# The lexicographic process revisited

Klosa-Kückelhaus, A.; Tiberius, C.P.A.

*International Journal of Lexicography*, 2024, **XX**, 1–12
https://doi.org/10.1093/ijl/ecae016
Advance access publication 1 November 2024
**Review Article**

OXFORD

# The Lexicographic Process Revisited

**Annette Klosa-Kückelhaus**[1] **and Carole Tiberius**[2]

[1]Leibniz-Institut für Deutsche Sprache, Germany (klosa@ids-mannheim.de)
[2]Instituut voor de Nederlandse Taal, The Netherlands (carole.tiberius@ivdnt.org)

## Abstract

Lexicographic practice is changing and so is the lexicographic process. This article therefore reconsiders the lexicographic process and takes into account the development towards centralised lexicographical databases which can be observed at lexicographic institutions and dictionary publishing houses in the last couple of years. We discuss how this affects the lexicographic process and its phases and how the process could be further changed by the systematic use of LLMs in lexicography or the further involvement of dictionary users in the process. The paper concludes by emphasizing the need for continuously observing the lexicographic process and redefining it if necessary.

**Keywords:** lexicographic process; digital dictionary; (centralized) lexical database; dictionary portal; computational lexicography; NLP

## 1. Introduction

The form and content of dictionaries have always been the focus of scientific dictionary criticism and the scientific study of dictionaries. Form and content are also easily accessible from the outside, because both are visible in the printed dictionary on paper or in the electronic dictionary on a screen. This is not the case for the creation process of a dictionary which takes place in the background. We call this the lexicographic process. By this we mean all the steps that are necessary for a dictionary to be published in print or electronically. Although the lexicographic process does not seem to be a particularly topical issue[1], it is important to keep adjusting our ideas about it, because the process is multi-layered (see the results of a study comparing 14 different projects in Europe in the context of the COST ENeL action with regard to their lexicographic process, cf. Tiberius and Krek, 2014) and constantly evolving.

A detailed understanding of the lexicographic processes for published dictionaries can also facilitate the planning of new projects: the better we understand the process, the better we will understand in which parts of the process new technologies can be of use and how they can be applied. In addition, the form and content of a dictionary can be better evaluated in light of what we know about its origins. Finally, a better understanding of the process can help make the lexicographic process more transparent to dictionary users and funding agencies.

In this paper, we revisit the question of the lexicographic process and present an extended definition of it against the background of new developments (cf. Section 2.).

Online publication has become the most popular publication medium for dictionaries in Europe. Furthermore, the era of stand-alone (paper) dictionaries seems to be over and more and more lexicographic institutions and dictionary publishers are moving towards producing a single resource containing all their linguistic/lexicographic data which feeds different (lexicographic and other) projects and applications (cf. the ELEXIS surveys on lexicographic practices in Europe (Kallas et al., 2019; Tiberius et al., 2022)).

We describe how this affects the lexicographic process and its phases for individual dictionaries, lexicographical databases, and dictionary portals (cf. Section 3.). The background to this development (e.g. technological innovations and the emergence of Linguistic Linked Data) is examined in Section 4 before Section 5 considers how the lexicographical process could be further changed by the systematic use of Large Language Models (LLMs) in lexicography or the further involvement of dictionary users in the creation of lexicographical information. Section 6 summarizes the results and finally argues in favour of observing the lexicographic process in the future and redefining it if necessary.

## 2. Re-defining the lexicographical process

The lexicographic process was first described exclusively for printed dictionaries (e.g. in Riedel and Wille, 1979; Landau, 1984; Schaeder, 1987; Dubois, 1990, and Zgusta, 1971), later extended to the use of computers in the compilation of dictionaries by Wiegand (1999) and to the production of electronic dictionaries by Müller-Spitzer (2004), and finally the process of compiling online dictionaries was also analysed (cf. Svensén, 2009, Klosa and Tiberius, 2016). Landau (1984, p. 227) identifies three phases that occur in any lexicographic process: "planning (30%), writing (50%), and producing (20%)", although the writing phase is usually much longer in academic dictionaries than in commercial dictionaries. It is to the credit of Herbert Ernst Wiegand (1999) that he was the first to consider and describe the lexicographic process in full theoretical terms. He defines the lexicographic process as follows:

> Ein abgeschlossener lexikographischer Prozeß [...] ist die Menge derjenigen prozeßzugehörigen Tätigkeiten, welche ausgeführt wurden, damit ein bestimmtes Wörterbuch entsteht. 'A completed lexicographic process [...] is the set of process activities carried out to create a particular dictionary.' (Wiegand, 1999, p. 134)

Wiegand distinguishes five phases in the production of dictionaries: the preparation phase, the material procurement phase, the material preparation phase, the material evaluation phase, and the typesetting and printing preparation phase. A distinction is also made between lexicographic processes with and without the use of computers. However, the goal of both processes is to produce a printed dictionary.

