Kernerman
DICTIONARY News
Number 23 • June 2015
kdictionaries.com/kdn

Multilingual dictionaries and the Web of Data

Jorge Gracia

1. Introduction

Nowadays, we are witnessing a growing trend in publishing language resources (lexicons, corpora, dictionaries, etc) as Linked Data (LD) on the Web. LD refers to a set of best practices for exposing, sharing and connecting data on the Web (Bizer et al 2009). In short, the LD paradigm requires that (i) resources are represented on the Web via HTTP URIs (Unique Resource Identifiers), (ii) once a resource is accessed via its URI, information about it is obtained, and (iii) such information contains links to other resources. The basic mechanism to support the representation of resources and their related information is the Resource Description Framework (RDF[1]), which follows the *subject-object-predicate* pattern. Resources can be anything, including documents, people, physical objects and abstract concepts. Following LD principles, a 'Web of Data' emerges in which links are at the level of data, as a counterpart to the "traditional" Web in which links are established at the level of documents (e.g. hyperlinks between webpages).

Publishing language resources as LD offers clear advantages to both the data owners and data users, such as higher independence from domain-specific data formats or vendor-specific APIs, as well as easier access and re-use of linguistic data by semantic-aware software agents. Further, we argue that reaching a critical mass of linguistic data as LD on the Web will set the basis for a new generation of LD-aware Natural Language Processing (NLP) services, with improved scalability and better interoperability level. The latter is, in fact, one of the motivations of LIDER[2], a European project that is driving a remarkable community effort in that direction.

In this context, the Ontology Engineering Group (OEG[3]) at Universidad Politécnica de Madrid has started converting a series of bilingual dictionaries and multilingual terminologies and publishing them as LD on the Web. In the following paragraphs we briefly present the RDF conversion process that we have followed, and report on our experience with two of these datasets: Apertium and Terminesp.

2. RDF generation of bilingual and multilingual dictionaries

Recently the W3C Best Practises for Multilingual Linked Open Data (BPMLOD) community group[4] has proposed a set of guidelines for the LD generation of language resources. In particular, the guidelines for bilingual dictionaries[5] identify five steps, namely: (i) vocabulary selection, (ii) modelling, (iii) URI

1 http://w3.org/TR/rdf11-primer/

2 http://lider-project.eu/
3 http://oeg-upm.net/
4 http://w3.org/community/bpmlod/
5 http://bpmlod.github.io/report/bilingual-dictionaries/index.html/

Editor | **Ilan Kernerman**

This issue is dedicated to the memory of **Adam Kilgarriff**

K DICTIONARIES LTD
8 Nahum Hanavi St.
Tel Aviv 63503 Israel
Tel: +972-3-5468102
kdl@kdictionaries.com
http://kdictionaries.com

ISSN 1565-4745

Jorge Gracia is post-doctoral researcher at the Ontology Engineering Group, Universidad Politécnica de Madrid. He got his PhD in Computer Science at University of Zaragoza in 2009 with a thesis about heterogeneity issues on the Semantic Web. His current research interests include multilingualism and Linked Data, linguistic linked data, and cross-lingual matching and information access on the Semantic Web. Currently he is exploring how to move language resources (lexica, dictionaries, corpora, etc) from their data silos into the multilingual Web of Data and make them interoperable, in support of future generation Linked Data-aware NLP tools.

http://jogracia.url.ph/web/

design, (iv) generation, and (v) publication. A similar approach can be followed for multilingual dictionaries as well. We are not going into the details here, but will just highlight key aspects.

In order to represent the lexical information contained in the original dictionaries, we relied on the LExicon Model for ONtologies (*lemon*[6]), a de-facto standard for representing ontology lexica. We used the *lemon translation module*[7] to represent explicit translations between languages (Gracia et al 2014). As a result of the conversion into RDF of a bilingual dictionary, a *lemon* lexicon is defined per language, where all the *translations* corresponding to a pair of languages are grouped under the same *translation set*. A *translation set* groups a set of translations sharing certain properties, for instance stemming from the same language resource, or belonging to the same organisation, etc.

To design the URIs of our RDF datasets, we adopted the patterns and recommendations proposed in the context of the ISA program (Archer et al 2012). In order to construct the URIs of the lexical entries, their senses and other elements, we preserved the identifiers of the original data whenever possible, propagating them into the RDF representation. This is, for example, the URI that points to the Apertium English-Spanish translation set: http://linguistic.linkeddata.es/id/apertium/tranSetEN-ES/.

One of the most interesting aspects of our conversion methodology is that, by being consistent with the generation rules of every URI assigned to every element in the model (lexicons, lexical entries, lexical senses, translations, etc), each time a bilingual dictionary is converted into LD, its monolingual lexicon is not created again if it already exists but its lexical entries are shared by two or more translation sets (see Figure 1). This allows the dynamic growth of monolingual lexicons and, more significantly, shared lexical entries serve as pivot nodes in the graph to allow getting indirect translations from two languages initially disconnected in the original dictionaries.

The generation step deals with the transformation into RDF of the selected data sources using the chosen representation scheme and modelling patterns. Depending on the format of the data source, there are a number of tools that can be used to support this task. In our work we have used Open Refine[8] (and its RDF plug-in) in a preferred way.

Finally, the generated RDF data has been loaded in a triple store and made accessible through a single SPARQL endpoint. In that way, all the data from the original dictionaries were made accessible as LD on the Web in a unified graph with lexical entries, senses, translations, etc, as nodes. All the nodes were identified with dereferenceable URIs. Such data can be accessed by means of SPARQL clients and RDF and HTML browsers.

3. Apertium

Apertium[9] is a free open-source machine translation platform. Its translation engine consists of a series of assembled modules that communicate with each other using text streams. One of the modules, the lexical transfer module, uses a bilingual dictionary to deliver the corresponding target lexical forms from a given lexical form in the source language. Many of the Apertium bilingual dictionaries are available in the Lexical Markup Framework (LMF) XML-based format[10]. We took such LMF dictionaries as a starting point in our process to publish the Apertium data as LD.

We used *lemon* and its *translation module* as representation schemes. Once the conversion into RDF was completed we published the Apertium data on the Web in accordance with the LD principles. The result is a set of 22 Apertium RDF bilingual dictionaries, which can be found in the LLOD cloud[11]. The following languages are currently represented in Apertium RDF: Spanish, Catalan, English, French, Italian, Romanian, Asturian, Aragonese, Basque, Galician, Portuguese, Occitan, Esperanto. The whole dataset contains 400,808 translations and, after its conversion into RDF, 8,842,510 RDF triples were created. The LD version of Apertium groups the data of the (originally disparate) Apertium bilingual dictionaries in the same graph, interconnected through the common lexical entries of the monolingual lexicons that they share. Figure 2 shows the network of interconnected languages in Apertium RDF.

We made all the generated information accessible on the Web both for humans (via a Web interface[12]) and software agents (with a SPARQL endpoint[13]).

6 http://lemon-model.net/
7 http://purl.org/net/translation/
8 http://openrefine.org/

9 http://apertium.org/
10 A complete list can be found at http://lod.iula.upf.edu/types/Lexica/by/standards/.
11 http://linguistic-lod.org/, and see also back cover.
12 http://linguistic.linkeddata.es/apertium/
13 http://linguistic.linkeddata.es/apertium/sparql-editor/

4. Terminesp

Terminesp is a multilingual terminological database created by AETER (Asociación Española de Terminología)[14] that contains the terms and definitions from Spanish technological norms (standards) including more than thirty thousand terms, many of them with translations available into another language (English, French, German, Italian, Swedish). The terminology contained in Terminesp is highly technical and specific of domains such as electrical engineering, aeronautics, marine technology, etc.

We converted the original data (an MS Access database) into RDF by using the *lemon-ontolex* model as representation scheme. The *lemon-ontolex* model is the next version of *lemon*, developed under the umbrella of the W3C Ontology Lexica (Ontolex) community group[15]. At the time of writing, the *lemon-ontolex* model is nearly finished and waiting for final corrections by the community to be officially released.

We established an automatic mechanism to extract the lexical entries from Terminesp database and instantiate the *lemon* lexicons. Translation sets were created between Spanish to French (13,996 translations), German (12,593), English (14,936), Italian (802) and Swedish (67). Differently from the Apertium RDF graph, the Terminesp RDF graph follows a star topology (see Figure 3), with Spanish as hub and the other languages as peripheral nodes.

In addition to accounting for explicit translations, we extended the original Terminesp dataset with part-of-speech and syntactic information that was not explicitly declared in the original data (e.g. nominal, prepositional and adjectival phrases), as well as some terminological variations (Bosque-Gil et al 2015). The whole conversion into RDF resulted in 1,095,051 triples. Since *lemon-ontolex* is still under development, the resultant RDF files are not published as LD yet. However, a preliminary version of Terminesp RDF, restricted to Spanish, English and German, and with less rich syntactical information, was already published in October 2013 as LD using *lemon*[16].

5. The emergence of a unified single graph of translations

The publication of the Apertium dictionaries as LD resulted in the creation of a large unified graph of linked lexical entries, senses and translations on the Web. The URIs of all these elements can be seen as the nodes of this graph. Every URI is dereferenceable, meaning that when it is accessed, a response is obtained with its attributes and links to other elements in RDF. There are several ways to access and explore the graph (both for software agents and humans), such as by querying through the SPARQL endpoint, by using dedicated search interfaces[17], or by following the links as they are represented in an LD interface such as Pubby[18].

Some of the advantages of having all the lexical information and translations in the same RDF graph are:

- As we have seen above (Figure 1), monolingual lexicons grow with time

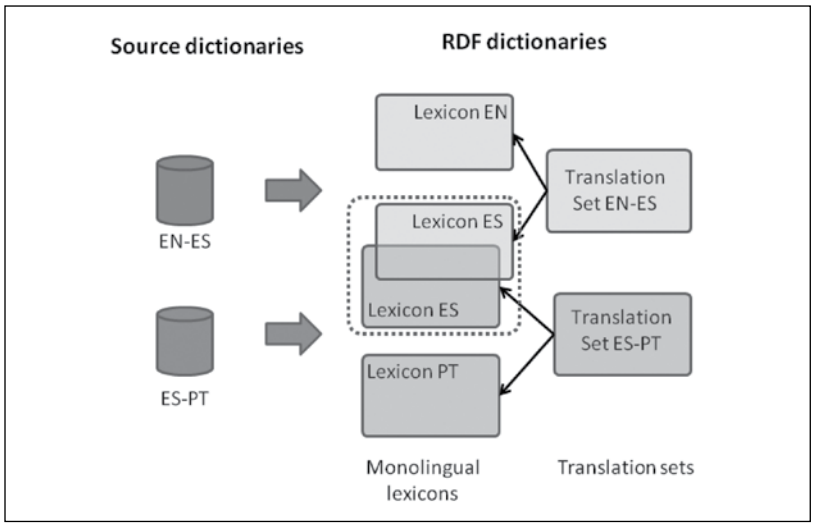


Figure 1: Example of the conversion of two bilingual dictionaries (EN-ES and ES-PT) into RDF

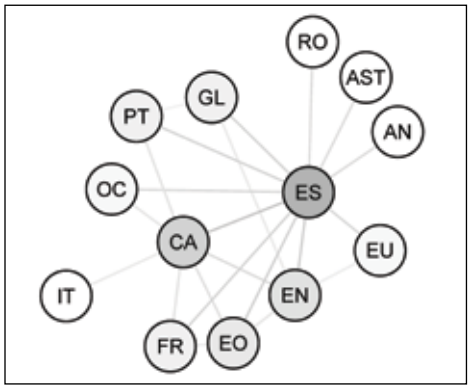


Figure 2: Network of languages in the Apertium RDF Graph (nodes are languages and edges are translation sets)

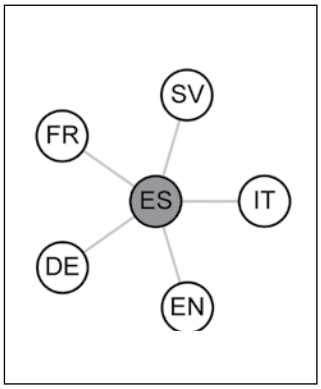


Figure 3: Network of languages in the Terminesp RDF graph

14 http://aeter.org/

15 https://w3.org/community/ontolex/

16 http://linguistic.linkeddata.es/terminesp/

17 http://linguistic.linkeddata.es/search/

18 For example, one can introduce the URI of a translation set in a Web browser http://linguistic.linkeddata.es/id/apertium/tranSetEN-ES/, and its properties and links will be shown in a human readable way (as the links are clickable, "manual" navigation through the graph is possible).

as more dictionaries of the same family are published as LD. There is no need to create a new and separate monolingual lexicon every time.

- Useful information can be obtained through a single SPARQL query in a manner that otherwise would be more difficult to get if the isolated original data sources must be queried. For instance, one can get all the possible direct translations of "network"@en into any available target language in the graph with just a single query (not needing to specify which dictionaries to look up), getting as a result the list of translated forms: {"xarxa"@ca, "red"@es, "rede"@gl, "reto"@eo, "sarera_konektatu"@eu, ...}.
- Indirect translations can be obtained between language pairs that were initially unconnected. For instance, a translation of the English term "network"@en to Italian can be obtained, with a single SPARQL query, by using Catalan as pivot language. The result is "rete"@it[19]. Notice however that some strategies have to be introduced in order to detect and exclude wrongly inferred translations. To that end we propose the use of the one time inverse consultation algorithm (Tanaka and Umemura 1994).
- Further, direct connections to other datasets in the Web of Data are possible so the original information can be enriched with additional relevant data. For instance, a high number of lexical senses in Apertium RDF have been linked to BabelNet (Navigli and Ponzetto 2012). In that way, additional descriptions, lexical relations, or even pictures, can be obtained by querying BabelNet to enrich the information that can be obtained in Apertium. For instance, one of the ontological references of "network"@en, when translated as "red"@es, is the babelsynset http://babelnet.org/rdf/s00030258n/i/, from which an English definition, not initially present in Apertium, could be obtained: "(electronics) a system of interconnected electronic components or circuits".
- Several "families" of bilingual dictionaries (Apertium and Terminesp in our case) can be published as LD under the same domain or default graph (http://linguistic.linkeddata.es/, in our case). In that way, we can have a common access point to all of them and unified SPARQL queries can be built to access these sub-graphs at the same time. For instance, a search for translations of "red"@es in Apertium could be extended with the languages covered by Terminesp, obtaining for example the German translation "Netz"@de, which is not available in Apertium originally.

