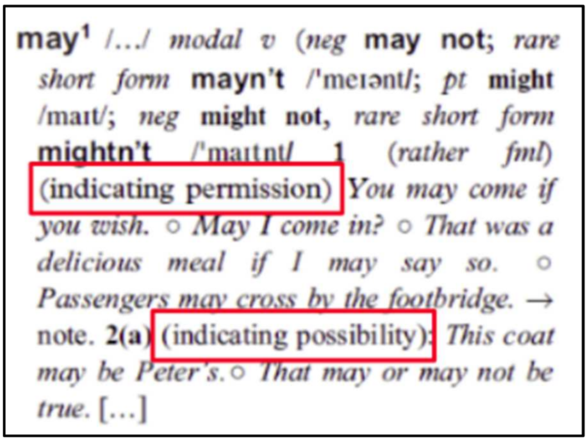
disturb (dɪ'stɜːb) *vb* (*tr*) **1** to intrude on; interrupt **2** to destroy or interrupt the quietness or peace of **3** to disarrange; muddle **4** (*often passive*) to upset or agitate; trouble *I am disturbed at your bad news* **5** to inconvenience; put out: *don't disturb yourself on my account* [c13 from Latin *disturbare*, from DIS-[1] + *turbare* to confuse] > **dis'turber**.
disturbance [...]
disturbed (dɪ'stɜːbd) *adj psychiatry* emotionally upset, troubled, or maladjusted.
disturbing (dɪ'stɜːbɪŋ) *adj* tending to upset or agitate; troubling; worrying > **dis'turbingly** *adv*

CED-8 (2006)

Figure 9: Entry for *disturb* in the Collins English Dictionary (2006) (Atkins and Rundell 2008: 36)

The text in the marked red box on the left hand side was considered a 'definition' by all lexicographic experts, the text in the marked red box on the right hand side by three only, while the others indicated that the information in brackets is grammatical or usage information.

We also included an extract containing a function word or what Atkins and Rundell (2008:196-198) call a grammatical word entry, as these entries often describe the function rather than the meaning.



may¹ /…/ *modal v* (*neg* **may not**; *rare short form* **mayn't** /ˈmeɪənt/; *pt* **might** /maɪt/; *neg* **might not**, *rare short form* **mightn't** /ˈmaɪtnt/ **1** (*rather fml*) (indicating permission) *You may come if you wish.* ○ *May I come in?* ○ *That was a delicious meal if I may say so.* ○ *Passengers may cross by the footbridge.* → note. **2(a)** (indicating possibility): *This coat may be Peter's.* ○ *That may or may not be true.* […]

Figure 10: Part of the entry for *may* in The Oxford Advanced Learner's Dictionary (1995) from Atkins and Rundell (2008: 197).

Seven experts would call the parts marked by the red box a 'definition', but three would not, as they considered these as semantic comments or comments on semantic implicatures.