Müller-Spitzer (2004, p. 161) extends Wiegand's concept for lexicographic processes with the use of computers: These are either computer-aided processes for print dictionaries, computer-lexicographic processes for electronic dictionaries, or media-neutral lexicographic processes designed to enable the publication of printed dictionaries or the publication of lexicographic data on electronic data carriers (e.g. Heidecke, 1996, p. 401 gives an example for the 'system specific realisation' of SGML files of bilingual dictionaries). Both a computer-lexicographic process and a media-neutral lexicographic process aim at the production of electronic dictionaries for human users or lexical resources for language technology (e.g. Viegas and Nirenburg, 1996 give an example for the computer-lexicographic process established to create a Spanish lexicon for a machine translation system).

de Schryver and Prinsloo (2000) propose to collect feedback from the target users during the compilation of any dictionary by "launching several small-scale parallel dictionaries from which informal as well as formal target user feedback is instantly channelled back into the compilation process of a main dictionary" (2000, p. 197). In this scenario of simultaneous feedback, we must assume a nesting of different lexicographic processes: one for the completed dictionary, in which other lexicographic processes are embedded for the smaller dictionaries. Gouws (2016, p. 80) recommends implementing a reflexive phase in the main process that can build on the data collected by simultaneous feedback.

Klosa (2013, p. 520) describes the lexicographic process for online dictionaries (being published and updated continuously) and identifies six phases: preparation phase, data acquisition phase, computerization phase, data processing phase, data analysis phase, and preparation for online release phase. In contrast to a certain linearity of the individual phases for the production of print dictionaries, here, the phases do not follow each other seamlessly (e.g. online dictionaries may be published in stages: Thus, the publication of an online dictionary may begin before the writing phase is completed) and it is sometimes difficult to identify precise boundaries between the different phases (Klosa 2013, p. 522). Tiberius and Krek (2014, p. 1) expand Klosa's model: "For a completed dictionary, a 7th phase can be identified, i.e. the afterlife which entails maintenance and preservation of the dictionary after the end of the project. This is similar to what Svensén (2009, p. 413) calls 'follow-up'." (cf. Section 3. for more details).

Klosa and Tiberius (2016, pp. 102–104) have presented thoughts on whether there is a lexicographic process for dictionary portals and how it might be defined. In the case of a dictionary portal, a web page is created through which various dictionaries can be accessed by human users. Although in this type of process, key phases of computer-lexicographic processes are omitted (in particular, the writing of dictionary articles), several steps (e.g. designing an access structure) still require lexicographic expertise. In this sense they suggest extending Wiegand's (1999) definition: according to their understanding, a computer-lexicographic process is the sum of activities that have to be undertaken so that a particular dictionary or a group of dictionaries (in a portal) comes into being on or in an electronic carrier medium. This process can be divided into different phases in which computers are used consistently (Klosa and Tiberius, 2016, p. 104).

The lexicographic process changes yet again as more and more lexicographic institutions and dictionary publishers are moving towards creating a single resource that feeds different lexicographic projects and possibly other (not necessarily) lexicographic tools and applications (for more information on the background cf. Section 4.). An early example for such a concept is described in Alexa et al. (2002) and (2016): In the Duden Ontology, all data of printed Duden dictionaries is contained in a structured and systematic way, and data can be exported from it to be printed, used in dictionary software, on mobile devices or online and in language applications such as a spell checker. In such a constellation, there is not just one lexicographic process, but there are at least two processes (one for the centralised lexicographic database and one of each product that is compiled out of the data, for more details cf. Section 5.).

We, therefore, propose to extend Wiegand's definition of the lexicographic process to lexicographic databases (instead of dictionaries) to better reflect the digital aspects: a (completed) lexicographic process is the set of process activities performed to create a lexicographic database. Lexicographic databases can be used in language technology, for creating dictionaries for human users, or can be integrated in a dictionary portal. Extending the graph of lexicographic processes using computers given in Müller-Spitzer (2004, p. 161), Figure 1 provides an overview. It is obvious that the digital revolution has left its distinct mark on the lexicographic process (cf. Leroyer and Køhler Simonson, 2020, p.184).