In conclusion, generating linguistic Linked Data is a growing trend in the community of language resources, with clear advantages such as standardised ways of representing and accessing the data, the possibility of linking to other resources on the Web of Data, and enabling enhanced ways of discovering and aggregating the data. In this article we have briefly reported our recent experiences with the LD generation of the Apertium bilingual dictionaries and the Terminesp multilingual terminological database, and commented on the benefits of publishing their information as unified RDF graphs.

19 Examples of queries in Apertium RDF can be found at http://dx.doi.org/10.6084/m9.figshare.1352066/.

Acknowledgement

This work is supported by the FP7 European project LIDER (610782) and by the Spanish Ministry of Economy and Competitiveness (project TIN2013-46238-C4-2-R).

References

Archer, P., Goedertier, S. and Loutas, N. 2012. *Study on persistent URIs. Technical Report, Interoperability Solutions for European Public Administrations*. ISA

Bizer, C., Heath, T. and Berners-Lee, T. 2009. Linked data – the story so far. *International Journal on Semantic Web and Information Systems* (IJSWIS), 5.3: 1-22.

Bosque-Gil, J., Gracia, J., Aguado-de Cea, G. and Montiel-Ponsoda, E. 2015. Applying the OntoLex model to a multilingual terminological resource. In *Proceedings of 4th Workshop of the Multilingual Semantic Web) at 12th ESWC, Portoroz, Slovenia* (MSW'15, to appear). CEUR-WS.

Gracia, J., Montiel-Ponsoda, E., Vila-Suero, D. and Aguado-de Cea, G. 2014. Enabling language resources to expose translations as linked data on the web. In *Proceedings of the 9th Language Resources and Evaluation Conference, Reykjavik, (LREC'14)*. Paris: European Language Resources Association (ELRA): 409-413.

Navigli, R. and Ponzetto, S.P. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193: 217-250.

Tanaka, K. and Umemura, K. 1994. Construction of a bilingual dictionary intermediated by a third language. In *COLING*: 297-303.

1st Summer Datathon on Linguistic Linked Data (SD-LLOD-15)

The 1st Summer Datathon on Linguistic Linked Data (SD-LLOD-15) will be held in Cercedilla (Madrid, Spain) from 15 to 19 June 2015. It is co-organized by Jorge Gracia from Madrid Polytechnic University and John McCrae from Bielefeld University (Germany) as part of the LIDER project. Its main goal is to offer persons from the industry and academia practical knowledge in the field of Linked Data applied to linguistics, and the final aim is to allow participants to migrate their own (or other's) linguistic data and publish it as Linked Data on the Web. This datathon is the first organized on this topic worldwide and is supported by the LIDER FP7 Support Action. It will join around seventy participants (including attendees, speakers and tutors) from all around the world and will constitute an invaluable forum not only for learning but also for the exchange of experiences and ideas related to linguistic Linked Data.

http://datathon.lider-project.eu/

Enhancing lexicography with semantic language databases

Bettina Klimek and Martin Brümmer

1. Introduction

The most recent major transition in the world of lexicography has occurred barely thirty years ago as part of the emergence of information technology. The introduction of computers into everyday life marked a medial change which is progressively taking over the traditional print dictionaries that were prevalent over the last centuries. The digitization of lexical language information has formed a new broad landscape of e-lexicography. The boundaries of the printed page have dissolved to unlimited virtual space that leads to online dictionaries, translation tools, large language networks, etc. As a result, more and more linguistic information such as pronunciation, word-form paradigms, syntactic relations and dialectal varieties accompany the lexical entry. The possibilities of data processing combined with large data storage capacities assist the lexicographer in compiling as well as enriching lexical content in a structured and multi-dimensional way. Moreover, new developments in Web technologies – namely the Semantic Web and Linked Data – offer unique potential to current e-lexicography by advancing the existing consumer-oriented linguistic data towards machine-processable semantic format that enables interoperable exchange of lexicographic and other resources on the Web. This article presents the outcome of research undertaken last year with the German language dataset of K Dictionaries (KD) within the realm of Linked Data technologies along three main topics: an introduction to Linked Data and its benefits for lexicography (section 2), *lemon* – the lexicon model for ontologies (section 3), and a presentation of the conversion of KD's data from XML to RDF (section 4). Finally, section 5 presents a conclusion with a summary of the findings.

2. Semantifying lexicographic resources with Linked Data

2.1 Linked Data principles

Linked Data describes a set of best practices for publishing structured data and linking it to other datasets, providing context and aiding discoverability as well as interoperability. The concept describes machine-readable data with explicitly defined meaning that links further data. When this data is published on the Web it is called Linked Open Data (Bizer et al 2007). Linked Open Data forms a Web of Data, which consists of a machine-readable semantic network of structured data, in contrast to the unstructured HTML documents that characterize the Web. Data that is published under an open access URL (Uniform Resource Locator, see 2.2) on the Web can profit from linking to other datasets, thus increasing interoperability and easing data integration. This linking process can be considered in parallel to publicly viewable Web content, which also allows inbound document linking independently of its content. In addition, the data of a lexicon can, for example, link references to concepts in an ontology to disambiguate the meaning of lexical entries, and multiple lexicons can then be integrated on the basis of these concepts. The core principles of Linked Data, according to Tim Berners-Lee (2006), consist of:

- Use URIs as names for things,
- Use HTTP URIs so that people can look up those names,
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL),
- Include links to other URIs, so that they can discover more things.

2.2 The Resource Description Framework (RDF)

RDF is a set of specifications developed by the World Wide Web Consortium (W3C[1]) as a data model that can be used to formally describe *resources*. A resource can be anything that is uniquely identified, ranging from digital documents like lexicons, to abstract concepts like parts of speech.

Resources are identified by URIs (Uniform Resource Identifiers), which are distinct strings with a uniform syntax. One kind of URIs are those that additionally describe the primary method of access to the resource. Most URIs are URLs that describe Web documents, e.g. http://kdictionaries.com/, which can be viewed to gain more information about a resource.

In the RDF data model resources are described by statements in the form of *subject-predicate-object*, called triples, which can be understood as metadata describing resources. The subject is the resource that is described by the statement, which is uniquely identified by its URI.

1 http://w3.org/RDF/

Bettina Klimek is a Ph.D. student at the Institute for Applied Informatics (InfAI e.V.) at the University of Leipzig in her first year of research. She has graduated in linguistics (M.A.) in 2013 and is investigating the interdisciplinary application of linguistic data and Semantic Web technologies. Over the last year she gained insights into lexicographic data during her internship at K Dictionaries, converting its German language database into RDF together with her colleague Martin Brümmer.

klimek@informatik.uni-leipzig.de

Martin Brümmer is a Ph.D. student at the Institute for Applied Informatics (InfAI e.V.) at the University of Leipzig in his second year of research. He is a contributor to the NLP2RDF and the DBpedia Project, as well as to the development of the Linguistic Linked Open Data Cloud. His research focus is on Linguistic Linked Open Data, NLP in the Semantic Web and Open Government Data.

bruemmer@informatik.uni-leipzig.de

The object expresses the content of the statement, the *meta datum* itself. It can either consist of a simple string, just as the orthographic representation of a lemma in a dictionary, or a resource as such, e.g. a lexicon. Finally, the predicate constitutes the semantic link between the subject and the object and describes the meaning of the relation between them.

In order to avoid ambiguity within these semantic descriptions, predicates also have URIs that can be looked up for further information and are then called *properties*. The additional benefit is that sets of properties can be defined and documented by institutions or developers, like the LExicon Model for ONtologies (*lemon*), then be reused by other users and thus increase their interoperability and reduce the work that is usually necessary for formal definitions.

These sets of properties and associated classes of things that are needed to create and interpret RDF triples are commonly called *vocabularies* or *ontologies*. A vocabulary or ontology is a set of classes and properties that models a conceptualization of a specific domain. A large number of these vocabularies already exist and can be reused.

RDF itself is only a data model, independent of the concrete serialization, which can be realized using different formats, such as RDF/XML, N3, Turtle or JSON-LD. All serializations contain the same information but differ in readability, size and ease of parsing.

2.3 Benefits of RDF for e-lexicography

RDF offers unique benefits for e-lexicography, first and foremost by increasing the interoperability of lexicographic resources on multiple layers. As a canonical data model for such resources, RDF provides syntactic interoperability and allows usage of RDF tools, such as databases, tools for data retrieval, querying and management, as well as visualisation and data integration. On the one hand, this is useful for small and medium-sized enterprises that deal with lexicographic data but don't have a large budget for tool development. On the other hand, data management tools such as OntoWiki[2] enable collaborative data editing and research.

The nature of RDF facilitates relatively generic use of these tools without any adaptations to the schema of the data, unlike what relational databases with rigid schemas do. In the same vein, RDF vocabularies are extensible without modifications to the tools themselves, allowing further data properties to be added during aggregation and maintenance.

The second layer of interoperability offered by RDF is semantic. Unlike XML structures that confine data modelling to hierarchical trees independently of the data, RDF graphs allow data modelling according to its content in an ontological way. Relationships between different classes of objects can be explicitly defined and expressed within the data. Sharing these definitions makes it possible to model data of the same domain in the same way. In the linguistic domain of lexicography, lexical data could become semantically interoperable among different lexicons, presenting lexicographic research with a broader and more consistent basis that could be merged and combined across dataset borders. Organizations dealing with lexicographic data can also expand their datasets more easily, without costly adaption of new data to their model.

Lastly, RDF offers access interoperability by its use of URIs and, in Linked Data, HTTP as an access layer. The nature of the resulting link graph can provide unique benefits to the users of lexical data. Interlinked data incites exploration of related data sources that can enrich the lexical data with pictures, articles and other media content.

Disadvantages of RDF include the still lacking stability of existing tools and the high skills required to use it to its fullest potential. Setting up a Linked Data access point for a dataset, a database and minimal tool support require either considerable time investment or IT support. However, the advantages to be realized by proper data modelling and management, as well as the potential for collaborative data aggregation, outweigh these hurdles.

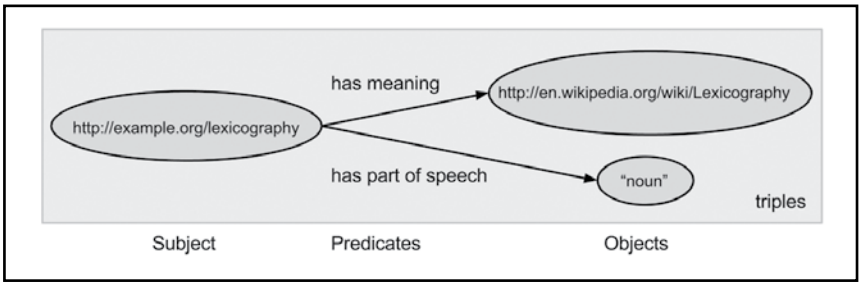


Figure 1: Example showing two triples

3. The Lexicon Model for Ontologies – *lemon*

Traditionally, standards for the design, structure and content of dictionaries have been set by established publishing houses. Now that lexicography is no longer tied

2 http://aksw.org/Projects/OntoWiki.html/

to the print medium, and is digitally transformed, the knowledge of data scientists significantly influences the way electronic language databases look like. However, just as the dictionary was bound to the limits of the book, the language database is tied to the limits of its format. This circumstance has been changed with the innovation of the Semantic Web and RDF. The reusable and interoperable character of Linked Data attracted rising numbers of participants in the compilation of lexicographic Linked Data resources. As a result, the Working Group on Open Data in Linguistics[3] collects many of them in the Linguistic Linked Open Data Cloud[4]. One significant dataset is DBnary (Serasset 2012), constituting of the RDF transformation of lexical data from Wiktionary for 13 languages and thus enabling these lexicons to be interlinked with other knowledge sources in the cloud. The model underlying DBnary is *lemon* (LExicon Model for ONtologies, McCrae et al 2011), which is highly specialized in representing lexicographic data. Other openly available datasets such as WordNet[5], PanLex[6] or Eurosentiment[7] also use *lemon* as underlying data format. Consequently, all of these datasets are interoperable and thereby pose a huge and valuable addition to any professional lexical content provider. With regard to the possibility of enriching existing resources with such open linguistic data in the future, we decided to convert the German dataset of KD by using *lemon* rather than designing a Linked Data model for lexicography completely anew. *Lemon* can be used in parts and is easily adjustable to any further data information if required. In the scope of transforming the XML format of the current database into RDF, we focused on the *lemon* core model that contains all basic elements necessary for a common dictionary entry. The layout is depicted in Figure 2.

As can be seen, the labels used to describe all lexicon elements differ slightly from those commonly used, e.g. "LexicalEntry" is also known as *headword*, *dictionary entry* or *lemma*. In order to understand the *lemon* vocabulary, all classes and properties are described within the corresponding *lemon*-RDF ontology file[8]. The *lemon* core is equipped with the necessary elements that are needed for a minimal dictionary entry. As an example serves the entry for "animal" in Figure 3 (McCrae et al 2010).

