The Future of Academic Lexicography -- A White Paper

Frieda Steurs, Tanneke Schoonheim, Kris Heylen, Vincent Vandeghinste (eds.)

Version 1.2
(Last Updated: 24.11.2020)

Background: The NIAS-Lorentz workshop (Leiden, 4-8 November 2019)

The basis for this white paper are the presentations and discussions during the NIAS-Lorentz-workshop "The Future of Academic Lexicography" that was held at the Lorentz Center in Leiden, 4-8 November 2019.[1] NIAS is the Netherlands Institute for Advanced Study in the Humanities and Social Sciences. The Lorentz Center is a workshop center that hosts international scientific meetings of typically one week.

Starting point of the workshop was the requirement of *academic* lexicography as **evidence-based lexicography**. Academic lexicographic institutes have a long tradition of analysing large amounts of language data in a scientific way, in order to compile concise, yet high-quality knowledge about words, and this for the benefit of the entire language community. Such an approach is, of course, not exclusive to academic institutes, and can also apply to scientific or scholarly commercial lexicography, for which many of the things discussed at the workshop and in this paper can be applicable equally well.

Lexicography faces **challenges** with respect to

(a) its **role** in society, science and the knowledge economy;
(b) the **scalability** of both the analysis and production process; and
(c) the **ability to customise and make the content accessible** for a diverse audience, including as a resource for information technology developers.

The workshop explored how each of these challenges can be fruitfully addressed through **interdisciplinary research and development** in collaboration with neighbouring disciplines such as computational linguistics, data analytics, artificial intelligence, citizen science, human-computer interaction, cognitive science and sociology.

The workshop aimed to devise **strategies to strengthen the position** of academic lexicography as a locus for multidisciplinary scientific research with a direct relevance for, and impact on, society. These strategies are summarised in this White Paper which aims to serve as an inspiration and guiding document for setting up new collaborative research platforms and projects in the longer term.

The White Paper: Aims and Timing

Although the white paper is inspired by the presentations and discussions during the workshop and the feedback from participants afterwards, it is not intended as a report. Rather, it is a **strategic document** that:

(a) spells out the **specific issues** that confront academic lexicography with respect to its societal and economic positioning, its scalability, and content customisation;
(b) outlines and compares **potential solutions** and their feasibility and impact on lexicographic institutes, both in the short, middle and longer term;
(c) charts the types of **expertise** from different disciplines that are **necessary** to implement the proposed solutions;
(d) proposes **formats for collaboration** between experts from the different scientific disciplines.

[1] https://www.lorentzcenter.nl/the-future-of-academic-lexicography-linguistic-knowledge-codification-in-the-era-of-big-data-and-ai.html

The paper aims to provide information and inspiration relevant for **setting out the future course for lexicographic institutes** in general, and more specifically for the *Instituut voor de Nederlandse Taal* (Dutch Language Institute -- INT).[2]

The writing of this white paper proceeds in **different phases**. A first version (version 1.0) of the white paper stated the scope of issues to be dealt with and a first suggestion of the type of solutions to be considered. This version was circulated among the original workshop organisers and the participants for review and additions. In the current phase, a revised version (version 1.2) will be circulated among the scientific, societal and economic stakeholders of lexicography at large for input and suggestions. A final version, incorporating feedback from all stakeholders is expected to be published by early 2021.

With the white paper as a strategic planning document, the organisers of the original Lorentz workshop, together with other participants and stakeholders will then explore the **formation of** one or more **project consortia** in 2021.

State-of-the-Art and Challenges

Academic Lexicography: Historic Background and State-of-the-Art.

Scientific or scholarly dictionaries belong to the great scientific endeavours of the 19th century that strived for the **systematic extension and improvement of human knowledge**. Projects like the *Deutsches Wörterbuch*,[3] the *Oxford English Dictionary*,[4] the *Woordenboek der Nederlandsche Taal*[5] or the *Svenska Akademiens Ordbok*[6] wanted to bring "the Scientific Method" to the study of words and their meaning.

These grand projects set out to provide a **systematic, complete and definite description** of a language's **entire vocabulary**, past and present, based on the empirical observation of authentic language use.

At the same time, these scholarly dictionaries were intended as **public monuments** testifying to the shared history and common purpose of a people and a nation and the greatness of their cultural heritage. Although these high-minded goals have gradually become more modest and practical, lexicography as a scientific discipline has endured and the idea of the systematic, data-driven and principled documentation of linguistic knowledge inspired further dictionary projects in the 20th century like *Den Danske Ordbok*[7] or the *Slovar slovenskega knjižnega jezika*.[8]

Its focus on data processing and knowledge creation made academic lexicography a natural **early adopter** of information and communication technology. Scientific dictionaries re-invented themselves, not only in terms of how they are compiled, published and distributed, but also with respect to the intended audiences and the way users interact with them.