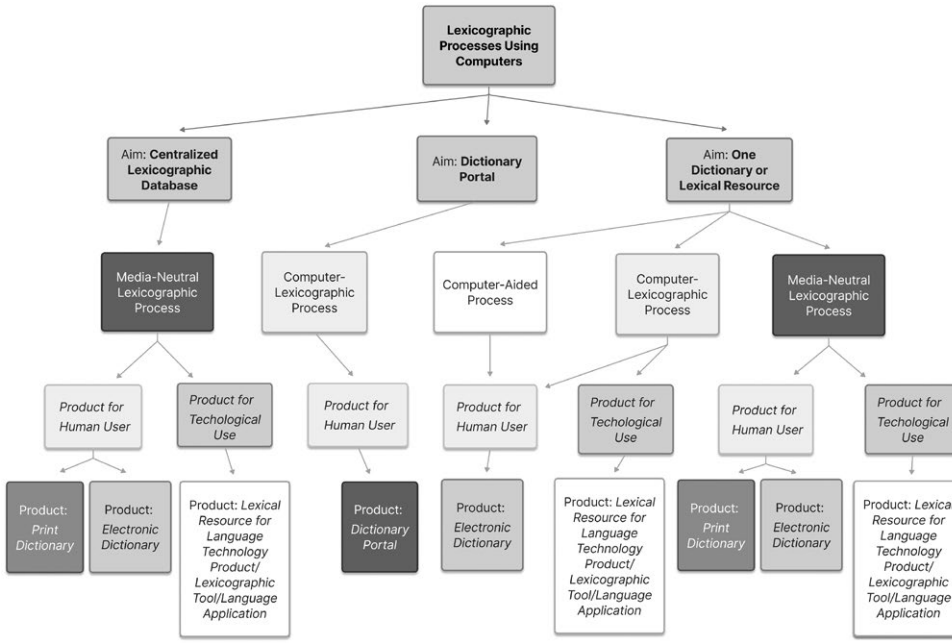


**Figure 1.** Overview of different lexicographic processes

## 3. Phases of the lexicographic process

We have already seen that the lexicographic process model needs to be extended. The same applies to the process phases. We certainly need to take into account that dictionary portals and electronic dictionaries, especially those published online, need to be maintained and preserved. It may also be necessary to better adapt the scheme to the reality of different dictionary projects. For this reason, an overview of the process of the different projects, as given in Tiberius and Krek (2014) or Hildenbrandt and Klosa (2016), is valuable. Finally, the survey on lexicographic practices in Europe conducted as part of the ELEXIS initiative (cf. Kallas et al., 2019, Tiberius et al., 2022) sheds light on potential ways of improving electronic dictionaries and their lexicographic processes. Based on this data, we distinguish the following phases of the lexicographic process (of dictionaries, dictionary portals, and centralised lexicographic databases):

1. phase of preparation
2. phase of data acquisition
3. phase of computerization
4. phase of data processing
5. phase of data analysis
6. phase of preparation for publication
7. phase of maintenance and preservation

In phase 1, the following steps can be taken: Designing the dictionary or dictionary portal or centralised lexicographic database; preparing the organizational plan; conducting a pilot study or studies, etc. Phase 2 involves providing the sources (corpora, data from other dictionaries, etc.) and obtaining data such as videos, illustrations, and sound files. While these data are needed for the creation of a single dictionary as well as a centralised lexicographic

database, the data acquisition phase for a dictionary portal includes the acquisition of the digitized dictionaries to be included in the portal. In the computerization phase (phase 3), corpora are annotated, software such as corpus query tools and dictionary writing systems are programmed, and the lexicographic database is created. Print dictionaries can be retro-digitized in the computerization phase (e.g. for inclusion in dictionary portals). The data processing phase (phase 4) is used to annotate audio or video files, and to process corpus data as well as data from other sources for tasks such as creating a candidate lemma list, determining frequencies, statistically computing co-occurrences or automatically compiling lexicographic information. These steps can also be taken if an already existing dictionary is to be improved (e.g. by adding information on frequencies). In dictionary portal projects, this phase includes tasks such as creating a combined lemma list for all dictionaries in the portal or indexing data to provide networked access structures. Phase 5 involves the writing of dictionary entries (depending on the result of phase 4). It can also involve post-editing lexicography (e.g. manually verifying (semi-)automatically compiled lexicographic information, such as example sentences, collocations, hyphenation). In the case of dictionary portals, this phase is usually skipped, since here lexicographic information is already available in the dictionaries of the portal. The data analysis phase usually takes the longest in dictionary projects (cf. Tiberius and Krek, 2014, p. 16), but building a centralised lexicographic database is also time consuming and may include a lot of manual work. In the preparation for publication (phase 6), testing, proofreading, and writing of texts such as user manuals take place in the dictionary and dictionary portal projects. In projects with a centralised lexicographic database, all data to be published in an individual end-product must be collected automatically, so programming such procedures and testing them is also necessary for this phase. Phase 7 is relevant to all projects, as individual electronic dictionaries, dictionary portals, as well as a centralised lexicographic database all need to be maintained. Arrangements must also be made for archiving different versions of the data and for version control.

Although this list of phases in the lexicographic process looks linear, it is not: The phases described above do not necessarily always run consecutively (see Klosa, 2013, p. 522, and Lemberg, 2001, p. 86; for examples of the overlapping of phases and the circularity of phases in different projects, see the contributions in Tiberius and Krek, 2014; Hildenbrandt and Klosa, 2016). In concrete dictionary projects, moreover, it is sometimes difficult to determine the exact duration of individual phases, since the work in a particular phase is not necessarily full-time, but can extend over longer periods (cf. Tiberius and Krek, 2014, p. 16).

