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.

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

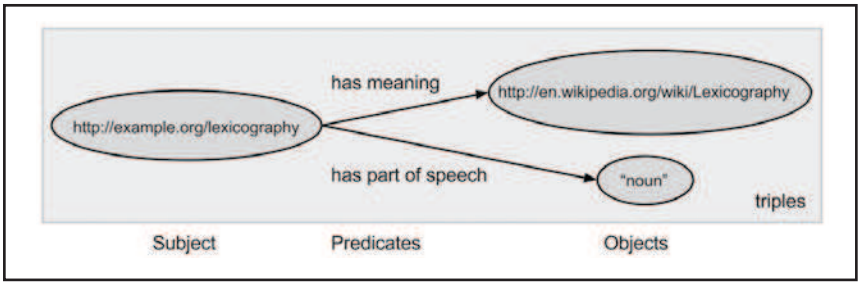

Figure 1: Example showing two triples

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

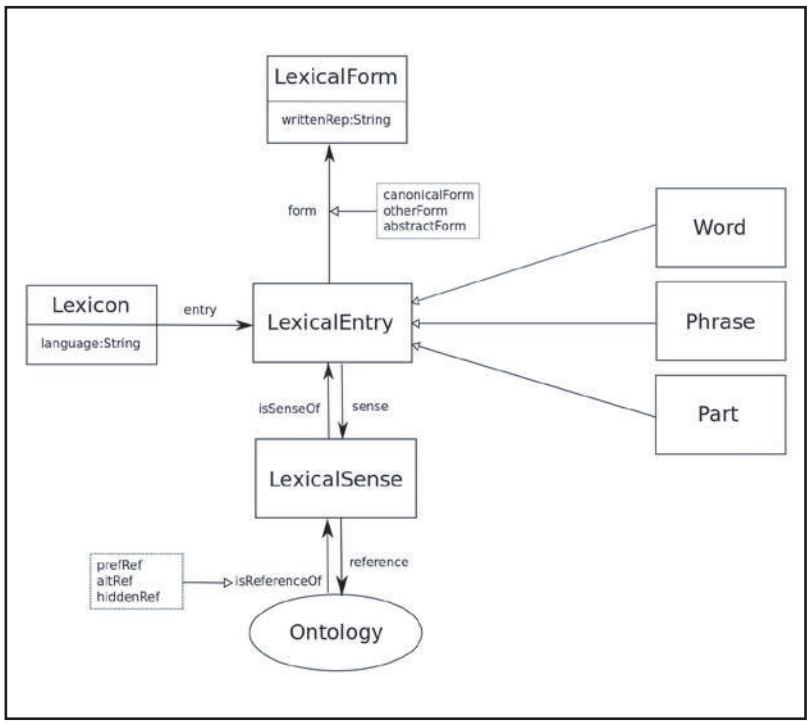

Figure 2: The *lemon* core path

```
@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"

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).

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

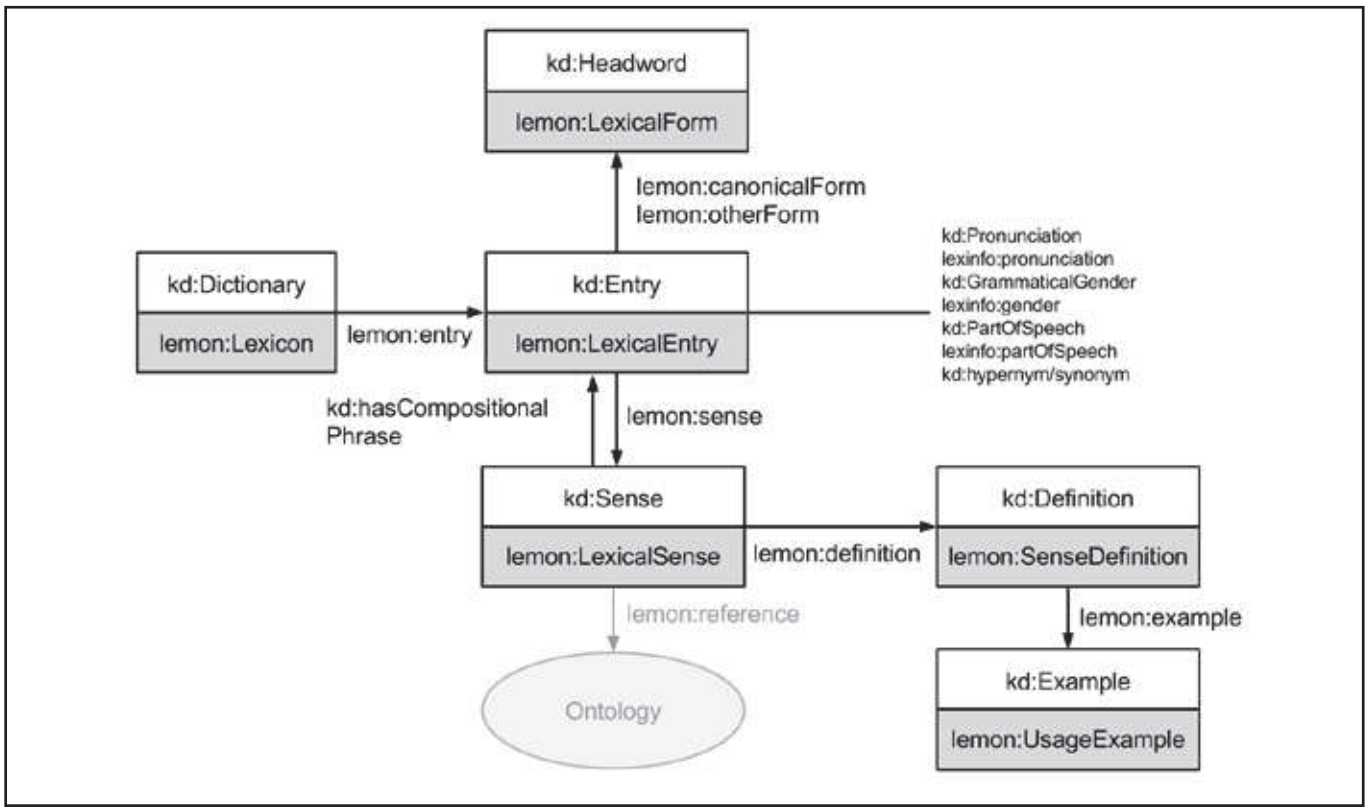

Figure 6: Mapping KD's XML elements and *lemon* resources (excluding the greyed out Ontology part)

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/.