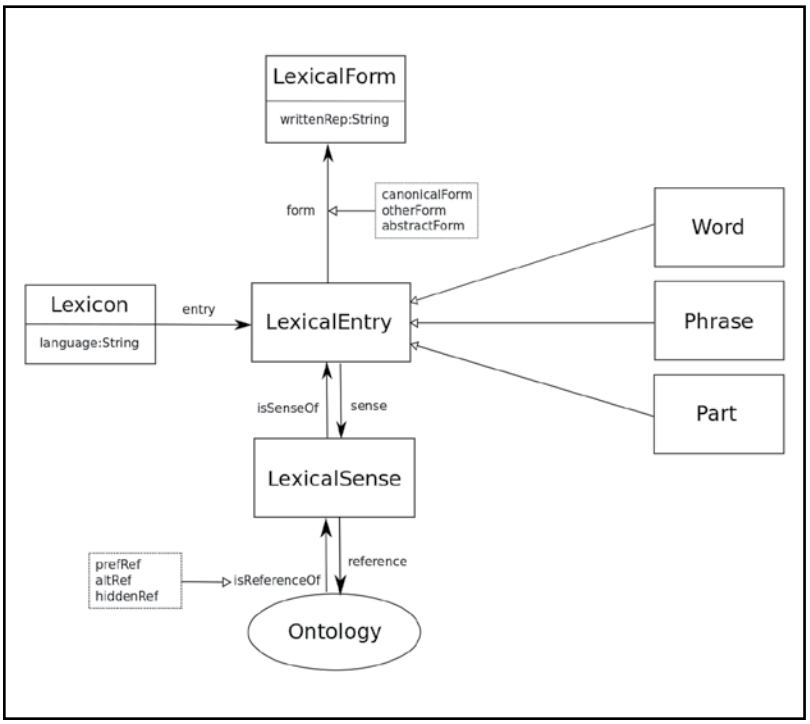


Figure 2: The *lemon* core path

```
@base <http://www.example.org/lexicon>
@prefix ontology: <http://www.example.org/ontology#>
@prefix lemon: <http://www.monnetproject.eu/lemon#>

:myLexicon a lemon:Lexicon ;
       lemon:language "en" ;
       lemon:entry :animal .

:animal a lemon:LexicalEntry ;
       lemon:form [ lemon:writtenRep "animal"@en ] ;
       lemon:sense [ lemon:reference ontology:animal ] .
```

Figure 3: *lemon*-RDF example for the lexical entry "animal"

What is encoded here are triples containing statements about the lexicon as such, the language of the lexical data, the orthographic or written representation of the lexical entry, and its meaning being a reference link to an external ontology. This conceptualization will be explained in section 4 in more detail. *Lemon* is designed to describe lexical content on different levels of granularity. The lexical entry, for instance, does not necessarily need to be a word. It can also be only a part of a word

3 http://linguistics.okfn.org/
4 http://linguistic-lod.org/llod-cloud/
5 http://wordnet-rdf.princeton.edu/
6 http://ld.panlex.org/rdf.html/
7 http://portal.eurosentiment.eu/home_resources?page=8/
8 http://lemon-model.net/lemon.rdf/, or visit http://lemon-model.net/lemon#/ for an HTML view of it.

```
<Entry hw="a" pos="letter" identifier="EN00000001">
  <DictionaryEntry identifier="DE00000001">
    <HeadwordBlock>
      <HeadwordCtn>
        <Headword>a</Headword>
      </HeadwordCtn>
      <HeadwordCtn>
        <Headword>A</Headword>
      </HeadwordCtn>
      <Pronunciation>a:</Pronunciation>
      <PartOfSpeech value="letter" />
      <GrammaticalGender value="neuter" />
    </HeadwordBlock>
    <SenseBlock>
      <SenseGrp identifier="SE00000001">
        <SidCtn identifier="SI00000001">
          <SenseIndicator>Buchstabe</SenseIndicator>
        </SidCtn>
        <Definition>erster Buchstabe des Alphabets</Definition>
        <ExampleCtn>
          <Example>Schreibt man das mit großem A / kleinem a?</Example>
        </ExampleCtn>
        <CompositionalPhraseCtn>
          <CompositionalPhrase>von A bis Z</CompositionalPhrase>
          <ExampleCtn>
            <Example>Das ist von A bis Z frei erfunden.</Example>
          </ExampleCtn>
        </CompositionalPhraseCtn>
      </SenseGrp>
    </SenseBlock>
  </DictionaryEntry>
</Entry>
```

Figure 4: Sample XML entry in the KD data

or a phrase. Likewise, next to the canonical orthographic written representations an abstract or other form can be given for the lexical entry. Just as classes can be extended by adding subclasses, also the properties stating the relations between them can be widened to the necessary level of description as desired. Hence, the *lemon* core model is open to any kind of structural adjustment, and even if the formal elements required are not stated in the extension of the core model an appropriate expansion can be undertaken with low effort, as will be shown in section 4.

Overall, lexicographic data modelled in *lemon* is concise and in RDF, so that it also allows for greater representation of linking between different sections of the lexicon (McCrae et al 2010).

Consequently, *lemon* offers the means to not only document lexicographic data but also to interconnect knowledge about the relations that hold between lexical entries of different linguistic description levels. Since it expresses all concepts necessary for lexical data documentation and beyond, it is powerful enough to serve as a foundation for the conversion of KD's XML data structure to RDF.

4. RDF transformation of KD's German dataset

To practically demonstrate the benefits of RDF, we converted sample data into *lemon*-RDF. KD supplied us with a small part of their German monolingual dictionary set, comprising around 5,000 entries. It came in valid XML files with a custom schema to represent the data, containing the entries in individual XML elements. Each entry element has a varying number of child elements representing additional data, such as the written representation of the entry, its pronunciation, associated meanings, examples of usage, semantic relations and part of speech labels. For visualization purposes an XSLT stylesheet is used to transform the data into HTML for user-friendly representation. Figure 4 shows an example of KD's XML.

As one of the RDF serializations RDF/XML is an XML format, the stylesheet could be modified to produce an RDF version of the dictionary. This procedure has the advantage that completeness of the transformation can be guaranteed, meaning that for every XML element, either an equivalent RDF resource could be established or its content would be expressed as a relation between two RDF resources. Figure 6 shows, analogously to the *lemon* core path in Figure 2, the XML elements of the KD data (on top, white background) that we mapped to *lemon* resources (below, grey background). Boxes represent resources in *lemon* and arrows represent relations between resources. These relations are expressed in XML as a relationship between a parent element and its child elements. For this reason, *lemon* relationships do not have a KD equivalent in the diagram. The RDF modelling thus explicates the semantic relationships that were implicit in the hierarchical structure of the XML data model.

Additional information was transformed using RDF properties of the LexInfo vocabulary (Cimiano et al 2011). These are common properties expressing lexical information, such as part of speech, gender or pronunciation. This step required some additional mapping. In the RDF model, information that can be categorized into a number of distinct

classes, such as *masculine*, *feminine* and *neuter* for grammatical gender, is generally expressed by assigning RDF resources to these classes. In classical dictionaries this information is expressed within standard strings. Thus, we mapped gender and part of speech information of the dataset to their respective resources in the LexInfo vocabulary.

During the transformation, gaps in the *lemon* model became apparent. The KD data contains compositional phrases (multiword units) for many senses, but there is no exact equivalent to express this relationship in *lemon*. So we established a new property, "hasCompositionalPhrase", and used it to link the senses to additional "CompositionalPhrase" resources. These phrase resources are, according to *lemon*, a subclass of LexicalEntries. Other gaps in the existing vocabularies concern properties to express semantic relations, such as hypernymy and synonymy. Again we established properties to express these relationships. This approach – of extending existing vocabularies with further properties adapted according to the expressivity of a new data source – is a standard procedure during RDF conversion. Thus, at the end of the transformation process, the added properties formed a small *lemon*/LexInfo extension, containing ten properties and ten classes. This extension vocabulary could now be published to aid the conversion of new lexicons into RDF and provide compatibility of these resources with KD's data, and vice versa. Figure 5 provides the *lemon* conversion of the original XML entry shown in Figure 4.

A persistent gap in the conversion is the missing *lemon:reference* property and the ensuing link to an external ontology. This link would disambiguate the meaning of the KD entry in an interoperable way. In addition to the common textual definition, the sense would point to a resource expressing its meaning, like the respective Wikipedia entry shown in Figure 1. This disambiguation could then be used to provide interoperability between disparate lexicons. Entries and senses in different lexicons could be compared by matching their links to external ontologies first, providing a way to find equivalent senses across lexicon borders. Such a mapping could be exploited for the enrichment of one lexicon with information from another, or for merging different types of dictionaries, such as picture with standard dictionaries. However, creating such a link automatically would imply automatic disambiguation of the senses of a lexical entry on the basis of a small textual description and few examples, which currently cannot be fulfilled reliably.

```
<http://kdictionaries.com/de/entry/DE00000001>
 a lemon:LexicalEntry ;
 lemon:canonicalForm [
  lemon:writtenRep "a, A" ;
  lexinfo:pronunciation "[a:]" ;
  a lemon:LexicalForm
 ] ;
 lemon:language "de" ;
 lexinfo:gender lexinfo:neuter ;
 lexinfo:partOfSpeech kd:letter ;
 lemon:sense <http://kdictionaries.com/de/sense/SE00000001> .

<http://kdictionaries.com/de/sense/SE00000001>
 a lemon:LexicalSense ;
 lemon:definition <http://kdictionaries.com/de/sense/SE00000001#def> ;
 lemon:example <http://kdictionaries.com/de/sense/SE00000001#ex1> .

<http://kdictionaries.com/de/sense/SE00000001#def>
 a lemon:SenseDefinition ;
 lemon:value "erster Buchstabe des Alphabets" ;
 kd:hasCompositionalPhrase <http://kdictionaries.com/de/compo/SE000000011> .

<http://kdictionaries.com/de/sense/SE00000001#ex1>
 a lemon:UsageExample ;
 lemon:value "Schreibt man das mit großem A / kleinem a?" .

<http://kdictionaries.com/de/compo/SE000000011>
 a kd:CompositionalPhrase ;
 lemon:canonicalForm [
  lemon:writtenRep "von A bis Z" ;
  a lemon:LexicalForm
```

Figure 5: *Lemon* version of the sample entry in Figure 4

Taking into account the possible advantages of such links for lexicography, it should be considered to add them manually in the process of lexical data creation.

5. Concluding remarks

The transformation of KD's German lexicographic XML data to a *lemon*-RDF lexicon resulted in the following outcomes. Firstly, the Linked Data principles were all fulfilled so that an integration of other RDF data is easily achievable. Secondly, all the lexical data elements are now identifiable via resource URIs and thus interlinkable with further relations within the dictionary and other external data. And thirdly, all XML elements could be mapped to an equivalent class or relation in the *lemon* model without decreasing the high quality of the data content. What is more, the whole *lemon* model that goes far beyond the *lemon* core comes with more fine-grained lexicographic conceptualizations that are

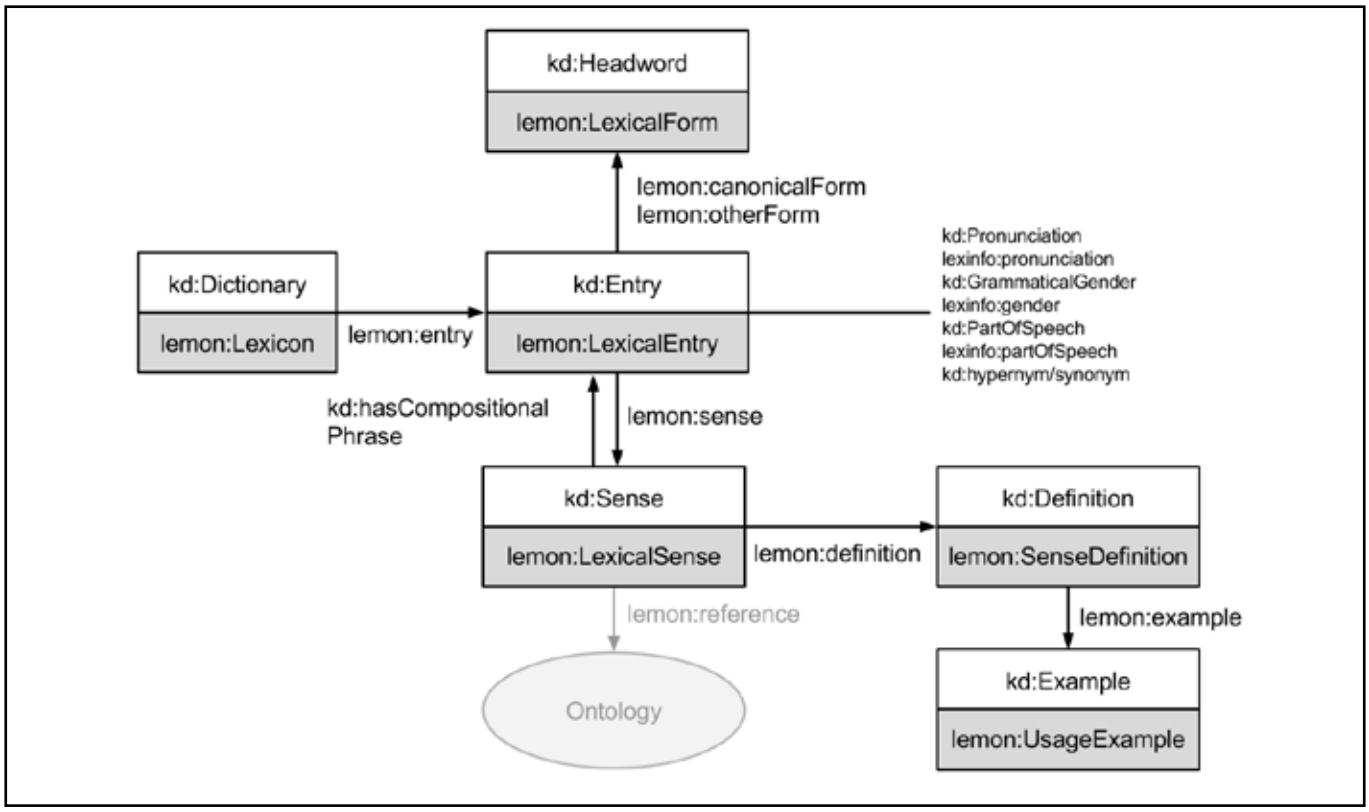


Figure 6: Mapping KD's XML elements and *lemon* resources (excluding the greyed out Ontology part)

worth considering in future data compilation or extension.

As a consequence, all possibilities of Linked Data in general can now be explored. With its underlying Linked Data format this dataset is equipped to express any considerable aspect of lexicography. Since the model is open for adaption, the complex and infinite nature of natural language can be documented to any desired extent. Existing open linguistic Linked Data resources such as lexicons of other languages, datasets including phonological, morphological or syntactic information, text corpora, and media content as well as all available Linked Data tools can be exploited and reused for specific lexical data compilations. In RDF all these usually isolated linguistic datasets become interoperable. It is such an interrelation of single pieces of data across various datasets without needing to make any change whatsoever in the data schema that will advance lexicography significantly in the future.

Acknowledgements

We would like to thank Dr. Sebastian Hellmann for giving advice and sharing his expertise during the compilation of the data conversion. Our gratitude also goes to Ilan Kernerman who revised this article and provided the great opportunity for this internship at K Dictionaries, thus making it possible to interconnect academic research with real industry data.

References

Berners-Lee, T. 2006. Linked Data – Design Issues. Retrieved 23 July 2014, http://w3.org/DesignIssues/LinkedData.html/.

Bizer, C., Heath, T., Ayers, D. and Raimond, Y. 2007. Interlinking Open Data on the Web. In: Demonstrations Track, 4th European Seman-tic Web Conference, Innsbruck, http://eswc2007/.org/.

Cimiano, P., Buitelaar, P., McCrae, J. and Sintek, M. 2011. Lexinfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9 (1): 29 (51) , http://sciencedirect.com/science/article/pii/S1570826810000892/.

McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez Pérez, A., Gracia, A. et al. 2010. *The lemon cookbook*, http://lemon-model.net/lemon-cookbook.pdf/.

McCrae, J., Spohr, D. and Cimiano, P. 2011. Linking lexical resources and ontologies on the semantic web with lemon. In *The Semantic Web: Research and Applications*. Heidelberg: Springer, 245-259.

Serasset, G. 2012. Dbnary: Wiktionary as an LMF based Multilingual RDF network. In: Proceedings of the 8th Language Resources and Evaluation Conference, Istanbul, (LREC'12), http://lrec-conf.org/lrec2012/.

Reflections on the concept of a scholarly dictionary

Dirk Kinable

The current age is frequently characterized as the era of information. Characteristics of this time are indeed the increasing dependence on information technology and the ever higher demands on information itself in terms of accuracy, completeness, interrelatedness, timeliness, etc. This development has strongly influenced dictionaries as containers/suppliers of lexical information. According to present-day standards of e-lexicography, the conception of dictionaries as merely linear, alphabetically-ordered sequences of self-contained entries has long since become outdated. Applications like the inter-connection of lemmas in more comprehensive semantic relationships such as hypernymy and hyponymy, or the introduction of the onomasiological search function from concept to corresponding lemmas, may suffice as examples here. For collections of dictionaries as well, the image of a linear arrangement on bookshelves is on the verge of becoming antiquated. Here, too, a three-dimensional virtual reality, as it were, is being developed by cross-connecting dictionaries by means of portals.

In other words, lexicography and dictionaries are undergoing a fundamental development at present. It is therefore at an opportune moment that the organization of European Cooperation in Science and Technology (COST) has established a platform named the European Network of e-Lexicography (ENel). ENeL aspires to play a stimulating role in bringing together lexicographers and linguists to reflect on building a comprehensive and modern Web portal for dictionaries of the European languages. The keyword therefore is widening the perspective. In line with this we like to avail ourselves of the opportunity to explore other means of communication, such as this newsletter, to draw the attention to one of the central issues that has to be dealt with in implementing the project.

At first instance, building such a portal implies reflection on its content. Even though a digital environment is always expandable, it is recommended to 'map' the area in advance. ENeL's own website describes its first aim in the following general terms: "to give users easier access to scholarly dictionaries and to bridge the gap between the general public and scholarly dictionaries". This entails the necessity to gain more insight into what is to be understood by a *scholarly dictionary*. Although the idiom occurs regularly in the professional literature, its definition is rarely at the centre of interest. Any definition attempt soon reveals that this concept is no exception to the general rule that defining is far from easy, which holds for both concrete and abstract nouns. Even for the former, which are generally easier to define, Landau states in his standard work *Dictionaries. The Art and Craft of Lexicography*: "There is no simple way to define precisely a complex arrangement of parts, however homely the object may appear to be. One obvious solution is not to define it precisely; but modern dictionary users expect scientifically precise, somewhat encyclopedic definitions" (2001: 167). This applies not in the least to abstract nouns, the complexity of which is usually more difficult to grasp. In the following, rare definition of *scholarly dictionary*, the shorter way according to Landau appears to have been followed. By means of only a *genus proximum* 'the next higher category' and two features, Hartmann and James (1998) give the following description: "a type of reference work compiled by a team of academics as part of a (usually long-term) research project, e.g. linguists working on a historical dictionary or dialect dictionary". In this definition, the distinctive semantic features are specifically related to the authors and to the research-related nature of the information offered. The previous definition marks the contours of the meaning of the idiom in a general way. Compared to this and consistent with the quotation from Landau above, the semantic features can be specified in a far more detailed way. This line was followed when the concept was the subject of a presentation at the Vienna meeting of ENeL last February. Participants had answered the call to send their views on the characteristics of a scholarly dictionary and their specifications fit in with the general definition above, concretizing it to a considerable degree. We summarize their views below.

Primarily, the scholarly dictionary was seen related most often to an academic environment, both on the production side and the demand side. The former was described as including 'academic editors or supervisors', 'academic publishing houses' and 'academic institutions', while among the ranks of the latter were counted 'linguistic researchers', the 'academic community', a

Dirk Kinable is one of the editors of the *Algemeen Nederlands Woordenboek*, an online scholarly dictionary of standard contemporary Dutch that is compiled in Leiden at the Institute of Dutch Lexicology, a bi-national institution of the Netherlands and Flanders (Belgium). His interests also comprise historical lexicography, to which he contributed previously as an editor of the comprehensive *Woordenboek der Nederlandsche Taal*, covering the period from 1500 to 1976. He has a PhD in Language and Literature from Leiden University, and also writes on lexicography and research subjects with a lexicographic focus on medieval texts.

dirk.kinable@inl.nl

This contribution is based on a presentation made at a meeting of the European Network of e-Lexicography (ENeL, WG1) held in Vienna on 11 February 2015, stemming from feedback by members to a call to define what a scholarly dictionary is.

http://elexicography.eu/

'scholarly audience' and 'users concerned with advanced linguistic studies and professionals on a fairly advanced linguistic level'. Indicative of this environment is also the notion that a scholarly dictionary is generally not produced on a commercial basis. The academic level of the authors and the primarily intended users accordingly implied high demands with respect to such dictionary's content. More specifically, the vocabulary had to be described on the empirical base of a processed corpus or of scholarly harvested examples, and several standards had to be met such as the pursuit of completeness in the scope of entries, comprehensiveness as to textual genre and language variation, and detailed information beyond the communicative support for reception and production purposes, all on an authoritative level. Regarding content, adequate room should also be reserved for encyclopedic information when relevant. Apart from the factors *author*, *content* and *user*, also the approach of the content was considered characteristic of a scholarly dictionary's profile. Based on the lexicographic standards of its time, analysis and description had to add new knowledge on the lexicon from a descriptive, not primarily prescriptive, perspective. This had to be realized using analytical definitions, scholarly terminology and the quotation of good dictionary examples as evidence, and the results had to be suitable for linguistic research. Finally, the last group of characteristics mentioned by respondents bore upon the contact with the user. Due to the often voluminous size of scholarly dictionaries, this is often established either digitally in the form of updates or in print by means of instalments. To convey the specialized information, the edition is often supported by a scholarly apparatus. In digital versions the user also often avails of functions giving access to many categories and also making the material collection searchable, and preferably expandable and linkable to other collections and tools.

Including this information according to Landau's previously-mentioned explicit description style, we can propose the following working definition of *scholarly dictionary*:

> knowledge-oriented dictionary compiled by (usually) academics to provide detailed word descriptions for the pursuit of lexical insight and research support according to the linguistic and lexicographic standards of their time, and traditionally designed with such main features as the pursuit of completeness with regard to the entries relevant to subject matters, a preference for analytic definitions, the use of an extensive corpus of observed discourse, the inclusion of documenting example sentences with bibliographic references, the availability of a scholarly apparatus like descriptions of method and project plan, a bibliography of sources, and, in digital specimens, the implementation of advanced search and application tools

The inclusion of an important definition element as "according to the linguistic and lexicographic standards of their time" indicates that a certain flexibility has been built into the definition. This chronologically relative point of view implies that not every scholarly dictionary can meet all the characteristics enumerated at any time. The tenor of the definition is in other words prototypical. The term is used here in the linguistic sense referring to the prototype theory. A prototype is the ideal example of a semantic category. The arrangement of a category may be conceived as follows: surrounding the core of the prototype are the instances of the category that share certain, but not all, of the characteristics of the prototype. Viewed from this angle the enumeration in the definition above is exemplary rather than exhaustive and certainly not meant as a list of necessary and sufficient characteristics. The latter is still often too narrow a way of characterizing definitions.

At present we carry out further research on this definitional issue with respect to the concept of a scholarly dictionary. A possible approach may consist of trying to specify what is at the centre of the category and resembles more the prototype and which dictionary types are more on the periphery.

Research of this kind is stimulated by the wealth of possibilities for discussion that are characteristic of the era of information mentioned in the introduction. Networks are not only devised between dictionaries, but the lexicographer as well is encouraged to consider his/her own position as a constituent part of a larger whole. While this development makes work more complex on the one hand, on the other hand it also makes communication easier both within the profession and outside it.

Comments and suggestions regarding the working definition above are welcome as cause for reflection (scholarlydictionary@inl.nl).

References

Hartmann, R.R.K. and James, G. 1998. *Dictionary of Lexicography*. London and New York: Routledge.

Landau, S.I. 2001. *Dictionaries. The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

The historical dictionary and the digital age: Steps of a transformation process

Nathalie Mederake

1. Media change and the dictionary

Dictionaries consist of many things at once, but first and foremost they offer a rich documentation basis for languages by providing a survey of their state via structured access to meanings and definitions. Historical dictionaries in particular reflect how the use of words has evolved. Admittedly, throughout their creation and editing process, printed dictionaries are affected by inconsistencies much more than might appear at first glance or than we would like them to be: as books, they seem to be most stable objects after all. Historical dictionaries, in particular, are end results of long-term academic projects and therefore predestined to show such inconsistencies. The reasons are manifold: an ever-changing project staff, an ever-updated corpus, the modification of metalexicographic standards during a decades-long working process, etc. After their publication in printed form, established historical dictionaries are often retro-digitized, which is when new discussions on the transformation of lexicographic data arise. Although this causes discomfort to traditional lexicographers, the changes brought about by the world of digital media have to be met in the context of historical lexicography too.

Media change has been experienced in various forms, for example, in the early days of cinematography artists began by filming theatre plays on stage before realizing that so much more was possible in cinema (and eventually TV, etc). Similarly, in the field of lexicography, there is no need for retro-digitized dictionaries to hide behind the traditional print product anymore: these dictionaries have a high potential to evolve into much wider-range information tools in the digital world.

This world is by all means a very complex one. As we can see, crowdsourced products or user-generated dictionaries are having quite an impact on the lexicographic landscape nowadays, which may also affect retro-digitized dictionaries. We expect the information we look up to be up-to-date, and do not bother to ask who made a change of what and when. Notwithstanding doubts and reservations, this development may be useful as future (historical) dictionaries won't exist as mere physical (and thus immutable) objects anymore. One could also argue that the storytelling of words is already happening entirely virtually. We are experiencing a golden age in terms of studying and appreciating words, and dictionaries are more accessible than ever, as Slate journalist Stefan Fatsis recently noted about Merriam-Webster[1].

Nowadays, lexicographers can benefit from the development of modern tools that expand possibilities in dealing with and exploring knowledge about language. Therefore, they have started to equip dictionaries with additional information, e.g. maps to show regional distribution or timelines to emphasize the occurrence of a word, i.e. anything that might be helpful for potential users. Indeed, the concept of a dictionary is being reinvented. However, it depends on the type of dictionary to what extent certain aspects can be taken on board in the creation of a multi-layered user interface. As the options are numerous, it can be challenging to make and actually *have a choice* that is not dictated by space availability. The new interpretation of existing media puts the lexicographer in the position of showing the public *all* collected information that became available thanks to decades-long work. However – while conciseness is valuable – constraints other than limitation on the number of characters with the paper model do exist. How is it possible to depict useful and relevant information for the user? Which access mechanisms should be implemented and what about the ongoing transformation in the first place?

2. Transformations within *Deutsches Wörterbuch*

The story of the *Deutsches Wörterbuch* (DWB) is a telltale of the transformation process in terms of the past, present and future of historical dictionaries. It is one of nearly two hundred years, and some turbulent times at that. Jacob Grimm and Wilhelm Grimm, the brothers who started to work on DWB in 1838, each had his own specific workflow and handling of the entry structure. Thus, inconsistencies existed from the very beginning. Also, the entire concept of the dictionary was influenced by numerous editors, who continued the

1 http://slate.com/articles/life/culture box/2015/01/merriam_webster_ dictionary_what_should_an_online_ dictionary_look_like.html/

Nathalie Mederake received her MA in German philology, history of law and sociology and PhD in German linguistics from the University of Göttingen. She has been editor and research associate of the *Deutsches Wörterbuch*'s revised edition at the Göttingen Academy of Sciences and Humanities since 2008, and is responsible for the production of manuscripts and prepress, leading development and digitization projects including the list of references, digital access to users and a cross-lingual vocabulary portal. Her other research topics concern the dynamics of Wikipedia entries from linguistic perspective and heading a task group on the accessibility and interoperability of retro-digitized dictionaries as part of the European Network for e-Lexicography.

nmedera@gwdg.de

Deutsches Wörterbuch
The *Deutsches Wörterbuch* (DWB, the German Dictionary) is the largest and most comprehensive dictionary of the German language in existence. Encompassing modern High German vocabulary from 1450 to the present, it also includes loanwords adopted from other languages and covers etymology, historical development of meaning, attested forms, synonyms, usage peculiarities and regional distribution within the German-speaking world. The dictionary's historical linguistic approach is illuminated by examples from primary source documents. The 32 volumes of DWB, published between 1854 and 1961, list more than 330,000 headwords in 67,000 print columns. Final steps to fully revise and update the letters A-F according to modern academic standards are underway at the Göttingen Academy of Sciences and Humanities and are due to be completed in 2016, with the new volumes published by S. Hirzel Verlag.

http://woerterbuchnetz.de/DWB/
http://www.uni-goettingen.de/de/118878.html/

The brothers Jacob and Wilhelm Grimm

Grimms' work, each leaving their own trace. The subject matter and objectives of DWB recurrently led to extensive discussions and workflow reorganization. Therefore, the first edition of this dictionary – if not the concept of a historical dictionary altogether – lived to see small but numerous transformations within roughly 120 years until its completion. The answers to what should be said and shown have been constantly subject to change in this first phase of its history.

The fact that the first edition of DWB has eventually turned into a 32-volume leviathan with a web interface, which now exists for over ten years, marks a second step in the transformation process. The volumes were interlinked and enhanced with detailed search options for the complete dictionary: apart from full-text searches, more complex searches within individual entries were made possible using an annotated database. The data was encoded in XML, meeting the guidelines of the ISO Text Encoding Initiative (TEI). Furthermore, the dictionary's data has become part of a network including other historical and dialect dictionaries of German[2] providing an extensive array of word-related information.

It goes without saying that this step in the transformation process is media-induced. Dictionary makers and computer linguists have adapted to the digital opportunities of lexicography. A new technology is, indeed, liberating. From a current perspective, it is more than obvious, though, that already very basic and self-evident amendments like the implementation of search options and integration into a Web interface reinvent the *dictionary system* we used to know (in print media). Regarding the consequences, Granger (2012, 10) states: "It shows that all facets of the field are undergoing a transformation so profound that the resulting tools bear little resemblance to the good old paper dictionary".