As the lexicographic process changes, the expertise needed to create lexicographic databases and publish dictionaries or dictionary portals grows, and many lexicographic institutions take on new roles. They become more of a data provider and less of a dictionary publisher and their lexicographers are commonly also involved in tasks which are not strictly speaking lexicographic, such as project management, data management, fundraising, teaching, and public relations (Tiberius et al. 2022, p. 519).

In addition to lexicographers, computer specialists (corpus linguists, computational linguists, and computer scientists) are being employed in lexicographic projects, and other expertise such as Internet design and website usability is also in demand. Therefore, certain tasks in a lexicographic process are often outsourced, such as developing an online application and user interface (Tiberius et al., 2022, p. 513, see also Kallas et al., 2019, p. 17), or retro-digitization tasks such as scanning, typing, and converting, developing a dictionary writing system or corpus retrieval system, setting up a database, or creating a mobile app (Tiberius et al., 2022, p. 514).

Users can be involved in many steps of the lexicographic process. Patrick Leroyer's scheme reflecting the involvement of users and experts in the lexicographic process of specialised dictionaries (cited in Tiberius and Krek, 2014, p. 15) illustrates that users can contribute to the data acquisition phase, should be involved in the testing phase (preparation for publication), and that user involvement is "definitely required in the post-publication phase". But users can also be involved as early as the planning stage when a pilot study asking

for their needs is conducted. Crowdsourcing and gamification are ways to involve users further on in the lexicographic process; both may increase motivation to participate in compiling dictionary content. Users can help in the phase of data processing, for example by taking over annotation tasks. They may even support lexicographers during the data analysis phase: Kallas et al. (2019, p. 47) show that European dictionary projects seek user input on "synonyms, word associations, neologisms [...] and the transcription of a particular dialect". Information on paradigmatic partner words, for example, will certainly be used by lexicographers when writing the entries. However, the ELEXIS surveys (Tiberius et al., 2024) indicate that crowdsourcing and gamification are not yet common practice in the dictionary compilation process. Only 11 institutions out of the 65 institutions participating in the surveys reported involvement in crowdsourcing. Even fewer institutions (i.e. only 3) are or have been involved in gamification. A possible explanation for the low figures for crowdsourcing and gamification may be that institutions are still searching for the best ways of including crowdsourcing methodologies in the lexicographic workflow or that there is not enough time and/or money to implement them.

As more and more experts, as well as users, become involved in an increasingly complex lexicographic process, the more important organization and control of this process become. Such complex lexicographic processes, whether for dictionaries, dictionary portals, or centralised lexicographic databases, must definitely be calculable, decomposable, controllable, adaptable, learnable, and testable (cf. Wiegand, 1999, p. 134). We also suggest that user studies should be used in the first phases of the lexicographic processes of dictionaries and dictionary portals so that they can fulfil their actual purpose and are oriented toward the needs of their users. But it is also true for later stages of the process as de Schryver and Prinsloo (2000) have shown. Applying this system consistently, leads to a kind of lexicographic democracy, where users give feedback with reference to new data types, new structure or the presentation of the data (cf. Gouws, 2016, p. 80). For centralised lexicographic databases, however, user studies are hard to implement as they are part of a larger infrastructure with many potential (and also yet unknown) end users.

## 4. The changing field of lexicography

Before the arrival of the digital age, lexicography was not really susceptible to standardization efforts and collaboration between lexicographic projects was limited. However, major changes such as the move from print to online, the shift from unstructured to structured data, technological innovations, and the emergence of Linguistic Linked (Open) Data, have led to an increased awareness of the need for more stable and established encoding formats for lexicographic data. As a result, different encoding formats have emerged, TEI Lex-0, Ontolex-Lemon, and the ISO LMF (Lexical Markup Framework) standard as well as ISO 1951:2007, and most recently DMLex.

Rooted in the Text Encoding Initiative (TEI) ecosystem, TEI Lex-0[2] has been developed, with the aim of establishing a baseline encoding and a target format to facilitate the interoperability of heterogeneously encoded lexical resources (Tasovac et al., 2018). The focus of TEI Lex-0 is on retro-digitized dictionaries. Within the Linked Data community, Ontolex-Lemon[3] (Cimiano et al., 2016) has been defined and forms the de facto standard for representing lexical information as RDF (Resource Description Framework). It includes a special module for lexicography, Ontolex-Lexicog[4]. The ISO standard LMF[5], a meta-model for representing data in monolingual and multilingual lexical databases used with computer applications, has been technically revised in 2019 and ISO 1951:2007[6], the standard dedicated to "Presentation/representation of entries in dictionaries – Requirements, recommendations and information" is currently being revised. As part of the ELEXIS project, DMLex[7] (Data Model for Lexicography), has been developed within OASIS, an open source standard targeted specifically at born-digital dictionaries.