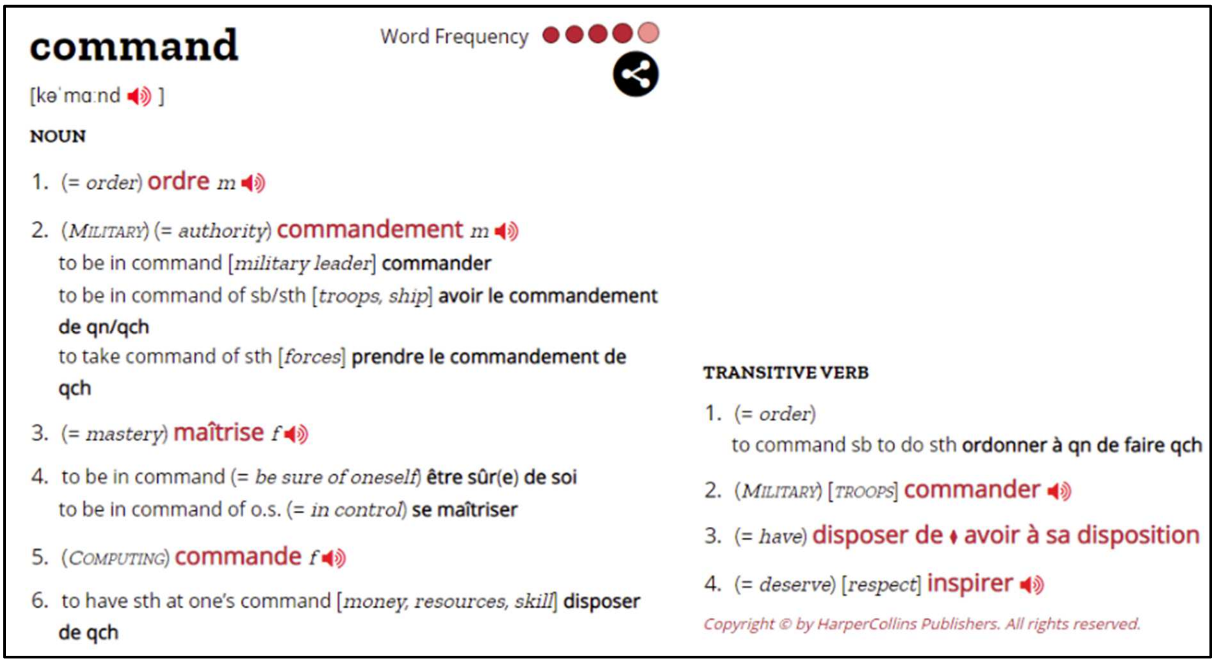
4.6 Translation and Example

For 'translation', an extract from the bilingual English-French Collins Dictionary[21] was selected (see Figure 11).

There was complete agreement among the experts. All considered the three items that were offered *ordre, être sûr(e) de soi*, and *disposer de, avoir à sa disposition* as 'translation'. One noted that the last one actually contains two translations.

For the 'example' element, one extract from a modern dictionary (the Collins Dictionary English-French) and one extract from a historical dictionary (Petit Larousse Illustré) were selected.

[21] https://www.collinsdictionary.com/dictionary/english-french/command



command Word Frequency ●●●●●

[kəˈmɑːnd]

NOUN

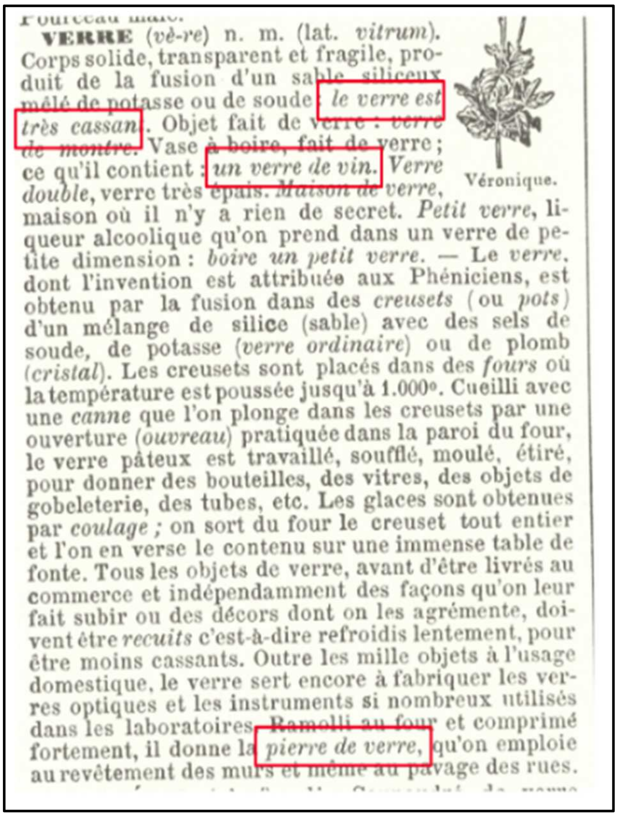
1. (= *order*) **ordre** *m*
2. (MILITARY) (= *authority*) **commandement** *m*
   to be in command [*military leader*] **commander**
   to be in command of sb/sth [*troops, ship*] **avoir le commandement de qn/qch**
   to take command of sth [*forces*] **prendre le commandement de qch**
3. (= *mastery*) **maîtrise** *f*
4. to be in command (= *be sure of oneself*) **être sûr(e) de soi**
   to be in command of o.s. (= *in control*) **se maîtriser**
5. (COMPUTING) **commande** *f*
6. to have sth at one's command [*money, resources, skill*] **disposer de qch**