Another aspect of transformation comes into play with DWB's revision. Thoughts about a second or revised edition has already been underway when the paper version of the first edition was almost finished in the late 1950's. Through the years the dictionary had become a fundamental work regarding German philology and historical sciences. Back in the 1950's, however, only the latest volumes met eligible contemporary scientific standards. The first six volumes written by Jacob and Wilhelm Grimm in the 19th century clearly needed to be updated to fit into the dictionary's concept as a whole. The revised edition project, starting in the 1960's, was then justified by being a so-called "repair solution" to the first six volumes. Conceptually and lexicographically tied to the other dictionary volumes, the revision project was designed to make DWB a coherent oeuvre. However, the revised edition is *not* a supplement to the first edition, but presents an adjusted storytelling form for the history of words. It is based on a new and extensive corpus that adheres to high quality standards. Its entries are structured in a succinct way in terms of etymology, notes of explanatory matter and usage, precise definitions and quotes to mark noteworthy usages throughout the centuries. Entries from the revised edition differ substantially in concept from the very first ones written by the Grimm brothers.

For the time being, the revised edition only includes the letters A-F in the printed version. Nevertheless, the importance of the first and the revised editions as sources for the history of dictionary making in Germany is unrivalled. They form a fundamental work for all users with questions on the origins of German words. However, in the case of the revised edition, a possible digital concept that may go beyond the Web interface (which already exists for the first edition) has not been a topic of discussion yet. Considering such a concept might lead to new and important developments beyond "the good old paper dictionary", while at the same time focusing on potential trouble spots: accessibility and usability of words' stories in historical dictionaries.

3. In need of new paths

It is not only the advance of search engines or search optimization that is becoming a focal point. What we see today is that digital enhancements are not concerned any more only with displaying the content of a dictionary, but with emphasizing the role of the potential (or even actual) user. Despite the fact that users of a historical dictionary in general (and DWB in particular) have not yet been the subject of considerable studies, it seems to be a worthwhile objective of a digital interface to catch the user's interest for the dictionary's content and suggest means of handling the lexical information. Lexicographers (of DWB) must be aware of the fact that, as Lew (2011, 248) states in a survey on online dictionaries of English, "without proper guidance, users run the risk of getting lost in the riches". Therefore, the goal is to fathom new ways of access that were not possible for the printed book. Not least do different kinds of media-induced performances provide

2 http://woerterbuchnetz.de/

different ways of explanation. It may thus be useful to refer to the possible fields of inquiry and usage as well as to draw the user's attention to matters of microstructure. Consequently, new possibilities for DWB should deal with the options of having a facultative meta-comment on the one hand and a navigation aide on the other, both of which reach beyond the realization of the Web interface of the first edition. Tools like these could be sufficiently implemented after or during a digitization process (and count among the options that stem from the age of electronic dictionaries), but could not have been included in the framework of the printed volumes.

Figure 1 shows how the pictured steps in the transformation process come together and lead to a new direction. The given suggestions toy with the idea of an additional didactic-oriented concept as they try to guide the user through the manifold (and interrelated) information aspects of a dictionary entry. In view of the numerous and comprehensive entries of dictionaries, it should make use of the electronic environment and deliberately low access thresholds should be established, e.g. by reducing complexity and giving illustrative examples, as demonstrated in Figures 2 and 3 for a potential electronic version of the revised DWB. Now the challenge is to pursue this path and to pool lexicographic competences and technical resources in order to develop and offer efficient solutions. That also means overthinking the displayed content, which in the case of a historical dictionary can sometimes be very heterogeneous. As a possible consequence, it seems quite a feasible solution to visualize dictionary search options, which may eventually lead to a better understanding of the dictionary-structure.

To sum up, dictionaries in the digital age are different – and should differ – from traditional ones in print. Media change requires new ways of access and usability of lexicographic content. In particular, makers of historical dictionaries need to reconsider the conception of their products to not risk becoming antiquated. Transformation is important in order to keep a lexicographic product competitive and sustainable, and it leads to enhanced versions. At this stage, one cannot deny that in the field of retro-digitized dictionaries – although best practices for relevant additional information categories are still missing – concepts for interoperability and accessibility are becoming a pressing issue. Nevertheless, reinventing the dictionary is not only a means of technical expertise. In fact, understanding language in context of culture emphasizes the middleman role of the lexicographer. Dictionaries are institutions of general importance as well as trusted authorities. With that also comes an obligation to observe ongoing media changes and exploit their options for lexicography-based information tools.

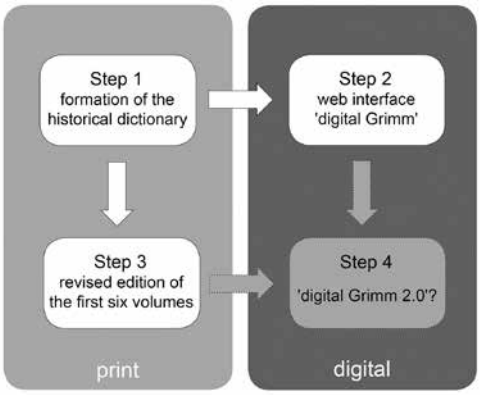


Figure 1: Steps in the transformation process of DWB

References

Granger, S. 2012. Introduction: Electronic lexicography – from challenge to opportunity. In Granger, S. and Paquot, M. (eds.). *Electronic Lexicography*. Oxford: Oxford University Press, 1-11.

Lew, R. 2011. Online Dictionaries of English. In Fuertes-Olivera, P.A. and Bergenholtz, H. (eds.). *e-lexicography*. London/New York: Continuum, 230-250.

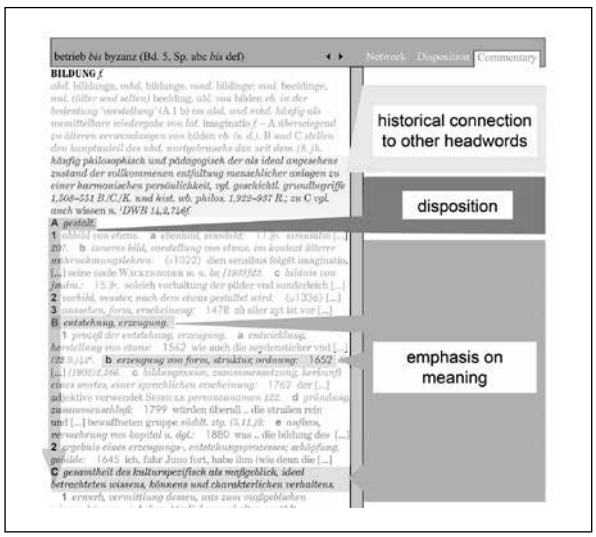


Figure 2: Possible focal points in the microstructure

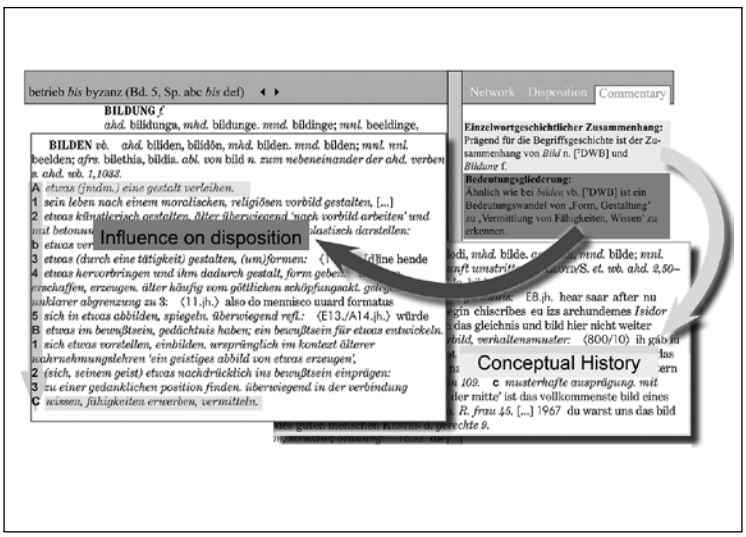


Figure 3: Pointing to interrelated information aspects

Recent developments in German lexicography

Alexander Geyken

Alexander Geyken works at the Berlin-Brandenburg Academy of Sciences and the Humanities since 1999, where he directs the long-term research project Digital Dictionary of the German Language (http://dwds.de/). He received his Ph.D. in Computational Linguistics at the University of Munich in 1998. His main research interests are computational lexicography, corpus linguistics, and the use of syntactic and semantic resources for the mining of large textual data.

geyken@bbaw.de

> The digital revolution is changing the way readers consume news and search for information. People are moving away from printed reference books and going online where, generally, they expect to get their information for free. (Press release by Chambers Harrap, 15 September 2009.)

This declaration by Chambers Harrap Publishers in 2009 was one of the rare public statements by a publishing house before closing its business. It points to the fact that the technological change is a decisive factor for the crisis of traditional dictionary production that has led to numerous staff reductions or insolvencies of dictionary publishers on an international level. Similarly, the national German dictionary market has been confronted with dramatic changes in the past years. Traditional dictionary publishers shrank dramatically (Duden, Langenscheidt) or even disappeared completely (Wahrig), and the largest academic dictionary, the *Deutsches Wörterbuch* (DWB, German Dictionary, by the Grimm brothers), compiled by the two Academies in Berlin and Göttingen, will cease its work in 2016. Except for Langenscheidt where the decline has a longer history, all this was announced to the public in one and the same year: 2013. The timing was pure coincidence since the momentous decisions were taken much earlier. To begin with, in 2009, the publishing house Langenscheidt with a tradition of more than 150 years of business in bilingual dictionaries, sold the prestigious Duden department to Cornelsen, a large company known for its text books in the field of education. This happened just a few months after Langenscheidt sold the Brockhaus encyclopedia to the Bertelsmann group. These sales were not the end of Langenscheidt's decline. In 2011 Langenscheidt also separated from Polyglott and in 2012 the 'adult education and school' department was taken over by its competitor Klett. Langenscheidt ended up as "Langenscheidt light"[1]. Instead of investing into a technological reorganization the company surprised the public by announcing a recent strategy shift that foresees a stronger focus on print products[2]. Another challenge for Langenscheidt was the advent of collaborative internet platforms. Based on large initial vocabularies donated by third parties and on a very active user community, the two most popular German dictionary internet platforms, Leo[3] and dict.cc[4], managed to compile very large translation databases of currently nine (Leo) and 26 (dict.cc) language pairs with German as the pivot language and became much more popular on the internet than Langenscheidt's rather traditional website. In addition Linguee[5] – a large database of paragraph-aligned translations where the translation quality of words or phrases in the sentence context can be rated by contributors – is becoming increasingly popular among users. Finally, Klett, who has invested very early in technological products, occupied a large market share. Under its brand PONS it has published a diligently curated set of bilingual dictionaries that have been made available free of charge on the internet since 2001 and continually extended to 13 language pairs at present with access to over 10 million words and phrases[6]. In 2009 Klett has also published a monolingual German dictionary that challenged Duden's spelling dictionary[7].

As a reaction to the stronger pressure from the competitors, Duden was also working on a more powerful internet platform. In April 2011 a website was launched where free access was given to the complete edition of Duden's flagship product, the *Großes Wörterbuch der Deutschen Sprache* (GWDS, Great Dictionary of the German Language, 1999[8]). GWDS is the largest dictionary of contemporary German. It was published as a print edition in 10 volumes with a total of 7,200 pages and 200,000 entries in 1999, and one year later in a CD-ROM version. In order to appreciate the full impact of the free internet version on the market strategy of Duden one has to remember that the CD-ROM was initially

1 http:// buchreport.de/nachrichten/verlage/verlage_nachricht/datum/2012/11/08/langenscheidt-light.htm/

2 http:// boersenblatt.net/949754/

3 http://dict.leo.org/

4 http://dict.cc/

5 http://linguee.de/

6 http://pons.com/

7 http://text-gold.de/praxistipps-fuer-online-redakteure/pons-online-woerterbuch-macht-dem-duden-konkurrenz-ein-praxistest/

8 http://duden.de/woerterbuch/

sold for an equivalent of 500 Euros. According to experts in the field, the entry of Duden into the market came too late. The sharp decrease in sales of the spelling dictionary, previously the no. 1 selling work of Duden, could not be counterbalanced. Only two years later, in 2013, Duden announced a dramatic reduction of staff from 190 to 30 employees[9]. Of course, plans for a complete revision of the GWDS were unrealistic under these conditions. Duden now concentrates on its one volume works, including the spelling dictionary, the grammar and the idiom dictionary.

Wahrig, the number two in the monolingual German dictionary market never managed to obtain a significant brand visibility on the internet. Being almost hidden among many other resources in Bertelsmann's large knowledge platform[10], it does not come as a surprise that Wahrig's dictionary was buried together with the Brockhaus encyclopedia: it was also in the year 2013 that Bertelsmann announced the discontinuation of their knowledge platform. The entire lexicographic staff was made redundant and since then, work on Wahrig's dictionary came to its end.

This crisis of lexicography in Germany is more than only an economic one. It is a well known fact among publishing houses that revenues of the large flagship dictionaries do not exceed their expenses. However, in the past these expenses could be cross-financed, for example by the revenues of print products derived from a flagship dictionary. This somewhat comfortable scenario stopped with the sharp decrease in sales of printed books and the triumph of the internet. Users do no longer rely on print products or, for that matter, on classical browser interfaces. Access via smartphones or tablets has become more and more common, and users are not willing to pay for these services. German dictionary producers have not been prepared for compiling tailored products for these new devices. The good old times when dictionaries were produced in three consecutive phases, i.e. planning the dictionary, compiling the dictionary and producing the dictionary (Landau, 1984), are definitively over. Nowadays dictionaries are not produced sequentially anymore but the various phases run in parallel or in cycles. Protagonists of dictionary production are no longer restricted to a team of lexicographers alone but prefer to work with an interdisciplinary team consisting of corpus linguists, computational linguists, IT specialists and lexicographers. Concerns that have to be addressed nowadays include the appropriateness of corpus compilation and its dynamic adaptation to new needs rather than the compilation of citation slips. Also, the automatic extraction of lexicographic information from corpora via statistics or machine learning techniques plays a major role in the dictionary production process today. Numerous papers on lexicography bear witness to these new challenges (e.g. Gouws 2011, Rundell 2012).