Set against this background, a parallel trend can be observed. More and more lexicographic institutions and dictionary publishers are moving away from stand-alone resources (in different encodings) towards centralised lexicographic knowledge bases from which different end-products can be derived. See for example, Duden dictionaries (Alexa et al., 2002, 2011, and 2016), KDictionaries in the context of their global multilingual dictionary series (cf. Kernerman, 2011) and OUP's Oxford Global Languages initiative (cf. Parvizi et al., 2016; Pearsall, 2018). At the same time, several lexicographic institutions throughout Europe have also started creating such centralised databases for their lexicographic data (cf. Tavast et al., 2018; Depuydt et al., 2019; Gantar, 2020[8]).

Having a single pool of data has several practical advantages. It supports reusability of the lexicographic data for different end products, dictionaries as well as language technology applications. It avoids having multiple dictionaries with duplicated and conflicting information and helps to minimize redundancies. It also ensures consistency which again leads to more efficient maintenance of the lexicographic data in a homogenous manner. These advantages are important for lexicographic institutions and publishing houses considering shorter project running times as well as the changing role of lexicographic institutions and publishing houses (cf. Tiberius et al., 2024) which are becoming more of a data provider and less of a dictionary publisher.

On the other side of the coin, having a single pool of data also presents new challenges. Integrating all lexicographic data into one centralised database naturally results in a more complex data model.[9] Moreover, this data model should be flexible and extensible to be able to deal with (future) technological innovations (e.g. LLMs). The situation is further complicated if heterogeneous content from different existing dictionaries (each with their own data model and their own peculiarities) needs to be integrated into the centralised database. Precisely because of this problem, we see that different institutes follow different routes to achieve the goal of a centralised lexicographic knowledge base. The Dutch Language Institute, for instance, started from a minimal core of shared data which is gradually growing overtime with projects feeding more and more data into the database (conforming the data model which is being refined as the database grows) (Depuydt et al., 2019). The Estonian Language Institute opted for the opposite approach and merged all data (both semasiological and onomasiological) at the institute into the centralised database right at the start and unification of the data is the long-term goal. Issues with the data – ambiguities, under-specification, inconsistencies and conflicts – have been and are being resolved on the way (e.g. Tavast et al., 2018).

Having a single pool of data also requires special editing tools which allow different views on the data and support a more relational encoding of lexicographic data (cf. Meyer and Eppinger, 2018; Měchura et al., 2023) to enable editing per 'feature' (e.g. edit all parts of speech for all headwords). Current Dictionary Writing Systems (DWS) are generally geared towards hierarchically structured editing per headword and do not support this. Organizing lexicographic work by task instead of by headword is also not something lexicographers are used to (cf. e.g. Tavast et al., 2018) and requires a turn-around in their way of working. Finally, it remains to be seen how much work is needed to customize the content from the centralised lexicographic knowledge base for specific end-products.

## 5. The changing lexicographic process

In the constellation exemplified in the previous section, there seems to be not just one lexicographic process, but at least two: one for the centralised lexicographic database and one for each end-product that is compiled out of the data. The compilation of a central database, which is by default media-neutral, requires a media-neutral lexicographic process. As shown in Figure 1, the data in the database can be used to produce reference works intended for human users (print dictionaries or electronic dictionaries for human users), or

**Table 1.** Phases in the lexicographic process of a centralised lexicographic database and its end-products

| Phase | centralised database | print dictionary from database | electronic dictionary from database | data for NLP from database |
|---|---|---|---|---|
| preparation | X | X | X | X |
| data acquisition | X | | | |
| computerization | X | | | |
| data processing | X | | | |
| data analysis | X | | | |
| preparation for publication | | X | X | X |
| maintenance and preservation | X | X | X | X |

lexical resources for language technology products, lexicographic tools, or other language applications. In this respect, this process is similar to that for individual resources, insofar as these arise from a media-neutral process. Table 1 shows that all the phases that were discussed in Section 3 are still relevant in the context of the centralised database except for the phase of preparation for publication as publication is expected to occur in the process for individual end-products. For the creation of the individual end products a subset of the phases is needed. In particular, for each individual end-product there will be a phase of preparation (to develop a concept for a specific end-product), a phase of preparation for publication and a phase of maintenance and preservation.

Furthermore, for the centralised database, the organization of the work in the different phases changes: generic content creation requires more task-based editing instead of editing per headword (e.g. editing morphological information, pronunciation, synonyms, and translations for a set of headwords).