[2] http://www.ivdnt.org
[3] http://woerterbuchnetz.de/
[4] http://www.oed.com/
[5] http://gtb.ivdnt.org/
[6] https://www.saob.se/
[7] https://ordnet.dk/ddo
[8] http://bos.zrc-sazu.si/sskj.html

Lexicography actively contributed to the development of computational corpus linguistics. In the new *e-lexicography* paradigm of the 1990s and early 2000s, computational ***corpus query systems*** enabled the analysis of word use in the ever growing body of digital and digitised texts.

Dictionary writing systems did away with paper slips and filing cabinets and introduced many of the functionalities of relational databases and the efficiency of digital publishing. But, this new approach introduced new challenges, such as the spread of information over many different database fields, leading to new inefficiencies, if there is no clear definition, description or protocol for which information belongs in which field.

Online publishing not only introduced many new ways to query dictionaries as alternatives to traditional alphabetical look-up, it also **obviated the size limitations** of print and allowed the inclusion of **multimedia content**, now requiring an information curation task which is no longer guided by size limitations but by avoiding information overload.[9]

Traditional scholarly dictionaries had often been "by academics for academics", only playing a role for a popular audience as an indirect traditional influence, as information provider and reference point for commercial dictionaries. But the digital content of these scholarly dictionaries appeal to the popular audience through a more direct, unmediated form, through their **integration into applications** like writing aids, language learning software or computer assisted translation tools, which in turn have reached completely new audiences. The inclusion of dictionary entries in the results returned by internet search engines implied, in a way, the ultimate accessibility of scientific dictionary content for all.

Over the last few decades, many countries have become more linguistically diverse. Many different linguistic communities exist within single countries and one might wonder whether the role of national language institutes should be broadened towards the linguistic heritage of the different linguistic communities living in the country, and to the **multilingual interactions** between the different communities within a country and internationally.

Additionally, the traditional national and language-specific orientation of scientific lexicography has been overcome by initiatives like the **European Network on e-Lexicography**.[10] Scholarly dictionaries from across Europe are currently being interconnected as Linked Open Data[11] in the **ELEXIS**[12] project so that they become available as an open data resource for researchers and developers. The **Prêt-à-LLOD** project,[13] which is closely aligned with ELEXIS, exploits the combination of linked data and language technologies, that is Linguistic Linked Open Data (LLOD), to create ready-to-use multilingual data. The **European network for Web-centred linguistic data science**[14] (Nexus Linguarum) aims at building and strengthening a European network for creators and users of linguistic data.

---

[9] R. Gouws and, S. Tarp (2017). Information overload and data overload in lexicography. *International Journal of Lexicography*, Vol. 30(4), pp. 389–415.

[10] http://www.elexicography.eu/

[11] https://en.wikipedia.org/wiki/Linked_data#Linked_open_data

[12] ELEXIS is a European Horizon 2020 research infrastructure which aims to harmonize dictionary development. https://elex.is/

[13] https://pret-a-llod.github.io/

[14] CA18209 https://nexuslinguarum.eu

Challenges and a Vision for the Future

Dictionaries as end-products are increasingly less relevant today whereas the **raison-d'etre** and any prospects for new lexicography lie in its content, which assumes new shapes on its own and in interoperability and multidisciplinarity with numerous other domains, primarily for data sciences.

True to its beginnings, academic lexicography's core business is still threefold:

(a) a scientifically underpinned **analysis of language use** for the purpose of
(b) the **systematic documentation of word meaning and use** so that
(c) this *word knowledge* is **useful and accessible for a wide audience**.

In principle, a digital implementation of its enduring mission should make academic lexicography **well placed to take up an important role** in the interconnected knowledge infrastructure that enables our increasingly knowledge-based and networked economies and societies. After all, most of humanity's knowledge is still expressed in words, making *knowledge about words* crucial for any type of knowledge-based activity, be it by humans or machines.

Yet, if academic lexicography is to live up to this potential, its **current state of digitisation may not be aligned well enough** with the increasingly interconnected and network-based nature of information and communication. This plays out in all of lexicography's core activities: Data processing, content creation and generation, and dissemination. Looking at these in reverse order can clarify these interconnectivity/interoperability challenges for lexicography.

Dissemination: Accessing lexicographic information in a networked world

Even in their digital form, most dictionaries are still conceived primarily as *stand-alone* reference works that **human users** intentionally access to look up information on specific words.