With the decreasing lexicographic staffs in publishing houses, further development of lexicography relies predominantly on institutional funding, namely the *Union der deutschen Akademien der Wissenschaften* (Union of German Academies) and the *Institut für Deutsche Sprache* (IDS, Institute for German Language). Both have a long tradition of compiling monolingual dictionaries. Currently there are more than 20 different dictionary projects funded by the Academies. However, the majority of these projects were started a long time ago with traditional methods and will run out of funding in the coming ten years. And given the above-mentioned technological changes it is not likely or desirable that new projects will start in the traditional way that is currently still typical of almost all these projects.

By contrast, there are currently two larger projects in Germany that recognize and implement the principles of the new era of e-lexicography. Both can hope for a sustainable funding: *elexiko* and DWDS.

Elexiko[11] started in 2000 as a long-term project of the IDS. The goal is to describe the German language from the end of the 1940's to the present in all its national variants. Practically, the focus in *elexiko* is set on the description since the 1990's corresponding to the text representation in the underlying corpus base, i.e. the DEREKO-corpus, a continually growing corpus of currently more than 25 billion words. A list of 300,000 lemmas has been selected for *elexiko*. Until the end of 2014, approximately 2,000 entries with high frequency in the corpora were manually edited by the lexicographers. Most lemmas consist of semi-automatically generated minimal articles with information about the spelling, the morphology and corpus examples. The hypertextual structure of the lexicon in *elexiko* played a role right from the beginning. Therefore particular emphasis is put on cross-referencing individual articles and providing links to external resources (Meyer 2014). The online presentation of *elexiko* is embedded into the

9 http://boersenblatt.net/543236/

10 http://wissen.de/

11 http://owid.de/wb/elexiko/start.html/

lexical information system OWID[12]. OWID grants access to a set of lexical modules including the lexicon of neologisms, the lexicon of paronyms and its core module *elexiko*.

The DWDS (*Digitales Wörterbuch der Deutschen Sprache*, Digital Dictionary of the German Language) began in 2007 as a long term academic project at the Berlin-Brandenburg Academy of Sciences and the Humanities (BBAW). The motivation to launch this project was threefold: firstly, there is no satisfactory account for the history of the German vocabulary since the end of the 19th century. Secondly, the *Grimmsches Wörterbuch* will remain outdated for the letters G-Z even after the completion of the second edition of the DWB that ends in 2016 after the completion of the letters A-F (by the way, 'Frucht' (fruit) was the last word compiled by the brothers Grimm). And thirdly, existing dictionaries at that time did not draw on large corpus data and computational methods right from the outset. Given the comparatively small project size of ten specialists, the goal of the DWDS project cannot be to compile a full historical dictionary. Instead it was decided to compile a large synchronic dictionary, to which diachronic modules could be added if such work will be funded in the future. More precisely, the aim of DWDS is to build an aggregated information system that draws on several complementary lexical resources, word statistics and corpora. The DWDS can make use of several lexical resources that are part of the heritage of the BBAW: the *Wörterbuch der Gegenwartssprache* (WDG), a synchronic dictionary of 4,800 pages with 90,000 keywords, compiled between 1961 and 1977, the *Etymologisches Wörterbuch des Deutschen* (Etymological Dictionary of German)) and the *Grimmsches Wörterbuch*. Moreover, some 60,000 dictionary articles were licensed from the Duden-GWDS for cases where the WDG articles are missing or outdated. The platform integrates an automatic collocation extractor and a good example finder (Didakowski and Geyken 2012, Didakowski et al 2012). Finally, the DWDS draws on large corpora with a size of 4 billion running words that cover the period between 1600 to the present. The results of this project are accessible under http://dwds.de/.

To sum up, the past decade has brought a shift in German lexicography away from private publishing houses to publicly funded institutions and collaborative internet platforms. The next years will show in what way the two institutional key players in Germany, namely the IDS and the Academies, are able to keep pace with the rapidly developing technology, thus being able to bring academic lexicographic knowledge to the public of the 21st century.

12 http://owid.de/

Acknowledgement
The author would like to thank Lothar Lemnitzer for his comments on an earlier draft of this manuscript.

References

Didakowski, J. and Geyken, A. 2012. From DWDS corpora to a German Word Profile – methodological problems and solutions. In *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie". Arbeiten zur Linguistik* 2/2012 (OPAL), Mannheim: Institut für Deutsche Sprache, 43–52.

Didakowski, J., Lemnitzer, L. and Geyken, A. 2012. Automatic example sentence extraction for a contemporary German dictionary. In Fjeld, R.V. and Torjusen, J.M. (eds.), *Proceedings of the XVth EURALEX Congress*. Oslo: University of Oslo, 343-349.

Gouws, R.H. 2011. Learning, Unlearning and Innovation in the Planning of Electronic Dictionaries. In Fuertes-Olivera, P.A. and Bergenholtz, H. (eds.), *e-Lexicography. The Internet, Digital Initiatives and Lexicography*. London: Continuum International Publishing Group, 17–29.

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

Meyer, P. 2014. Meta-computerlexikografische Bemerkungen zu Vernetzungen in XML-basierten Onlinewörterbüchern – am Beispiel von *elexiko*. In Abel A. and Lemnitzer, L. (eds.), *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern. 5. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie". Arbeiten zur Linguistik* 2/2014 (OPAL). Mannheim: Institut für deutsche Sprache.

Rundell, M. 2012. It works in practice but will it work in theory? The uneasy relationship between lexicography and matters theoretical (Hornby Lecture). In Fjeld, R.V. and Torjusen, J.M. (eds.), *Proceedings of the XVth EURALEX Congress*. Oslo: University of Oslo, 47–92.

A standardized wordlist and new spellchecker of Frisian

Anne Dykstra, Pieter Duijff, Frits van der Kuip and Hindrik Sijens

1. Introduction

In 2015 the Frisian Language Web saw the light of day online (FLW, http://taalweb.frl/). It has been developed by the lexicography department of the Fryske Akademy and consists of a newly created standardized wordlist of Frisian, an online spellchecker, a machine translator and a dictionary portal. These four applications together and individually offer a unique language tool to help Frisian native speakers and learners to write proper Frisian. The FLW is the first step towards a Frisian language platform that makes use of modern web technology. The standardized wordlist of Frisian is an important part of FLW and will not only be a firm base for future dictionaries but also the backbone of a special spellchecker. In this article we will give a brief overview of the compilation process of the standardized wordlist and how it is applied in the spellchecker.

2. Standard Frisian

Frisian has two major dialects and a number of smaller ones. Traditionally, variant pronunciations from minor dialects are not considered standard and thus are not included in dictionaries (anymore)[1]. Nevertheless there is not a fully-fledged standard yet. Today's standard is mainly based on the two major dialects: Clay Frisian and Wood Frisian. In practice the present dictionaries do not always make a clear and well-reasoned choice between the two. In daily practice it appears that the variant that in the dictionaries is given first, is and has always been regarded as the standard or preferred variant by users. Yet the dictionaries' ambivalent attitude towards standard Frisian causes some uncertainty with those very same users. It appeared that educators, authors and civil servants had a strong wish for more guidance as to the choice of preferred variants[2]. This doubt regarding correct usage is rooted in a lack of education and routine in writing Frisian. Frisian only became a compulsory school subject in the second half of the last century. In addition, because this obligation was applied only to primary schools and because written Frisian plays a minor role in daily life, most Frisians are not proficient in writing their own language. Language learners, as well as native speakers, are insecure in their language use, fearing to make mistakes. Even language professionals like journalists, editors, translators and novelists experience such problems.

Since the lack of a standard was felt to be a major obstacle in furthering the position of written Frisian, the Province of Friesland asked the Fryske Akademy to develop a standard for Frisian. It should be stressed here that the Provincial government does not have the legal means, nor the wish, to make the wordlist of Frisian a mandatory language standard. It is merely meant to be an aid to those who seek guidance when it comes to the choice of preferred variants.

3. Standard wordlist

The selection process for the wordlist has been guided by different criteria, though they have not been applied simultaneously:

(a) tradition

Though not fully-fledged, there is a certain standard Frisian language, which has been codified in Frisian dictionaries over the years. Though the dictionaries have most of the time not selected variants from peripheric dialects, they do not always make clear which variant from the two major dialects should be prioritized. Yet it has been common practice for a long time to consider the variants given first in the dictionaries as the preferred ones. That is why, for instance, Clay Frisian variants in *-om* or *-omme* (*rom* 'spacious, large, wide', *tomme* 'thumb') are given preference over Wood Frisian variants in *-ûm* of *-ûme* (*rûm*, *tûme*).

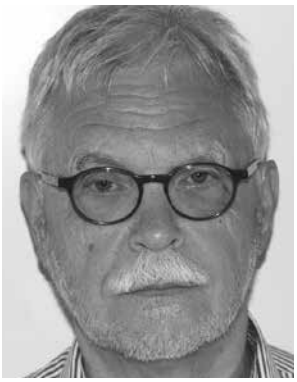
(b) distancing

One element in the standardization of a language can be distancing, which implies moving away from another (dominant) language. In the case of Frisian, that other language is Dutch. An illustrative example is the preferred variant *hiel* 'whole', though it is only used in a small part of Friesland. In by far the larger part of Frisians use *heel*, which is identical to the corresponding Dutch word and therefore disqualifies as a preferred variant.

Another example is the deletion of *e* in words like *kiste* 'box' and *mûtse* 'hat, cap'.

1 Only the scholarly *Wurdboek fan de Fryske taal* (http://gtb.inl.nl/) provides data from all dialects.

2 It must be said that there has also been resistance towards further standardization of Frisian, the main reason being fear of losing dialect variation.



Anne Dykstra has recently retired from the Fryske Akademy, where he has taken part in various Frisian language projects, including the scholarly *Wurdboek fan de Fryske taal* (Dictionary of the Frisian language). He was the editor the *Frysk-Ingelsk wurdboek* (Frisian-English Dictionary), has worked on the Frisian Language Database, and was project manager of the Frisian Language Web. Currently, he is editor of the *International Journal of Lexicography*.

dykstraanne@gmail.com

Pieter Duijff is lexicographer at the department of linguistics of the Fryske Akademy. He was editor of a legal dictionary (2000), co-editor of the *Wurdboek fan de Fryske taal / Woordenboek der Friese taal* (1984-2011), one of the editors-in-chief of the monolingual *Frysk Hânwurdboek*, and editor of word lists of Town Frisian dialects (1998) and modern Frisian (online, 2015). Since 2014 he is one of the editors of a new online Dutch-Frisian dictionary.

pduijff@fryske-akademy.nl

Frits van der Kuip has been a lexicographer at the Fryske Akademy since 1986. He was one of the editors of the 25-volume *Wurdboek fan de Fryske taal* (Dictionary of the Frisian Language) and one of the editors-in-chief of the *Frysk Hânwurdboek*. In 2003 he completed his Ph.D. thesis on 17th century proverbs. More recently he was involved in the Frisian orthography and spellchecker project, and at present he is preparing an online Dutch-Frisian dictionary.

fvdkuip@fryske-akademy.nl

Hindrik Sijens is a lexicographer at the Fryske Akademy, and an editor of the *Frysk Hânwurdboek* (monolingual Frisian dictionary).

hsijens@fryske-akademy.nl

In a part of the language area it is perfectly normal to say *kist* or *mûts*, yet these forms are not selected as preferred variants because of their similarity to Dutch.

(c) uniformity

The principle of uniformity can also determine the choice of the preferred variant. Words like *died* 'deed', *ried* 'advice', *sied* 'seed', *goed* 'good' or *paad* 'path' are often pronounced and written in Frisian without the final *d*. However, the *d* always emerges in inflected and plural forms and in compounds, both in spoken and in written language (*dieden*, *riedsgearkomste*). For that reason *died*, *ried*, etc. are written with the final *d*.

Another example is the Northern variant *baarne* 'burn', which up till now is the form that is given first in most dictionaries, but which has to give way to the Southern *brâne* because of the uniformity criterion: *brân* 'fire', *brâne* 'burn', *brânfersekering* 'fire insurance'.

(d) common usage

Frequency is a natural selection criterion when compiling a list of preferred variants. Highly frequent adverbs like *eigentlik* 'in fact, actually' and *eins* (idem) were included in the wordlist, while the less frequent *eigenlik*, *einlik* and *eink* were not.

(e) etymology

Sometimes the spelling reflects a less careful pronunciation: *abslút* 'absolutely', *bibleteek* 'library', *bommedearje* 'to bomb'. The wordlist only contains the full forms: *absolút*, *biblioteek*, *bombardearje*. Exceptions to the etymological principle are lexicalized forms such as *knyn* 'rabbit' and *plysje* 'police' (instead of *konyn* and *polysje*) and word pairs having different meanings like *krupsje* 'a/any disease' and *korrupsje* 'corruption'.

4. Spellchecker

It stands to reason that the standardized wordlist is the core of the spellchecker that is part of FLW. As explained, the wordlist is meant to be a practical tool to help those who want to use preferred variants in their written text. The term 'preferred variant' entails that variants that have not been included in the wordlist are not to be considered wrong. To further facilitate use, the wordlist has been incorporated into a database underlying the spellchecker in which non-standard words are linked to preferred variants. This feature makes it possible for the user to focus on the use of the preferred variant.

The following example will demonstrate how the spellchecker works:

Hy skamme sich faor syn tiim.
'He was ashamed of his team.'

The spellchecker will return the sentence with three underlined suggestions, each in a different colour:

Hy skamme sich faor syn tiim.

Red [*faor*] indicates a typo, green [*sich*] a Dutchism, and blue [tiim] a variant. The user can decide what he or she wants the spellchecker to do. For instance, to avoid seeing variants underlined the user can tick the blue box and get the following spellchecked text:

Hy skamme sich faor syn tiim.

Faor is a typo of *foar*, which is a mistake that any spellchecker would find. Since the wordlist usually lists English words in their original spelling, the user is advised to opt for *team*. A regular spellchecker might have come up with this suggestion as well.

Sich is quite a different case, clicking it renders the following list of alternatives:

har
harren
him
himsels

The correct Frisian word for the Dutchism *sich* in this case is *him*. The user will be able to replace *sich* by *him* by a simple click. This is where the FLW spellchecker stands out from regular ones, for which it would have been impossible to go from *sich* to *him*.

5. Conclusion

The Frisian Language Web has many functions, not the least of which is language maintenance. The first ever standardized wordlist of Frisian is supposed to encourage Frisians to write their native language with more confidence. When writers apply the spellchecker, they do not have to consult the wordlist directly, the spellchecker will make clear what the preferred variant of a particular non-standard form is. It is important to note that users have a choice in what they want the spellchecker to do. The Fryske Akademy will be happy to share its spellchecker technology with other (small) languages. Those interested are welcome to contact us.