Technological innovations (e.g. LLMs) will undoubtedly result in more changes in the lexicographic process. Since the launch of ChatGPT at the end of 2022, there has been an increased interest in the use of LLMs in lexicography (see de Schryver (2023, p. 5) for an overview of papers on ChatGPT and lexicography). Although these papers reveal that ChatGPT can be used for a number of lexicographic tasks and is actually quite good at some of them (especially for English), all papers also mention a number of concerns (e.g. memorization, authorship, propagation of negative generalisations, the non-deterministic nature of LLMs, hallucinations). One of the main concerns, though, is that there is no direct link between the generated output and the data that these models have been trained on. With corpora lexicographers can go back to the underlying data, whereas this is not the case with the output of LLMs. For many lexicographers (e.g. Rundell, 2023) this is currently the main deal-breaker.

To counteract some of the concerns, the creation of lexicography-specific LLMs is being advocated (e.g. Jakubíček and Rundell, 2023; McKean and Fitzgerald, 2023; de Schryver 2023). As McKean and Fitzgerald (2023, p. 25) note this could lead to the creation of lexicographical service providers for LLMs so that institutions and projects can then take advantage of these models in their specific lexicographic processes (similar to the way corpora and corpus tools are used now). McKean and Fitzgerald (2023, p. 26) also mention the importance of creating evaluation task sets for dictionaries that can be used to tune or train models to do lexicographic work. Lexicographic data could thus provide ground truth data for benchmarking.

It is difficult to predict the future, but a number of potential roles for LLMs can already be identified in the lexicographic process. First, LLMs can potentially be used for automatic content creation (i.e. in the phase of data processing). Second, they can be used as a secondary source that lexicographers consult whilst editing dictionary entries, preferably integrated in the DWS (i.e. in the phase of data analysis[10]). TLex (aka TshwaneLex),

for instance, already has an OpenAI section that lexicographers can use to prompt any of the OpenAI functionalities to generate dictionary content (de Schryver 2023, p. 2). Finally, with more and more lexicographic data that needs to be processed, analysed and (post-)edited, it would be interesting to explore further whether LLMs can be used to replace or complement human judgment, similar to the use of crowdsourcing and gamification in lexicography (in the phase of data acquisition and data analysis). Currently crowdsourcing and gamification are not yet common practice in the dictionary compilation process, but, as noted in Tiberius et al. (2024), relying at least in part on the wisdom of the crowd seems unavoidable. Relying at least partially on LLMs may well be equally unavoidable.

## 6. Conclusion

In this paper we have revisited the lexicographic process. We have extended Wiegand's definition of the lexicographic process to lexicographic databases (instead of dictionaries) to better reflect the digital era. The lexicographic process is the set of process activities performed to create a lexicographic database. Lexicographic databases can be used in language technology or for creating dictionaries for human users or can be integrated in a dictionary portal.

The paper shows that the process is multi-layered and constantly evolving. We noted the shift from one database per project to having one database for multiple projects. In particular, lexicographic institutions and dictionary publishers have started creating centralised databases containing all their linguistic/lexicographic data which then feed different (lexicographic and other) projects and applications. In this new constellation, there are at least two lexicographic processes: one for the centralised lexicographic database and one for each end-product that is compiled out of the data. These processes are, however, not the same. The phase of preparation for publication does not apply to the process for the centralised database as it is not very likely that the whole database will be published as such. The phase of publication will be relevant in the context of the individual end-products that can be derived from the database. Furthermore, the phase of data analysis will change. Traditionally, lexicographers organize their work by headword and they may continue to do so in single dictionary projects. However, generic content creation for the centralised database requires a more task-based approach. Lexicographers will have to adapt to a new way of working.

We also noted the changing role of lexicography. Lexicographic institutions and dictionary publishers do not just publish dictionaries anymore, but have become more of a content provider. Lexicographic data can provide useful content for various language applications and lexicographic datasets can be used for research and development, for instance, with LLMs.

Future innovations will further automate the process and relieve lexicographers from intellectually undemanding tasks, leading to new changes in the process. To best understand in which parts of the process new technologies can be of use and how they can be applied, it is important to keep monitoring the lexicographic process.

## Notes

1 Searching for 'lexicographic(al) process' in the title of all 6.482 articles listed in ELEXIFINDER (see https://elex.is/tools-and-services/elexifinder/) only gives two hits. There are also only two hits for 'compilation process', three hits for 'dictionary making process', and one hit for 'computational lexicon making process' in article titles as of 02 February 2024).
2 See https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.
3 See https://www.w3.org/2016/05/ontolex/.
4 See https://www.w3.org/2019/09/lexicog/.
5 See https://www.iso.org/standard/68516.html.

6 See https://www.iso.org/standard/36609.html.
7 See: https://github.com/oasis-tcs/lexidma/releases.
8 See https://hrcak.srce.hr/file/356570.
9 See, for instance, the data model of the Slovene Digital Dictionary Database: https://wiki.cjvt.si/books/digital-dictionary-database/page/data-model.
10 It should be noted though that there is an overlap with the phase of data processing as the LLM will suggest content, for example a definition, an example or a translation.