Although this type of usage is not likely to disappear soon, its relative importance is declining. Human access to codified word knowledge increasingly occurs **in the context of online search engines or language-centred applications** like dedicated software for authoring, reading, language learning etc. To the extent that computers access structured lexical data for other applications, like information retrieval, text mining etc., they do so via database calls or application programming interfaces (APIs). In such use cases, dictionaries serve as a background resource for the applications developed by others.

In both cases, lexicographic information needs to be **accessible** to non-lexicographic applications via computer networks and different information fields need to be identifiable, in order to retrieve the type of information needed by a specific use case.

If academic lexicography wants to fulfil its mission to make its word knowledge widely available and usable, **lexicographic data needs to be suitable for different purposes**, in different forms, depending on the application. It therefore needs to make its data flexibly accessible to different types of applications, otherwise its relevance will steadily decline further.

The Delpher Lexicon Service

Delpher[15] is a large, searchable archive of 120 million pages of Dutch books, newspapers and magazines. Users, by default, opt for query expansion to historical spelling variants of their queries. This expansion is done through an API search in a relational database, containing the INT GiGaNT Historical Lexicon.

Lexicography is challenged to remain relevant, reinvent itself, adapt to changes, and (continue to) serve society. A number of **challenges** remain that will be revisited in the other sections of this White Paper:

- **Business model**: If lexicographic data becomes a basic infrastructure that can be used for both traditional dictionaries and as data for applications, what is the model for funding lexicographic projects? Should such projects receive structural public funding like any basic infrastructure? Or should they develop new revenue models similar to other online content?

- **Genericity versus specificity**: If lexicographic data is used by a variety of applications, it becomes much harder to define a specific audience and user type for which the lexicographic content is intended and that lexicographers have in mind when creating lexicographic content. Usage data and user feedback also become harder to monitor and harder to take into account for content improvement. How should lexicographic institutes manage this mediated relation to users?

- **Information organisation**: Within a Linguistic Linked Open Data (LLOD) framework, the basic lexicographic unit is no longer the traditional dictionary entry linked to a word lemma, but rather the separate fields, i.e. the different pieces of information about the lemma within each entry. On the one hand, this leads to new standardisation issues around (meta)data formats to ensure retrievability and accessibility. On the other hand, new data exploitation models have to be developed that allow the different types of lexicographic information to be recombined in a customised and relevant way for many different types of users and usage.

Content Creation: Integrating Lexicographic Data

Digital dictionaries were often conceived as stand-alone reference works, their underlying databases self-contained. They included the information that goes into the dictionary's entry for each word. These databases are often **not well integrated with other linguistic resources** that may contain data on these same words, such as corpora and lexical databases with semantic relations between words (thesauri). Words may also be linked with their occurrence in multiword units and syntagmatic patterns (constructicons), and with frequency information about usage across socio-linguistically stratified corpora, with numeric representations of frequency distributions or semantics over contexts (word vectors, embeddings), and with psycholinguistic properties of words like their associations or valency, neurolinguistic properties of words in terms of brain activation patterns.

It is exactly this integration and linking of different types of linguistic data on a large scale that holds the greatest promise both for scientific research and for the development of new practical

[15] http://www.delpher.nl

applications. Several lexicographic institutes, such as the *Instituut voor de Nederlandse Taal* for Dutch, the *Institute of the Estonian Language* for Estonian and the *Jožef Stefan Institute* for Slovene, are already making the move towards reorganising their data with the purpose of **integrating different linguistic resources and different usage scenarios**. Such smart use and reuse of lexicographic data was even the central theme of the eLex conference of 2019,[16] and it is one of the aims of the ongoing European ELEXIS project to link up different dictionaries through existing multilingual lexical resources like BabelNet.[17]

Examples of Integration of linguistic resources

The *MentalLex* project, hosted at the INT, tried to link different lexicographic and psycholinguistic lexical databases for Dutch. The *Sõnaveeb Language Portal* for Estonian links many different resources for Estonian.[18] Another example of such linking is the *Historical Thesaurus of English*[19] which is linked to the Oxford English Dictionary.

One of the ongoing ELEXIS project's goals is to turn existing dictionaries into resources that are connected as part of the Linguistic Linked Open Data (LLOD) infrastructure. Through the LLOD infrastructure access to lexicographic data can be offered as a service (Lexicographic Data as a Service -- LDaaS).

The Prêt-à-LLOD project aims to exploit the combination of linked data and language technologies, i.e. LLOD, to create ready-to-use multilingual data.

The W3C Ontolex group has worked on a **Lexicon Model for Ontologies** (LeMOn).[20] The aim of *LeMOn* is to provide rich linguistic grounding for ontologies. Rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e. the meaning of these lexical entries with respect to an ontology or vocabulary.