TRANSITIVE VERB

1. (= *order*)
   to command sb to do sth **ordonner à qn de faire qch**
2. (MILITARY) [TROOPS] **commander**
3. (= *have*) **disposer de** ♦ **avoir à sa disposition**
4. (= *deserve*) [*respect*] **inspirer**

Copyright © by HarperCollins Publishers. All rights reserved.

Figure 11: Entry for *command* in the English-French Collins Dictionary



VERRE (*vè-re*) n. m. (lat. *vitrum*). Corps solide, transparent et fragile, produit de la fusion d'un sable siliceux mêlé de potasse ou de soude : *le verre est très cassant*. Objet fait de verre : *verre de montre*. Vase à boire, fait de verre ; ce qu'il contient : *un verre de vin*. *Verre double*, verre très épais. *Maison de verre*, maison où il n'y a rien de secret. *Petit verre*, liqueur alcoolique qu'on prend dans un verre de petite dimension : *boire un petit verre*. — Le *verre*, dont l'invention est attribuée aux Phéniciens, est obtenu par la fusion dans des *creusets* (ou *pots*) d'un mélange de silice (sable) avec des sels de soude, de potasse (*verre ordinaire*) ou de plomb (*cristal*). Les creusets sont placés dans des *fours* où la température est poussée jusqu'à 1.000°. Cueilli avec une *canne* que l'on plonge dans les creusets par une ouverture (*ouvreau*) pratiquée dans la paroi du four, le verre pâteux est travaillé, soufflé, moulé, étiré, pour donner des bouteilles, des vitres, des objets de gobeleterie, des tubes, etc. Les glaces sont obtenues par *coulage* ; on sort du four le creuset tout entier et l'on en verse le contenu sur une immense table de fonte. Tous les objets de verre, avant d'être livrés au commerce et indépendamment des façons qu'on leur fait subir ou des décors dont on les agrémente, doivent être *recuits* c'est-à-dire refroidis lentement, pour être moins cassants. Outre les mille objets à l'usage domestique, le verre sert encore à fabriquer les verres optiques et les instruments si nombreux utilisés dans les laboratoires. Ramolli au four et comprimé fortement, il donne la *pierre de verre*, qu'on emploie au revêtement des murs et même au pavage des rues.

Figure 12: The extract from Petit Larousse Illustré 1905

The answers to the question with the extract from the modern dictionary did not reveal anything unexpected. For the historical dictionary there was a little uncertainty on whether the last item marked by a red box in Figure 12 was an 'example' or something else.

Only seven experts gave an answer for *pierre de verre* and only three of those considered this an 'example'. The "reluctance" to answer may also suggest that some simply did not know what to answer.

The pilot survey clearly showed certain bottlenecks and as such provided useful feedback on the common vocabulary. The elements 'secondary headword', 'part of speech', and 'sense' in particular need further work. The survey also emphasised the importance of supporting the common vocabulary with concrete examples. In the near future, we will extend the survey to all elements from the ELEXIS common vocabulary and to a larger audience.

5. Summary and further work

In this paper we described ongoing work on the ELEXIS data model. We focussed on the description of the common vocabulary and discussed the results of a pilot survey that was conducted among lexicographic experts. In the near future, the pilot survey will be extended to all elements from the common vocabulary and a larger audience so that we get a more complete insight into the understanding of the core elements in the lexicographic community. This will undoubtedly lead to revisions and refinements in the work on the data model.