## References

Alexa, M. (2011). Modellierung eines semantischen Wissensnetzes für lexikographische Anwendungen am Beispiel der Duden-Ontologie. In A. Klosa & C. Müller-Spitzer (Eds.), *Datenmodellierung für Internetwörterbücher. 1. Arbeitsbericht des wissenschaftlichen Netzwerks "Internetlexikografie"* (pp. 61-70). OPAL – Online publizierte Arbeiten zur Linguistik 2/2011. Mannheim.

Alexa, M., Konopka, A., & Wind, L. (2016). Gibt es einen medienspezifischen lexikografischen Prozess für das Online-Wörterbuch? – Ein Werkstattbericht. In V. Hildenbrandt & A. Klosa (Eds.): *Lexikographische Prozesse bei Internetwörterbüchern* (pp. 8–19). OPAL - Online publizierte Arbeiten zur Linguistik 1/2016. Mannheim: Institut für Deutsche Sprache.

Alexa, M., Kreissig, B., Liepert, M., Reichenberger, K., Rostek, L., Rautmann, K., Scholze-Stubenrecht, W., & Stoye, S. (2002). *The Duden Ontology: An Integrated Representation of Lexical and Ontological Information*. In: LREC Workshop on WordNet Structures and Standardisation, and How These Affect Wordnet Applications and Evaluation 2002. Las Palmas, Gran Canaria. Accessed on 7 June 2024. https://bultreebank.org/wp-content/uploads/2017/05/OntoLex02Paper01.pdf.

Cimiano, P., McCrae, J. P., & Buitelaar, P. (2016). *Lexicon Model for Ontologies: Final Community Report, 10 May 2016*. Accessed on 16 January 2024. https://www.w3.org/2016/05/ontolex/.

Depuydt, K., Schoonheim, T., & de Does, J. (2019). *Towards a More Efficient Workflow for the Lexical Description of the Dutch Language*. Accessed on 16 January 2024. http://videolectures.net/elexisconference2019_depuydt_dutch_language/.

de Schryver, G.-M., & Prinsloo, D. (2000). Dictionary-making process with "Simultaneous feedback" from the target users to the compilers. In U. Heid, S. Evert, E. Lehman & Chr. Rohrer (Eds.), *Proceedings of the Ninth EURALEX International Congress* (pp. 197–209). Stuttgart: Institut für Maschinelle Sprachverarbeitung.

de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography* 36(4), 355-387. https://doi.org/10.1093/ijl/ecad021.

Dubois, C. (1990). Considérations générales sur l'organisation du travail lexicographique. In F.-J. Hausmann, O. Reichmann, H. E. Wiegand & L. Zgusta (Eds.), *Dictionaries. An international handbook on lexicography* (pp. 1645-1672). Berlin/New York: de Gruyter.

Gantar, P. (2020). Dictionary of Modern Slovene: From Slovene Lexical Database to Digital Dictionary Database. *Rasprave Instituta Za Hrvatski Jezik i Jezikoslovlje*, 46(2), 589–602. https://doi.org/10.31724/rihjj.46.2.7.

Gouws, R. H. (2016). Aspekte des lexikographischen Prozesses in Print- und Onlinewörterbüchern. In V. Hildenbrandt & A. Klosa (Eds.), *Lexikographische Prozesse bei Internetwörterbüchern* (pp. 70-81). OPAL – Online publizierte Arbeiten zur Linguistik 1/2016. Mannheim: Institut für Deutsche Sprache.

Heidecke, St. (1996). SGML-Tools in the Dictionary-making Process – Experiences with a German-Polish/Polish-German Dictionary. In M. Gellerstam, J. Järborg, S.-G. Malmgren, I. Norén, L. Rogström, & C. Röjder Papmehl, C. (Eds.), *Proceedings of the 7th EURALEX International Congress, Part I* (pp. 395-403). Novum Grafiska AB. Göteborg.

Hildenbrandt, V., & Klosa, A. (2016). *Lexikographische Prozesse bei Internetwörterbüchern*. OPAL – Online publizierte Arbeiten zur Linguistik 1/2016. Mannheim: Institut für Deutsche Sprache. http://dx.doi.org/10.14618/opal_01-2016.

Jakubíček, M., & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, & M. Jakubíček (Eds.), *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century* (pp. 508-523). Brno: Lexical Computing.

Kernerman, I. (2011). From Dictionary to Database: Creating a Global Multi-Language Series. In I. Kosem & K. Kosem (Eds.), *Proceedings of the eLex 2011 Conference: Electronic Lexicography in the*

*21st Century* (pp. 113-121). Bled Slovenia. Accessed on 16 January 2024. https://elex2011.trojina.si/Vsebine/proceedings/eLex2011-14.pdf.