However, some important **challenges** remain, amongst which:

- **Genericity vs. Specificity:** When lexicographers become content contributors to an integrated generic lexical database, project specific specifications no longer play their traditional role. Nevertheless, when a specific project is executed, newly entered (project specific) lexical data (e.g. word lists for a learner's dictionary) should be compliant with the generic lexical database structure and compatible with already existing lexical entries. How can lexicographic content be generic and specific simultaneously? How can we find the right balance between genericity vs. specificity? How can the content creation process cater to a multitude of applications and stay open-ended, so that future, as yet unknown, applications can also draw from it?
- **Linking information at different levels of abstraction**: Dictionaries include *word knowledge* at different levels of abstraction. Whereas word forms are directly observable in corpus

[16] E-lex conferences are conferences on electronic lexicography. https://elex.link/elex2019/

[17] https://babelnet.org

[18] K. Koppel, A. Tavast, M. Langmets, and J. Kallas (2019). Aggregating Dictionaries into the Language Portal Sõnaveeb: Issues With and Without Solutions. eLex 2019.

[19] https://ht.ac.uk/

[20] https://www.w3.org/2016/05/ontolex/

data, meanings, as codified in dictionary definitions, are high-level, schematic abstractions over usage data that differ from dictionary to dictionary. Other types of lexical resources, like thesauri or psycholinguistic word data, may also contain lexical information on a schematic level. These types of abstract knowledge representations are much harder to link and map onto each other than directly observable word forms. What techniques are available for such fuzzy mappings and how and in which format should they be included in an integrated, linked lexical database?

- **Systematicity and exhaustiveness**: Different resources will cover a non-perfectly overlapping vocabulary. Externally provided data will involve just a subset of the LLOD vocabulary. How can this issue be systematised when integrating several sources. When working with corpus data, not all items culled from the corpus can receive exhaustive treatment; so decision criteria should be described.

- **Explicitation of the content creation process:** Dictionaries aim to record the most essential knowledge about words in a concise way (for a given audience). This content is created by experts (lexicographers) through a process of selection and abstraction over data contained in other language resources, primarily corpora but also in existing reference works. However, this creation process is rather scarcely documented and only the end product, the dictionary entries are published in a digitally processable format. Because the expertise and *human* intelligence going into the creation process is largely hidden in the available output, (partial) automation of the creation process with *artificial* intelligence might be severely hampered. How can the content creation process be made more explicitly documented in a way that is workable and beneficial for the lexicographers involved? In what way is the content creation process transparent, and how can the end user in the digital era check the evidence -- what is the contemporary equivalent of checking the evidence in a corpus quotation?

Data Processing: Corpora in the Internet-era

Academic dictionary projects traditionally are based on a scientifically underpinned analysis of a well-delineated, carefully compiled electronic corpus that aims to be representative of general language use. However, the advent of the internet and social networks has revolutionised **the way language users communicate** and the **amount and types of language data** that is available.

The "general language use" that academic lexicography aims to capture is not well represented by a corpus that is compiled at the beginning of a dictionary project and that is based on traditional categorisations of styles, registers and text types.[21] Instead, academic lexicography needs to adjust to an empirical basis that is constantly shifting and exponentially growing, and is facing the following **challenges**:

- **Monitoring language use:** If compilation of a well-delineated, general reference corpus is so difficult, which strategies are available to academic lexicography (and corpus linguistics at large) to monitor important trends and changes in language use?

[21] Lee, David. (2002). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Teaching and Learning by Doing Corpus Analysis.* Brill Rodopi, pp. 245-292.

- **Scalability:** The sheer amounts of language data available make it impossible to process corpus data in traditional work flows. What does a *big data* approach for lexicography look like? How can data processing be automated as much as possible? And how do you provide the lexicographer with tools necessary for proper data analysis?
- **Balanced Data:** Even though the amount of (online) language data grows exponentially, not all data is easily available to lexicographic institutes. Some types and registers of language use are not well represented in web-crawled corpora. Other types of data are difficult to get by, due to copyright, IPR, GDPR and other use restricting regulations. How can we address these issues?

Setting up a Monitor Corpus

At INT we receive weekly data dumps of several Dutch and Flemish newspapers, providing us constantly with new data, allowing us to monitor language change, at least for newspaper language. We should look into how to obtain similar monitor data for other language registers.

We are implementing an automated corpus processing approach for such new data, so that they automatically become available in the Corpus Hedendaags Nederlands, which serves as the basis for building the Algemeen Nederlands Woordenboek. Processing currently consists of converting the data in TEI format, part-of-speech tagging and lemmatisation. Additional processing, like full syntactic parsing, is still in the pipeline.