In the next phase, it will also be necessary to express the ELEXIS data model in a formalism like UML, in order to realise the serialisation to the two ELEXIS interoperability formats, i.e. Ontolex-Lemon and TEI Lex-0. When the model is finished, a full mapping will also be provided with the related models (TEI Lex-0, Ontolex-Lemon and LMF).

The work on the ELEXIS data model and the common vocabulary is ongoing, and a lot remains to be done, but we hope that it will inspire a constructive debate on standardisation in the lexicographic community and related fields.

6. Acknowledgements

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

7. References

Atkins, S.B.T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Bosque-Gil, J & Gracia, J. (eds.) (2019). *The OntoLex Lemon Lexicography Module Final Community Group Report 17 September 2019.* Accessed at: https://www.w3.org/2019/09/lexicog/. (9 April 2021)

Cimiano, P, McCrae, J.P. & Buitelaar, P. (2016) *Lexicon Model for Ontologies: Community Report, 10 May 2016 Specification.* Accessed at: https://www.w3.org/2016/05/ontolex/. (9 April 2021).

Depuydt, K., Schoonheim, T. & de Does, J. (2019) Towards a More Efficient Workflow for the Lexical Description of the Dutch Language. Accessed at: http://videolectures.net/elexisconference2019_depuydt_dutch_language/. (9 April 2021)

Ide, N. & Pustejovsky, J. (2010). What does interoperability mean, anyway? Toward an operational definition of interoperability. *Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010).* Hong Kong, Available at: http://www.cs.vassar.edu/~ide/papers/ICGL10.pdf.

ISO 1951:2007 *Presentation/representation of entries in dictionaries – Requirements, recommendations and information.*

ISO/CD 24613-1:2018(E) *Language resource management — Lexical markup framework (LMF) — Part 1: Core model.*

ISO/CD 24613-2:2019(E) *Language resource management — Lexical markup framework (LMF) — Part 2: Machine Readable Dictionary (MRD) model.*

ISO/WD 24613-3:2020(E) *Language resource management — Lexical Markup Framework (LMF) — Part 3: Etymological Extension.*

ISO/WD 24613-4:2020 *Language resource management — Lexical Markup Framework (LMF) — Part 4: TEI serialisation.*

ISO NP 24613-5:2018 *Language resource management — Lexical markup framework (LMF) — Part 5: Lexical base exchange (LBX) serialization.*

Kallas, J., Koeva, S., Langemets, M., Tiberius, C. & Kosem, I. (2019). Lexicographic practices in Europe: Results of the ELEX survey on user needs. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 Conference, 1–3 October 2019, Sintra, Portugal.* Available at: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_30.pdf.

Kernerman, I. (2011). From Dictionary to Database: Creating a Global Multi-Language Series. In I. Kosem & K. Kosem (eds.) *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2011 Conference, 11-12 November 2011, Bled Slovenia.* Available at: https://elex2011.trojina.si/Vsebine/proceedings/eLex2011-14.pdf.

Kosem, I, Navigli, R., McCrae, J. P. & Jakubíček, M. (2021). Intermediate interoperability report. ELEXIS Deliverable 6.3. Available at: https://elex.is/wp-content/uploads/2021/02/ELEXIS_D6_3_Intermediate_interoperability_report.pdf.

Krek, S., McCrae, J. P., Kosem, I., Wissik, T., Tiberius, C., Navigli, R., & Pedersen, B. (2018). European Lexicographic Infrastructure (ELEXIS). In J. Čibej et al.

(eds.) *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts, Ljubljana, Slovenia, 17-21 July 2018.* Available at http://doi.org/10.5281/zenodo.2599902.