Fryske Akademy was founded in 1938 in Leeuwarden, the capital of the Province of Friesland in the Netherlands, and concentrates on fundamental and applied academic research into the Frisian language, culture and society. While the results of its activities are primarily of academic interest, they have significant social meaning and relevance to Friesland and beyond, particularly in the context of minority and "small" languages. The academy's major lexicographic accomplishment is the *Woordenboek der Friese Taal* (Dictionary of the Frisian Language, http://gtb.inl.nl/). Its Linguistics Department has developed the Frisian Language Database (http://fryske-akademy.nl/tdb/) and the Frisian Language Web (http://taalweb.frl/). Currently, a Dutch-Frisian online dictionary is in preparation.
http://fryske-akademy.nl/

Howard Jackson, ed. *The Bloomsbury Companion to Lexicography*

The Bloomsbury Companion to Lexicography presents a broad overview of contemporary research and trends in lexicography. It contains some twenty substantive chapters by eminent scholars in their fields, and includes additional reference materials. Although the *meta-* prefix is not attached to *lexicography* in the book's title, sometimes the frame of *metalexicography* is helpful in emphasizing the distinction which is repeatedly stated in the text: the *Companion* is meant to accompany not the practical craft of dictionary-making, but the theoretical work of lexicographical criticism and dictionary research. In day-to-day life, the two disciplines are probably not truly separable, but given the number of manuals intended for practitioners, a theoretically-oriented introductory compendium is a promising prospect.

Chapter Overview

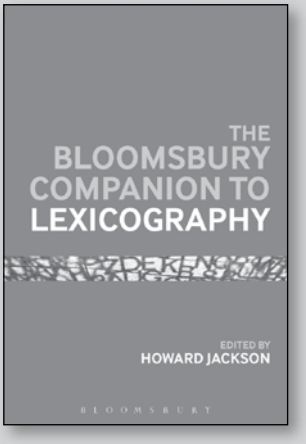
The introduction explains that the *Companion* "is aimed primarily at students of lexicography who are proposing to undertake research in one of the areas covered by 'lexicography'." It "aims to give a broad overview of the discipline, dealing with the main trends and issues in the contemporary study of lexicography" (1). Lexicography is a big enough field that reasonable people may have differing opinions about all sorts of questions, large and small. The *Companion* makes no attempt to offer a unified point of view, but puts forth a menu of perspectives from which its readers may launch or expand their own research.

After the editor's introduction, the late Paul Bogaards' "A History of Research in Lexicography" gives a historical overview. Beginning in the mid-twentieth century, Bogaards covers the development of studies in lexicographical history, criticism, and typologies; dictionary macrostructure and microstructure, usage, and corpus methods. The chapter runs for ten pages of text, plus three full pages of references to works in English, French, and German. This breadth of references is a reassuring opener for anyone who suspected that a *Bloomsbury Companion* might prove Anglo- or English-centric, and its diversity of perspective is further broadened throughout the book. Bogaards' typology of lexicographical scholarship does not completely correspond to the one presented by the *Companion* as a whole: for example, Bogaards (by way of Béjoint (2010) mentions research in encyclopedic, children's, and onomasiological dictionaries. These are indeed fruitful areas of lexicographical study, but they are not discussed in the *Companion*.

Next, Lars Trap-Jensen's "Researching Lexicographical Practice" is a reasonable textbook account of major topics in lexicographical practice: conceptualization, design, semantic description, dictionary writing systems, interfaces, and the specter of a future where all reference is mediated by Google (a major topic indeed!). Compared with Bogaards' chapter and many of the others, this chapter is light on connections to ongoing research. I was puzzled to find no references to any of the detailed manuals to lexicographical practice. Not that the reader likely needs to be told that they exist (nine are listed in the book's Annotated Bibliography) but a connection with these manuals could have provided an understanding of areas of relative consensus and divergence. Trap-Jensen begins by asserting a focus on monolingual native-speaker dictionaries, but the overview is unspecific enough that it can apply just as well to bi- and multilingual resources.

Kaoru Akasu's "Methods in Dictionary Criticism" describes the team-review methods used by the Iwasaki Linguistic Circle. As described by Akasu, these methods appear to be an excellent way to perform an intensive analysis of a dictionary by combining multiple reviewers' expertise; Akasu argues convincingly for rigorous procedure and comparative reviewing. Along the way, Akasu points to an interesting challenge of dictionary criticism. It is vanishingly rare for dictionaries to document their editorial practices and style guide in any detail (Sinclair (1987) being a cherished exception). As a result, critics must reverse-engineer a dictionary's intent in order to guess what its goals were, and thence evaluate its success.

Hilary Nesi's "Researching Users and Uses of Dictionaries" is a thorough overview of usage studies to date, with references to a broad array of user studies. Potential areas of study, and potential approaches, are so varied that it is not feasible to exhaustively survey user research in the space allotted. Nevertheless Nesi's account provides a clear, generously cited map to the enormous territory currently covered, categorizing existing work by its focus on user types, usage contexts, user preferences, or usage strategies.



The Bloomsbury Companion to Lexicography
Howard Jackson, ed.
London: Bloomsbury Publishers. 2013
ISBN (hardback) 9781441145970
ISBN (paperback) 9781474237376
ISBN (epub) 9781441144140

The Slownik Norweski website was launched by DTC PRO in 2012 and serves mainly Polish learners of Norwegian. It features KD's GLOBAL Polish/Norwegian dictionaries, and offers also specialized dictionaries along with study and test materials for free and to subscribers.

http://slownik-norweski.pl/

Published in cooperation with K DICTIONARIES

Adam Kilgarriff's "Using Corpora as Data Sources for Dictionaries" draws heavily from the author's own work. This is unavoidable, since the Sketch Engine (Kilgarriff et al 2004) has been preeminent in spurring a technological shift that he describes: "from [a methodology] where the technology merely supported the corpus-analysis process, to one where it pro-actively identified what was likely to be interesting and directed the lexicographer's attention to it" (85). Beyond his own work, Kilgarriff also cites research by others, some of which was new to me despite my own focus on corpora. The chapter could very well serve as a practical introduction for new corpus lexicographers. Its fundamentally future-looking orientation leaves an impression that this chapter will stand the test of time quite well: it describes practices that will surely continue to develop and gain ground.

Verónica Pastor and Amparo Alcina's "Researching the Use of Electronic Dictionaries" is an expanded version of their 2010 IJL paper (Pastor and Alcina 2010) and presents a classification of electronic-dictionary search methods. It is a descriptive study of existing facilities, rather than a speculative wish-list of potential new features. Although it describes the present state of an art which is constantly developing, Pastor and Alcina's paradigm is quite likely to accommodate yet-unforeseen dictionary features. As a result, much like the previous chapter, this one, too, ought to remain useful long past its publishing date.

John Considine's "Researching Historical Lexicography and Etymology" is exemplary in conveying a specialist's thorough survey of the subject matter, supported with extensive references to resources for deeper understanding. Although the OED is obviously the 137-pound gorilla of historical dictionaries, Considine does not neglect historical dictionaries in many other languages, and he makes the case for creating even more.

Amy Chi's "Researching Pedagogical Lexicography" is another outstanding contribution; right around the midpoint of the book, Chi frames her subject in ways that reach far beyond the chapter's nominally pedagogical focus. For example, Chi describes user studies showing that "most users exploit only a narrow range of dictionary items in their consultations, focusing predominantly on meanings," while ignoring guidance on abstractions like syntactic patterns or count/mass distinctions. She suggests that future studies ought to ask whether "English language curriculum and/or teaching…[has] promoted this narrow usage" (180). This question is about usage at least as much as it is about teaching. Any user is likely to bring established habits with them when they use new dictionaries, and to look only for the information they are accustomed to finding. As we create resources with far more data behind them, our efforts may be squandered if users never explore deeply enough to benefit from novel dictionary developments.

Shigeru Yamada's "Monolingual Learners' Dictionaries – Where Now?" details the history, present and future of learner dictionaries, with the author's characteristic comprehensiveness. Although I do not always agree with the way that Yamada evaluates individual features or frames particular dichotomies in this chapter, I heartily agree with his ultimate conclusions and vision of a possible future. The conclusion draws from Yamada (2011) to show a 'dismembered' LDOCE entry in an improved electronic layout, in which I saw very promising implications for the underlying data. Most of the other contributions use endnotes only for references or supplementary information, but Yamada's enjoyable endnotes sometimes convey bolder positions than the ones he takes in the main text.

Arleta Adamska-Sałaciak's "Issues in Compiling Bilingual Dictionaries" digs deep into the most challenging and interesting issues in bilingual lexicography. Although the chapter purports to focus largely on print dictionaries, much of the discussion – audience, scope, directionality, resource planning, microstructure, data sources, and challenges of inter-cultural conceptual equivalence in general – is highly illuminating for both print and electronic work, and much of it arguably for monolingual work as well.

The next two chapters, Danie J. Prinsloo's, "Issues in Compiling Dictionaries for African Languages" and Inge Zwitserlood et al's, "Issues in Sign Language Lexicography" are extraordinarily welcome and eye-opening. These two chapters are the deepest explorations of these challenging-but-important topics that I have seen in a single-volume lexicography book. Typically, if these topics are addressed in generalist books, it is with passing citations to other sources for specialists in those languages. Their inclusion here gives them a rightful position at the core of things that lexicographers should be concerned with, rather than as fringe topics for specialists. They merit broader attention not exactly for the languages in themselves: a monolingual lexicographer focuses on a single language, after all. But lessons learned from other languages may enrich everyone else's work, and these languages have some very

challenging things to tell us about unsolved lexicographical problems.

Robert Lew's "Identifying, Ordering and Defining Senses" is very satisfying and on point. It touches all the urgent and relevant issues of lexical and semantic analysis, frames interesting problems in an engaging way, and, like almost all the chapters, has excellent references. It is not hard to imagine future-dictionary scenarios where 'ordering' of senses is not a crucial task – a contextually-disambiguated word lookup doesn't need to tell you about senses *b* or *c* if it knows that you've come for sense *d* – but even in that future, Lew's treatment of structure in sense-enumerative semantics is excellent.

Tadeusz Piotrowski's "A Theory of Lexicography – Is There One?" is concerned with a question many of us have heard before. The fantastically incisive thing here is that Piotrowski frames the question in a way that allows for a positive answer, instead of the traditional rejection or minimization of the question. "Lexicography produces dictionaries, not theories, while metalexicography does not produce dictionaries but general statements about them. Accordingly, metalexicography can be a science, while lexicography is not" (309). Piotrowski notes that existing, practically-oriented lexicographical theories "have a strong prescriptive bent," as distinguished from scientific theories, which aim at description and prediction. The chapter does not go so far as to actually formulate a theory of lexicography, but it demarcates a space where such a more general lexicographical theory would be meaningful for critics and practitioners alike.

Piotrowski is a tough act to follow. In the next chapter, "e-lexicography: The Continuing Challenge of Applying New Technology to Dictionary Making", Pedro A. Fuertes-Olivera argues that many online dictionaries are really just print dictionaries recapitulated on a screen, and that only a few dictionaries are really conceived from scratch in the electronic domain. Although I fully agree with that evaluation, I find that when Fuertes-Olivera gets into specifics, much of his discussion of e-lexicography still feels conceptually grounded in print dictionaries. I fear Fuertes-Olivera takes insufficient account for the ways that computational intermediation can more deeply change the products and processes of lexicographical consultation.

Charlotte Brewer's "The Future of Historical Dictionaries, with Special Reference to the Online OED and Thesaurus" addresses the insights made possible by digitization. Fast search over massive resources can expose both the history of a language and the history of its lexicography. Brewer speaks from experience about the other edge of this sword, whereby digital editorial workflows can erase parts of this history and distort the historical record of the dictionary at a keystroke: another efficiency that was not possible in print.

On his way to describing "The Future of Dictionaries, Dictionaries of the Future", Sandro Nielsen takes a moment to define what he means by dictionary. This definition is a useful exercise for anyone talking about the future of dictionaries, since the future of 'hardbound printed books with speckled edges and thumb indexes' is very different from the future of 'semantic tools to aid linguistic production and reception.' Unfortunately Nielsen's proposed definition moves the goalposts just a few meters: "dictionaries are reference tools made up of several surface features" (356). These surface features are subsequently described, but the meaning of 'reference tool' is not. I am not feigning ignorance when I say I don't know what precisely a 'reference tool' might be. The relevance of the question for this review is that some of what Nielsen discusses, around different interfaces to electronically-mediated information, is not unique to dictionaries and can potentially enhance any information channel previously mediated through print. Enhancements like voice search, video results, and the intriguing "three-dimensional form, including holograms" (368) could enrich newspapers and gasoline pumps just as well as dictionaries, but newspapers are not prototypical reference tools. Lexicography has some unique features that Nielsen does not consider, but many of them are covered elsewhere in the *Companion*; in return, Nielsen offers several useful handles on contemporary problems that are not covered elsewhere in the book. His discussion of "information costs" (369) is a good frame for the distinctive tradeoffs of lexical reference, where a user's main task can be assumed *not* to be consulting the dictionary, but instead learning an answer so they can get back to what they were doing. Nielsen's conclusion that "dictionaries are in a transitional phase from the manufacturing sector into the service sector" (370) is also quite well taken.

The three remaining sections are reference material. Reinhard Hartmann's catalog of "Resources" is a general overview of societies, corpora, journals, and the like, with brief expository descriptions to accompany each section. In a book that appears otherwise carefully copyedited, this chapter has unusual inconsistency

Apps of the semi-bilingual PASSWORD English Dictionary for Android, iOS and Mac OS, developed by Paragon Software:

- English Afrikaans
- English Arabic
- English Bulgarian
- English Chinese Simplified
- English Chinese Traditional
- English Croatian
- English Czech
- English Danish
- English Dutch
- English Estonian
- English Farsi
- English Finnish
- English French
- English German
- English Greek
- English Hebrew
- English Hindi
- English Hungarian
- English Icelandic
- English Indonesian
- English Italian
- English Japanese
- English Korean
- English Latvian
- English Lithuanian
- English Malay
- English Norwegian
- English Polish
- English Portuguese Brazil
- English Portuguese Portugal
- English Romanian
- English Russian
- English Serbian
- English Slovak
- English Slovene
- English Spanish
- English Swedish
- English Thai
- English Turkish
- English Ukrainian
- English Urdu
- English Vietnamese

https://play.google.com/store/apps/details?id=com.kdictionaries.container/

https://itunes.apple.com/app/password-semi-bilingual-english/id672144357/

in the formatting of URLs and names in its informational tables, but this is not an obstacle to getting the useful information out of them.