Challenge 1. Business Model, Ecosystem and Partnerships

Scientific role and funding

Traditionally, lexicography has been seen as an applied branch of linguistics, and more specifically, of lexicology. However, as in linguistics in general, the advent of computers and large amounts of corpus data have given lexicography a **strong relation with information technology and computer science**. As a digital knowledge resource, dictionaries are also important tools in the digital humanities, cognitive, educational, data, knowledge and neighbouring sciences.

To adapt to this changing ecosystem requires **additional funding**, apart from the traditional governmental support for lexicography. To make the move towards a truly modular, multifunctional lexicographic resource that can cater to many different functionalities and users requires several research and developmental innovations that go beyond the traditional lexicographic work.

Such funding can be obtained by submitting research projects to funding agencies and governments, in cooperation with other scientific organisations, such as university research groups. Such research could be fundamental in nature, but it is more likely that practically useful results will be obtained through applied interdisciplinary research, in which specific research questions about lexicographic innovation are addressed, or which result in answers to lexicographic research questions as a side effect or side product of the main project goals.

Based on the recommendations expressed in the workshop discussions, representatives of the different lexicographic institutes have also committed themselves to **strengthening the ties to the**

neighbouring disciplines, by organising workshops and training about big data and artificial intelligence, both at lexicographic conference venues (e.g. eLex, EuraLex, Globalex), in graduate training programmes (e.g. European Master in Lexicography) as well as within institutes.

INT Intention

INT will adjust its long term policy plan to organise additional training and knowledge exchange between lexicographers, computational linguists and software developers and to become an associated member of the European Master in Lexicography.

Projects should be set up in **cooperation with university research groups**, in the framework of master and doctoral theses and internships for domains such as linguistics, digital humanities, educational science, cognitive science, information science, computer science and artificial intelligence. As such, artificial intelligence and natural language processing specialists also gain insight into lexicographic products and the lexicographic production process.

Cooperation with universities

INT is discussing with the Institute of Computer Science of the University of Leiden which master thesis subjects in computer science could be set up. INT is also cooperating with KU Leuven, setting up internships in artificial intelligence and master thesis subjects in linguistics, through cooperation with the Leuven Institute for Artificial Intelligence, and the Linguistics department. These collaborations could be further explored, with other departments and other universities.

The workshop discussions clearly indicated the need for a **permanent body** at the **European level**, to **structurally support joint, cross-disciplinary research and development** in lexicography, for all European languages (and not only English). The organisers and participants have committed themselves to actively explore the options for such a permanent body as a follow-up of the ongoing (but time-limited) EU research infrastructure project ELEXIS. How such a permanent body relates to the European Federation of National Institutes of Language (EFNIL)[22] needs to be discussed. Further cooperation through Horizon Europe[23] projects is a complementary option, but remains non-structural.

Already during the workshop, the first steps were undertaken to **form the nucleus of a consortium** between the Dutch Language Institute (INT - main organiser) and other scientific lexicographic institutes (a.o. *Zentrum für digitale Lexikographie der deutschen Sprache -- ZDL*, *Danske Sprog- og Litteraturselskab -- DSL*, *Dansk Sprognævn*).

International cooperation

On the 3rd of March 2020, a delegation of INT paid a visit to ZDL to mutually gain more insight into the operation of both institutes and to start up a consortium of (initially) three parties (ZDL Berlin, INT Leiden and DSL Copenhagen), in order to intensify collaboration in the implementation of new techniques in digital lexicography.

[22] http://www.efnil.org/
[23] https://ec.europa.eu/info/horizon-europe-next-research-and-innovation-framework-programme_en

The importance of **special interest groups** and of informal meetings was also discussed. This allows us to share experiences and exchange the *do's* and *don'ts.*

Societal role and funding

The traditional role of academic lexicographic institutes was to document and provide a **historic record of a *national* language** through the description of its vocabulary that was based on authoritative sources of attested language use. They provided society with a public monument of cultural heritage.

However, in the last decades, the role of these lexicographic institutes has expanded. At their core, they still document language use for the benefit of the entire language community but **the scope and purpose has diversified**.

They do not only make dictionaries, but also **compile other linguistic resources** like corpora, thesauri and language processing tools. Their purpose is not just to create a historic record but also to provide practical resources for researchers, companies and other users. These resources are not only useful in the field of linguistics, but also in the much broader field of the *digital humanities*.

However, this expanded role makes it sometimes **difficult to delineate** exactly what activities language institutes should engage in and how these activities are financed.

Policy makers should decide which activities are financed structurally with public means and which can be financed on a competitive basis. National language institutes can appeal to a **sense of responsibility** with regard to the language and its users.