Kallas, J., Koeva, S., Kosem, I., Langemets, M., & Tiberius, C. (2019). *ELEXIS deliverable 1.1 Lexicographic Practices in Europe: A Survey of User Needs*. Accessed on 16 January 2024. https://elex.is/wp-content/uploads/2020/06/Revised-ELEXIS_D1.1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf.

Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (Eds.), *Dictionaries. An international Encyclopedia of Lexicography. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography* (pp. 517-524). Berlin, Boston: de Gruyter.

Klosa, A. & Tiberius, C. (2016). Der lexikografische Prozess. In: A. Klosa & C. Müller-Spitzer (Eds.), *Kompendium Internetlexikografie* (pp. 65–110). Berlin/Boston: de Gruyter.

Landau, S. (1984). *Dictionaries. The Art and Craft of Lexicography*. New York: Charles Scribner's Sons.

Lemberg, I. (2001). Aspekte der Online-Lexikographie für wissenschaftliche Wörterbücher. In I. Lemberg, B. Schröder & A. Storrer (Eds.), *Chancen und Perspektiven computergestützter Lexikographie* (pp. 71-91). Tübingen: Narr.

Leroyer, P., & Køhler Simonson, H. (2020). Reconceptualizing lexicography: the broad understanding. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (Eds.), *Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress, Vol. I* (pp. 183-192). Alexandroupolis: Democritus University of Thrace.

McKean, E., & Fitzgerald, W. (2023). The ROI of AI in Lexicography. In *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users"* (pp. 18-27). Seoul: Yonsei University.

Měchura *et al.* (2023). *Relations, relations everywhere: an introduction to the DMLex data model*. Video presentation at eLex conference 2023, Brno. Accessed on 8 June 2024. https://www.youtube.com/watch?v=b6Lkv-3D5C0.

Meyer, P. & Eppinger, M. (2018). fLexiCoGraph: Creating and Managing Curated Graph-Based Lexicographical Data. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts* (pp. 1017-1022). Ljubljana University Press, Faculty of Arts.

Müller-Spitzer, C. (2004). Ordnende Betrachtungen zu elektronischen Wörterbüchern und lexikographischen Prozessen. *Lexicographica*, 19(2004), 140-168. https://doi.org/10.1515/9783484604568.140.

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)* (pp. 441–448), Portorož, Slovenia. European Language Resources Association (ELRA).

Pearsall, J. (2018) One Model, Many Languages? An approach to multilingual content. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Book of Abstracts of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts* (p. 209). Ljubljana University Press, Faculty of Arts.

Riedel, H., & Wille, M. (1979). *Über die Erarbeitung von Lexika*. Leipzig: Bibliographisches Institut.

Rundell, M. (2023). Automating the creation of dictionaries: are we nearly there? In *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users"* (pp. 9-17). Seoul: Yonsei University.

Schaeder, B. (1987). *Germanistische Lexikographie*. Berlin/Boston: de Gruyter Mouton. (Lexicographica. Series Maior 34).

Svensén, B. (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*. Accessed on 7 June 2024. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

Tavast, A., Langemets, M., Kallas, J. and 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* (pp. 749-761). Ljubljana University Press, Faculty of Arts.

Tiberius, C., & Krek, S. (2014). *Workflow of Corpus-Based Lexicography*. *Deliverable COST-ENeL-WG3 meeting*. Accessed on 7 June 2024. https://www.elexicography.eu/wp-content/uploads/2015/04/LexicographicalWorkflow_DeliverableWG3BolzanoMeeting2014.pdf.

Tiberius, C., Kallas, J., Koeva, S., Langemets, M., & Kosem, I. (2022). An insight into lexicographic practices in Europe. Results of the extended ELEXIS Survey on User Needs. In A. Klosa-Kückelhaus, St. Engelberg, Chr. Möhrs & P. Storjohann (Eds.), *Dictionaries and Society. Proceedings of the XX EURALEX International Congress* (pp. 509-521). Mannheim: IDS-Verlag.

Tiberius, C., Kallas, J., Koeva, S., Langemets, M., & Kosem, I. (2024). A Lexicographic Practice Map of Europe. *International Journal of Lexicography*, 37(1), 1–28. https://doi.org/10.1093/ijl/ecad023.

Viegas, E. & Nirenburg, S. (1996). The Ecology of Lexical Acquisition: Computational Lexicon Making Process. In M. Gellerstam, J. Järborg, S.-G. Malmgren, I. Norén, L. Rogström, & C. Röjder Papmehl, C. (Eds.), *Proceedings of the 7th EURALEX International Congress, Part II* (pp. 489-496). Novum Grafiska AB. Göteborg.

Wiegand, H. E. (1999). *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. 1. Partial volume. Berlin/New York: de Gruyter.

Zgusta, L. (1971). *Manual of Lexicography*. The Hague/Paris: Mouton.