Barbara Ann Kipfer's "Glossary of Lexicographic Terms" includes terms from lexicography, publishing, and parts of linguistics relevant to lexicography. Like all the other chapters, it represents its author's own viewpoint. In the case of a glossary this means that some of its terms are not used in the *Companion* itself (*back-formation*; *bogey*; *density*; *Sprachgefühl*). Considering that Kipfer's (1984) *Workbook on Lexicography* included Jennifer Robinson's (1983) glossary of lexicographical terminology, it was interesting to compare the two approaches some 30 years apart. The two glossaries have some overlap in their headword selection, and sometimes in the substance of the definitions and sense divisions. Robinson's glossary has example sentences taken from a reading list of lexicographical writing, and frequently uses *index entries*, variant headwords that are simply cross-references to a fully-defined synonym (a term I couldn't remember but that I found in Kipfer's glossary). Kipfer does away with illustrative examples and also with index entries, instead repeating definition content at variant headwords with small amounts of supplemental information at one entry or the other. I find the current approach more user-friendly and well-suited to the *Companion*.

Howard Jackson's "Annotated Bibliography" concludes the book. The whole book may be seen, in a certain light, as an annotated bibliography to the consistently great references sections of its individual chapters, and Jackson's bibliography presents a different kind of general overview. It turns out that both bibliographical streams are needed for the fullest picture of the available work. For example, Jackson's annotated bibliography lists a practical manual for field workers in indigenous languages (Bartholomew and Schoenhals 1983) but lists no work focusing on either any African languages or Sign languages. This is an honest reflection of their neglected place in mainstream lexicographical thinking, even as this book has done well in bringing them greater attention.

Evaluation

Many of the chapters address some of the same sub-topics from different perspectives and in varying levels of detail, and this to the book's great credit and advantage. Unfortunately, the *Companion* has vanishingly few cross-references between chapters. As a result, it is not possible to know beforehand that, say, corpus-driven headword selection, discussed in general terms by Trap-Jensen on pages 40-41, is explored more concretely by Kilgarriff on pages 79-83. Nor is the index much help: it runs for only two pages of this 420-page book, and lists Kilgarriff's headword-selection pages under 'headword' but not at 'lemma selection', yet conflates twelve references to 'headword' without any subcategorization (e.g. between headword selection and headwords as part of access structure).

The limited coordination among the authors also leads to a certain unevenness between chapters. Again on the subject of headword selection, Prinsloo concludes a stunning section about lemmatization challenges in Bantu languages (246) by mentioning frequency cutoffs as a potential method for lemma selection. References to either Kilgarriff or Trap-Jensen would have been helpful here, but it would have been most interesting if Prinsloo had been able to engage with their positions, and to discuss the consequences of frequency cutoffs from the perspective of Bantu-family language users and lexicographers.

The stand-alone chapters and skimpy index create an obligation to read the whole book in order to be sure that one has read everything that its contributors have to say about a subject. A professor who wished to assign selected readings from the *Companion* might need to assign two or more chapters to get full coverage of various issues that span subdisciplines. To be clear: it is a great strength of the book that it contains these complementary perspectives; it is regrettable only that the connections are not more accessible. The book is of manageable length and often illuminating, so 'reading the whole book' is in no way a burden.

It is clear from the start that the *Companion* is deliberately latitudinarian, permitting leading scholars to introduce their specialties and to describe their cutting-edge research on their own terms. The book covers far more intellectual territory than the average researcher could hope to have at the front of their mind all the time; making it available at arm's reach is surely part of what qualifies it as a *companion* rather than an *introduction*.

Is it necessary to distinguish between lexicography and metalexicography? We say that one field is concerned with practice and the other with theory, but the two are never truly separable. As Chi suggests in her chapter, people use dictionaries in certain ways because lexicographers have historically made dictionaries in certain ways. The study of users is therefore also the study of lexicographers. Beyond

being cultural artifacts, dictionaries are technological artifacts in a major transition, as all of the book's contributors would surely agree. We do not yet know what will be the end point of this transition, but Piotrowski makes me feel that theory can be our guide.

Piotrowski says that a theory of lexicography is not like a scientific theory, because it cannot successfully predict unobserved phenomena. Although this is probably true in absolute terms, it is interesting to consider what kinds of things theoretical lexicography can at least infer, if not predict outright. Akasu and the Iwasaki Linguistic Circle can guess at a lexicographical team's underlying principles based only on their finished dictionary: finding the proof of the pudding in the eating. Trap-Jensen's chapter somewhat frustratingly describes an array of possible lexicographical practices without much accounting for how people choose among them in practice. Yet every working lexicographer makes complicated choices in practice every day, and these choices are motivated by some kind of theoretical orientation, even if it is largely implicit or unexamined convention.

The chapters on Sign and African languages, where aspects of traditional practice are impossible, throw stark contrasts that help to reveal the shadow theories behind mainstream lexicography. As we work to document under-resourced languages at a level of quality that approaches that of resource-rich languages like English, we encounter features that cannot fit into the familiar paradigms of lexicography for Indo-European languages. It may turn out that a solution to a distinctively Xhosa or ASL challenge – be it lemmatization, gestural search, or semantic compositionality – could be usefully applied to lexicography of familiar western languages and enrich the entire lexicographic discipline, in both theory and practice.

Conclusion

Enough theorizing. These thoughts have been spurred by the *Companion*, but no doubt other readers will seize on different aspects and reach their own conclusions. The important thing is that I expect this book will be a strong catalyst for lexicographers of every stripe. It presents contemporary research, summarized for review at a readable scale, with the happy outcome that both specialists and new researchers may reach a clearly contextualized understanding of the trajectories of subfields other than their own.

References

Bartholomew, D.A. and Schoenhals, L.C. 1983. *Bilingual dictionaries for indigenous languages*. Mexico: Summer Institute of Linguistics.

Béjoint, H. 2010. *The Lexicography of English. From Origins to Present*. Oxford: Oxford University Press.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. 2004. The Sketch Engine. *Proceedings of EURALEX 2004, Lorient*, 105-116.

Kipfer, B.A. 1984. *Workbook on Lexicography*. Exeter Linguistic Studies, Vol. 8. Exeter: University of Exeter.

Pastor, V. and Amparo A. 2010. Search Techniques in Electronic Dictionaries: A Classification for Translators. *International Journal of Lexicography* 23 (3): 307-354.

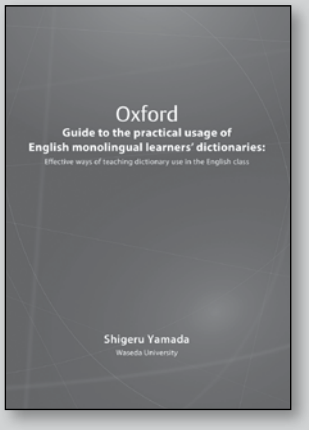
Robinson, J. 1983. A Glossary of Contemporary English Lexicographical Terminology. *Dictionaries* 5: 76-114.

Sinclair, J.M. (ed). 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins ELT.

Yamada, S. 2010. Layout matters. Paper presented at the Dictionary Society of North America XVIII Biennial Meeting, McGill University, Montreal, 8-11 June.

Orion Montoya
MDCCLV
orion@mdcclv.com

Shigeru Yamada. Oxford Guide to the practical usage of English monolingual learners' dictionaries: Effective ways of teaching dictionary use in the English class



Oxford Guide to the practical usage of English monolingual learners' dictionaries: Effective ways of teaching dictionary use in the English class
Shigeru Yamada
Tokyo: Oxford University Press. 2014
http://oupjapan.co.jp/teachers/resources/oup_guide_to_dictionary_use_2014_E.pdf/

This short guide combines a general introduction to dictionaries and their receptive and productive functions (pp1-10), with a set of classroom activities in dictionary use, all designed to demonstrate the value of using a dictionary rather than any of the more alluring alternatives on offer (pp10-16).

There is much to like here, and the guide is full of sensible advice. Yamada acknowledges that, for most learners, a monolingual dictionary of their target language can look intimidating, and he recognises that "some may find the task of consulting a dictionary troublesome". In the past, learners have generally preferred bilingual dictionaries to monolinguals, and now the Web offers a range of other options too, such as automatic translation sites or forums like Word Reference (http://forum.wordreference.com/). All of this creates tough competition for the traditional monolingual learner's dictionary (MLD). But Yamada makes a spirited case for the benefits of MLDs, seeing dictionary use as "a learning opportunity".

This gets to the heart of the matter, and it's useful to think about this in terms of learners' short-term and longer-term goals. In the short term you may need to decode an unfamiliar word encountered while reading, or resolve a communicative problem in order to complete an assignment. This is where a "quick fix" is in order, and MLDs are not always well-adapted to this role. But if your longer-term goal is to become proficient in a second language, the process of consulting an MLD brings benefits in terms of learning (as opposed to merely problem-solving). As noted here, "by reading English-language definitions, learners get greater exposure to English and learn the language within its own system". This comes up in the section where the author compares definitions in MLDs with translation equivalents in bilingual dictionaries, and demonstrates that – given the anisomorphism of language systems as different from each other as Japanese and English – one-to-one equivalents rarely tell the whole story. But as he concedes, "this actually presents no major problems when confirming the meaning of specific things such as flora and fauna". This observation perhaps points to future developments. Digital dictionaries can use different media for different purposes, and for certain types of word a conventional definition is less helpful than, say, an image, a video, or a sound file. The function of definitions, as Bolinger observed 50 years ago, is "to help people grasp meanings [by supplying] a series of hints and associations that will relate the unknown to something known" (Bolinger, D. 1965. The atomization of meaning. *Language* 41: 555-573). And if that job can be done more efficiently through nonverbal media, perhaps that is what we should focus on in such cases.

The guide also explains the features that distinguish MLDs from other kinds of monolingual dictionary, emphasizing the "approachability" of the definitions and examples. This is an important argument – more so than ever now, when every type of dictionary is freely available. A language learner who looks up *condescension* in Wiktionary (en.wiktionary.org), for example, is unlikely to get beyond the first definition:

> The act of condescending; voluntary descent from one's rank or dignity in intercourse with an inferior

Even I am having problems working out what this means, and in the unlikely event of a learner successfully decoding the definition, it wouldn't solve any problems because it fails to correspond to any normal use of the word. (The definition is in fact lifted, verbatim, from an ancient Webster's dictionary.) This is the kind of thing that gives monolingual dictionaries a bad name, and Yamada is right to stress the superiority of corpus-based MLDs over many of the free offerings on the Web.

Classroom activities for dictionaries typically focus on specific data types (information on meaning, collocation, and the like), in order to familiarize users with the way different kinds of information are conveyed and thus to facilitate dictionary use. What is interesting (and original) here is that the process of consulting a dictionary is framed in terms of seven distinct steps, and activities are proposed for most of these. Dictionary consultation is seen as a "complex intellectual activity" (even if proficient users perform it unconsciously), which proceeds from recognising the communicative problem and determining what the problematic word or multiword unit is, through finding the "right" information

in the dictionary, extracting the data you need, and applying this information in order to resolve your problem. Along the way, a number of definition conventions are helpfully explained. Some of the advice on finding the appropriate entry is less applicable to digital dictionaries than to traditional print-based ones: finding phrasal verbs and idioms, for example, is far easier in a well-structured online dictionary, where the trend is for these to be separate entries (rather than "nested" at the end of a base form). Intelligent search algorithms take you straight to an idiom even if you don't know the exact canonical form. (Locating *close/shut the stable door after the horse has bolted* in a paper dictionary was as big a problem for users as deciding where to put it was for lexicographers. No more.)

One task lists a number of common English words and expressions (such as *not bad*), and asks users to compare the English definition with a corresponding translation equivalent in an English-Japanese dictionary. This is a neat way of showing how items like these don't always map conveniently from one language to another, and again makes the case for using an MLD.

The guide is aimed at Japanese learners of English, but much of it would be useful for teachers and learners with other first languages. And though produced for a particular dictionary publisher (OUP), it is far more than a mere promotional tool. The advice it gives is refreshingly even-handed and all the main MLDs are referred to at different points. (One quibble is that the URL given for the *Macmillan Dictionary* site is for a long-defunct version: the correct address is http://macmillandictionary.com/.)

There is occasionally an elegiac feel about the guide, in that some of the advice relates to using print dictionaries, and it is hard to imagine the average high-school student in Japan (now by definition a digital native) consulting one of these (unless forced to!). And perhaps more could have been said about some of the excellent complementary resources available on the Web. While students are advised (p13) to use the "example banks" in dictionary CD-ROMs to match examples to word senses, many would feel more at home with a Web resource such as SKELL (http://skell.sketchengine.co.uk/).

In the end, we are left with the question of whether teaching dictionary use is a worthwhile project in itself. Yamada believes that, when a user's search for information is unsuccessful, "either the dictionary or the user is to blame". My default position is that if users can't readily find what they are looking for, the fault lies squarely with the dictionary. Consequently, the onus is on dictionary producers to ensure that information is easy to locate and easy to digest – an approach which feels more in tune with the way that software products are designed nowadays so that no instruction manual is needed. Few students will be fortunate enough to have a teacher who understands dictionaries as well as the author of the guide. In most cases, they must rely on their dictionary being well enough designed to make its use intuitive. Having said that, this guide will give teachers who are not especially dictionary-aware the resources to demonstrate to their students the benefits of using a monolingual learner's dictionary.

Michael Rundell
Lexicography Masterclass
and Macmillan Dictionary
michael.rundell@lexmasterclass.com

MULTILINGUAL

Semi-Automatically Generated Multilingual Glossaries

KD's updated English Multilingual Dictionary (EMD, see p25) now serves as a base for developing new multilingual glossaries for other languages. The process begins by reverse-engineering the Password dictionary data (which is at the heart of the EMD) in order to produce a raw index for each language to English. Next, a dedicated software tool is used to manually edit and refine the index, including the linking of each L1 headword to the corresponding sense(s) of the original English entries. Finally, the translations from all other languages to every particular sense in the EMD are associated automatically, turning the L1-English index into an L1 multilingual glossary with translations to 43 languages. So far, multilingual glossaries were created for these 20 languages:

Catalan | Chinese Simplified | Danish | Dutch | Estonian | French | German | Hungarian | Indonesian | Italian | Japanese | Norwegian | Polish | Portuguese Brazil | Portuguese Portugal | Romanian | Russian | Slovene | Spanish | Swedish

The Linguistic Linked Open Data Cloud