If creating resources required to support digital presence for a language or sublanguage is **not commercially viable**, then it is the role of the national institute to build these required linguistic resources to support digital presence for that language, and to ensure avoidance of digital extinction.

The specific challenges for **sign language lexicography** were an eye-opener and a concrete starting point for further interdisciplinary cooperation.

INT and Sign Language

INT is involved in the Horizon 2020 SignOn project about automated sign language recognition and translation, which should result in improved accessibility and annotation of sign language corpora and sign language dictionaries, for, amongst others, Flemish and Dutch sign language. The Flemish Sign Language Centre (VGTC) is another consortium partner in this project, so cooperation between VGTC and INT will become formal when this project starts (early 2021).

Challenge 2. Scalability: Removing the bottlenecks

Improving the lexicographic workflow

In order to determine the bottlenecks in the current lexicographic workflow, and to see whether automation through the application of techniques from artificial intelligence can remove them, it is important for the researchers in artificial intelligence to thoroughly **understand what lexicographers**

do and how they do it, what bothers them in their current workflow, and what the lexicographers would like to see as improvements in their workflow.

Observing lexicographers at work

At INT, we set out for a detailed observation by technologists of the daily work flow of lexicographers, who were showing how they work right now, explaining what bothers them and which features they would like. Based on these observations, a report was written and discussed with a wider group of technologists and lexicographers. An estimation was made which suggestions had the highest return on investment, and these were treated first.

Once such an understanding is established, technologists can suggest points where automation is feasible, or where automation would be an interesting research topic. When new technologies like *deep learning* and *big data analytics* can partly automate the lexicographic workflow, this has consequences for the **role of lexicographers**. Which new skills for lexicographers and which new process workflows does this entail? Importantly, will partial automation not simply shift the bottleneck?

When e.g. dictionary definitions are automatically generated, lexicographers, like translators, could become mainly **post-editors** of automatically created content and apply their expertise to quality assurance. Their role could also include the creation of gold standard data, for training or testing artificial intelligence methods.

Or, alternatively, highly interactive integration of human expertise with *artificial intelligence* could lead to so-called ***augmented intelligence*** *or* ***hybrid intelligence,***[24] similar to *interactive machine translation*, in which the lexicographer starts typing the definition, and the artificial intelligence tries to automatically complete the typing of the lexicographer.

A more **high level role** could be that lexicographers oversee lexicographic projects, by mediating between use cases and technologists, helping with the appropriate translation of the lexicographical task into a technological solution with task specifications.

A joint team of lexicographers and technologists

At INT we have created a team of everyone who is working on contemporary Dutch. This team includes lexicographers, computational linguists and computer programmers. On a regular basis, the team holds meetings, in which issues in the current workflow and suggestions for automation and automation research are discussed.

With the introduction of computers, important **parts of the scientific dictionary making process have already been automated**. These include corpus compilation, lemmatisation, neologism detection, identification of key word usage characteristics like collocations, syntactic behaviour, stylistic or regional preferences, and the selection of good dictionary examples.

[24] https://www.hybrid-intelligence-centre.nl/

Although so-called One-Click Dictionary[25] approaches already offer some preliminary automatic identification of word senses, semantic relations and definitional contexts as input to the lexicographer, the in-depth lexicographical analysis and definition writing step is **still a mainly manual, time-consuming enterprise** by highly trained specialists, and, as a consequence, the main bottleneck to the continuous updating of scientific dictionaries.

We discussed how realistic and feasible it is to **automate these *knowledge abstraction*** steps using innovations in artificial intelligence, such as deep learning and big data analytics. To know whether deep learning and other state-of-the-art techniques from machine learning, trained on expert-generated lexicographic data, can automate other steps in the dictionary process, research projects have to be set up. If such automation proves feasible, it might allow us to compile constantly updated, high quality dictionaries.

Nevertheless, such approaches face the fact that neural networks are often black boxes, and that certain outcomes might be ***unexplainable***.

Research proposal example

We could set up research on automated dictionary definition writing,[26] using existing dictionaries as training data (keeping a certain percentage of definitions apart, for evaluation), and using e.g pre-trained BERT embeddings[27] for Dutch. BERT embeddings encode word meaning in a continuous semantic space, taking into account the specific sense of the word in a certain context. Several sets of BERT embeddings for Dutch are already available, trained on large Dutch corpora including SoNaR.[28] We could use a *generative adversarial network* (GAN)[29] in which two neural networks are trained which play the following game: one network generates dictionary definitions, while the other network learns to classify whether definitions are *synthetic* or human-made. The definition generating network will try to outsmart the classifier, and, as such, learn to generate better definitions .[30] Another option would be to look into the potential of the GPT family.[31]

Not all applications of artificial intelligence have to be of such a grand ambition. We should also look into how techniques from artificial intelligence can be used efficiently in a production environment, and how these techniques can contribute, in collaboration with a lexicographer, to a high quality product.