Krek, S., Declerck, T., McCrae, J.P. & Wissik, T. (2019). *Towards a Global Lexicographic Infrastructure.* Presented at the Language Technology 4 All Conference. Available at http://doi.org/10.5281/zenodo.3607274

McCrae, J.P. (2020). Interoperable interface for Lemon and TEI resources. ELEXIS Deliverable 2.2. Available at: https://elex.is/wp-content/uploads/2020/02/ELEXIS_D2_2_Interoperable_Interface_for_Lemon_and_TEI_resources.pdf.

McCrae, J.P., Tiberius, C., Khan, A.F., Kernerman, I., Declerck, T., Krek, S., Monachini, M. & Ahmadi, S. (2019). The ELEXIS interface for interoperable resources. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 Conference, 1–3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o.* Available at: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_37.pdf.

Měchura, M. B. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference, 19-21 September 2017, Leiden, The Netherlands.* Available at: https://www.lexonomy.eu/docs/elex2017.pdf.

Parvizi, A., Kohl, M.,Gonzàlez, M. & Saurí, R. (2016). Towards a Linguistic Ontology with an Emphasis on Reasoning and Knowledge Reuse. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia. Available at: https://www.aclweb.org/anthology/L16-1071/.

Pedersen, B. S., McCrae, J. P., Tiberius, C. & Krek, S. (2018). ELEXIS - a European infrastructure fostering cooperation and information exchange among lexicographical research communities. In F. Bond, T. Kuribayashi, C. Fellbaum & P. Vossen (eds.) *Proceedings of the 9th Global WordNet Conference (GWC 2018), Global Wordnet Association, Singapore.* Available at: http://doi.org/10.5281/zenodo.2599954.

Repar, A. & Krek, S. (2020). Tools for the automatic segmentation and identification of lexicographic content. ELEXIS Deliverable 1.3. Available at: https://elex.is/wp-content/uploads/2020/02/ELEXIS_D1_3_Tools_for_the_automatic_segmentation_and_identification_of_lexicographic_content.pdf.

Romary, L. (2015). TEI and LMF crosswalks. *JLCL - Journal for Language Technology and Computational Linguistics*, 30 (1).

Svensén, Bo (2009). *A handbook of lexicography. The theory and practice of dictionary-making.* Cambridge: Cambridge University Press.

Tasovac, T, Romary, L., Banski, P., Bowers, J., de Does, J., Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Petrović, S., Salgado,

A. & Witt, A.. (2018). *TEI Lex-0: A baseline encoding for lexicographic data.* Version 0.8.6. DARIAH Working Group on Lexical Resources. Available at https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

Tavast, A., Langemets, M., Kallas, J. & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana, 17-21 July 2018.* Ljubljana University Press, Faculty of Arts, pp. 749—761. Available at: https://euralex.org/publications/unified-data-modelling-for-presenting-lexical-data-the-case-of-ekilex/.

Dictionary titles used in the survey:

The American Heritage Dictionary of the English Language. Fifth Edition. Accessed at: https://www.ahdictionary.com. (9 April 2021)

Collins English Dictionary (2000) Fifth Edition, HarperCollins Publishers, Glasgow, UK

Collins English Dictionary (2006) Eight Edition, HarperCollins Publishers, Glasgow, UK

Collins Dictionary English-French. HarperCollins Publishers. Accessed at: https://www.collinsdictionary.com (9 April 2021)

Dictionary.com. Accessed at: https://www.dictionary.com/. (9 April 2021)

DWDS *Digitales Wörterbuch der Deutschen Sprache.* Accessed at: https://www.dwds.de. (9 April 2021)

Lynyole dictionary. Accessed at: https://www.webonary.org. (9 April 2021)

MacMillan English Dictionary for Advanced Learners (2002) First Edition, MacMillan

Oxford Advanced Learner's Dictionary (1995) Fifth Edition, Oxford University Press, Oxford, UK

Oxford-Hachette French Dictionary (1994) First Edition, Oxford University Press, Oxford, UK

Petit Larousse Illustré 1905

The Right Rhymes Dictionary. Accessed at: https://therightrhymes.com. (9 April 2021)