Deep neural networks and word embeddings could also be used on **smaller problems**, such as

[25] https://www.youtube.com/watch?v=TaC8sTFWkqs provides a video of the presentation M. Jakubíček, V. Kovář, M. Měchura and P. Rychlý. (2017). One-Click Dictionary. eLex. Electronic lexicography in the 21st century: Lexicography from Scratch. Leiden.

[26] Possibly similar to Sørensen, N. H., & Nimb, S. (2018). Word2Dict–Lemma Selection and Dictionary Editing Assisted by Word Embeddings. In The XVIII EURALEX International Congress.

[27] J.Devlin, M. Chang, K. Lee and K. Toutanova, (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.

[28] A frequently used framework for BERT models is HuggingFace. A number of models for Dutch are available, such as Delobelle et al. (2020) RobBERT: a Dutch RoBERTa-based Language Model. https://arxiv.org/abs/2001.06286 and de Vries et al (2019). Bertje: A Dutch BERT model. https://arxiv.org/abs/1912.09582

[29] https://en.wikipedia.org/wiki/Generative_adversarial_network

[30] Much like the dictionary game described in https://news.sophos.com/en-us/2018/07/06/generative-adversarial-networks-explained/

[31] https://openai.com/projects/

- improving consistency of definitions through spell checking and resolving spelling variation over different definitions;
- outlier form detection, to spot where things went wrong, and to indicate where it's probably worth for lexicographers to take a close look;
- sense clustering, automatically grouping corpus examples with the same or similar sense

Word embeddings could also be used to create or **extend existing symbolic knowledge bases**, or to assist in manual extension of such knowledge bases. One could think of ordinary wordnets, but also of multilingual resources, and collections of multiword units, which could even be discontinuous. While lexicography is often word-based, there is no need to restrict the analysis to the level of the individual words.

Research Internship

INT has proposed an internship at KU Leuven for students in Artificial Intelligence, in which classifiers are trained to predict Open Dutch Wordnet sense relations (synonymy, hypernymy, ...), testing the accuracy of such predictions for different algorithms. Once a sufficiently accurate approach is found, the automatically created relations can be used to automatically extend the existing Open Dutch Wordnet.

INT will also (try to) involve Dutch universities in similar internships.

One can also consider new tasks for lexicographers, instead of word-by-word descriptions, such as trend analysis and other tasks which were not feasible in the pre-digital era. While such tasks have not been widely discussed at the workshop, they are worth a discussion. Is it the ambition of academic lexicographic institutes to go beyond traditional lexicography?

Crowdsourcing

Next to automation, involving a wide **group of volunteers** is another potential avenue to overcome the bottlenecks in scientific lexicography, while at the same engaging the wider language community.

Up to now, mainly ***micro-tasks*** like the (correction of) corpus annotation or collection of word familiarity judgements have been crowdsourced. We discussed whether higher level and conceptually more abstract tasks like definition writing and the analysis of sense relations can be crowdsourced as well. It was debated whether Wikipedia, with its crowd-sourced knowledge summarisation, provides a good model, or whether the additional step of abstraction-from-corpus data in lexicography requires a fundamentally different approach.

It was furthermore discussed how **validation** mechanisms can be set up, depending on specific crowd group properties (does it concern e.g students, (ex-)teachers, or other groups of volunteers) through e.g. reputation markings, and how the crowd's interest is kept at a high level, by introducing motivational features such as gamification and scoring mechanisms, ranking users.

Challenge 3: Genericity vs. Specificity

Diverse End-Users

In the last decades, society has grown much more diverse and now includes **language users with myriad immigration and linguistic backgrounds, education levels and language skills**. Their needs and expectations with regard to linguistic aids and reference works, both in educational and professional settings, vary extensively.

Although scientific dictionaries have been reaching out beyond academia to new audiences for some time, the ability to **customise** them to specific user needs is still quite limited.

On a more **fundamental level**, it is an unresolved issue whether the detailed and fine-grained descriptions from scientific lexicography can be easily "simplified" or customised or whether these "academically biased" reference works are even the best starting point to do so.

In the workshop, we discussed whether and how it is possible to set up a **generic lexicographic infrastructure** that can be used as the basis from which usage-specific knowledge bases can be derived. This aspect is linked with the workshop section on new technologies.

It is also worth investigating whether data analytics for **user and usage modelling** can be integrated with innovations in lexicographic automation to produce highly customised lexicographic tools.

Lexicography as an infrastructure

To enable reuse of lexicographic data in a wider context, we have to change the goal of academic lexicography from building a specific dictionary to building a central database and infrastructure that can be adapted for specific use cases. This includes considering how such an infrastructure can be useful for human users as well as NLP applications. This is a **changing ecosystem for lexicography**: lexicographical knowledge is used for different functions and in different shapes. Building such an infrastructure requires rethinking the lexicographical process as a modular approach.

INT Example

INT is linking all its dictionaries and databases into a central lexicon, called GiGANT, a computational lexicon of the Dutch language from the sixth century up to the present. This lexicon is a collection of words and word groups, including named entities (names of persons, places, organisations), showing every possible variant of spelling and form.[32]

The integration of scientific dictionaries into the web's knowledge infrastructure requires the representation of dictionary entries in a standardised format, such as **Linked** (Open) **Data.** It requires making this data available so that NLP applications for text mining and data processing can use this in-depth lexicographic knowledge.

Although projects like ELEXIS, that include creating lexicographic linked data, are underway, it is unclear how the unique content of scientific dictionaries with their fine-grained meaning and usage descriptions for specific languages can be fully linked up and integrated into a **multilingual**

[32] https://ivdnt.org/onderzoek-a-onderwijs/corpora-lexica/gigant

knowledge infrastructure. Projects like BabelNet try to link lexicographic sense descriptions across languages via existing ontologies like WordNet.[33] However these are mainly based on English and, like a procrustean bed, such a linking could force the sense descriptions of other languages to align with how English, and more specifically WordNet, carves up the world into concepts, words and meanings. A solution could be to look for universal concepts and universal semantic frames and start building multilingual knowledge bases from there.

In the workshop, we discuss which innovations in data linking would allow us to do justice to the **unique world views offered by specific languages** and their scientific dictionaries and to leverage this rich diversity in the multilingual knowledge web.

Infrastructure Cooperation

A common interest in a Europe-wide cooperation in historical lexicography on, among other topics, a **shared etymology database** became so apparent that further cooperation between historic lexicographers from Oxford, Leiden and other universities and institutes has already been initiated.

Conclusions and Recommendations

During an entire week we set up discussions on a whole plethora of topics related to the present and the future of lexicography. Apart from the inspiration and the consolidated network of the participants of the workshop, this paper is **the main *deliverable*** of the workshop.

It is clear that the art, or rather craft, of lexicography is changing from the old school dictionary writing handicraft into a more contemporary, **efficient and reusable effort**. We discussed how the scientific and societal role have changed and how this affects funding, how artificial intelligence and natural language processing can benefit from lexicographic resources, and how building lexicographic resources can, in turn, benefit from the state-of-the-art models and tools developed in artificial intelligence and natural language processing. We talked about how this can affect the daily work of lexicographers, and how some of the knowledge acquisition bottlenecks might be removed through automation and *offshoring* tasks to the crowd, and how quality can still be assured in such cases. We agreed upon the fact that we have to build our lexicographic resources as an infrastructure, as such that they become modular and reusable, allowing us to recuperate efforts on one use case for the development of other use cases. After the discussions, it is time for action.

For further inspiration and understanding of what the possibilities are and how they can be used in practical production tasks, it would be good if the examples in this document could be extended by other lexicographic institutes.

Recommendations

1. **Strengthening the ties between disciplines**
    a. Communication between lexicographers and computational linguists at conferences and training programs

[33] WordNet is, of course, not a dictionary, though incorporating some lexicographic aspects.

   b. Cooperation with university research groups in artificial intelligence and natural language processing in order to set up applied research focused on lexicographic issues
   c. Need for a permanent body at the European level to structurally support joint, cross-disciplinary research and development in digital lexicography

2. **Building a generic lexicographic infrastructure instead of writing specific dictionaries**

   a. Link a generic lexicon (generic lexicographic resource) with different, already existing lexicographic resources
   b. Embed the generic lexicon in an international, multilingual context
   c. New, use-case-specific lexicographic resources can reuse information from the generic lexicon, improving consistency and productivity
   d. Quality assurance methodologies are needed for linking of lexicographic resources

3. **Realism combined with ambition**

   a. Small, realistic improvements can make a huge difference in practical lexicographic work
   b. Automation and crowdsourcing can remove bottlenecks, but require adequate quality assurance methodologies, involving the lexicographer

Acknowledgements

The editors of this paper thank the people that have reviewed and commented upon previous versions of the paper, especially Katrien Depuydt, Carole Tiberius, Suzan Verberne, Dirk Geeraerts, Thierry Declerck and Ilan Kernerman.

This paper is tributed to Tanneke Schoonheim, who tragically traded the temporal for the eternal on the 25th of August 2020.