Applying the OntoLex-*lemon* lexicography module to K Dictionaries' multilingual data

Dorielle Lonke and Julia Bosque-Gil

Introduction

In recent years, K Dictionaries (KD) has been working on representing its multilingual lexicographic resources in Linked Data (LD) format (Bizer et al. 2011), adhering to the RDF Ontolex-*lemon* model for lexical resources (cf. McCrae et al. 2017). The latest iteration, which began in 2019, follows two previous rounds, and focuses on adapting the newly added *lexicog* module[1], which addresses the need to preserve the original structure of a lexicographic dataset. In cooperation with Julia Bosque-Gil and Jorge Gracia from University of Zaragoza, the KD team has been involved in implementing *lexicog* in the Global series of multi-language, multi-layer resources (cf. Bosque-Gil et al. 2019). The contributions of this effort are twofold: primarily, offering the first use case of real-world lexicographic data represented entirely in LD format; secondly, addressing previously reported limitations of the Ontolex-*lemon* model and offering new solutions, allowing a more agnostic approach to a graph representation of lexicographic data.

The first two iterations of this data conversion were carried out under a strict principle of 'round-tripping', wherein every element in the original XML structure must be accounted for within the new corresponding RDF structure in terms of hierarchy and order, thus enabling one-to-one reconstruction of the lexicographic resource from RDF back to XML (Klimek and Brümmer 2015, Bosque-Gil et al. 2016). In addition to obtaining perfect matching back and forth between the two data forms, this principle served to assure perfect validation of the data conversion from either format to the other. However, this principle was abandoned in the latest iteration due to the understanding that each type of data format should be applied freely and fully in accordance with its own nature and not be restricted by characteristics of the other. In consequence, the alternative validation process provided in this iteration introduced a new incremental approach, which actually proved to be more efficient in validation and error catching. In section 2 we describe the revised

1 https://www.w3.org/2019/09/lexicog

Dorielle Lonke has been working at K Dictionaries for the past three years, involved in different projects pertaining to linguistic data design and managment, and leading conversion from XML to RDF. Currently she is graduating in Linguistics and completing her Philosophy degree at Tel Aviv University. Her main fields of interest include language technologies, computational processes in natural languages and ontological representation of data.
dorielle@kdictionaries.com

pipeline which features the incremental approach. The modelling improvements that were performed as part of the new *lexicog* module adaptation are presented in section 3, and section 4 reports on the ensuing validation process, including queries and results, followed by a brief summary of the process in section 5.

The pipeline

The incremental approach saw a step-by-step modelling process, in which each component was first modelled to fit the *lexicog* module, and then manually described as RDF triples serialized in Turtle format (TTL)[2]. Based on the manual description, a generalization was applied in the automatic conversion process, allowing the simultaneous conversion of numerous entries in the dataset. The final stage of this process was to upload the dataset onto a triple store, enabling querying and detection of errors, and fixing such errors in the conversion. The process was repeated for each cluster of components. Instead of committing to a restrictive structure that requires a one-to-one conversion, this has enabled a looser, more flexible workflow that facilitated casting off excessive information that encumbers the model, while still fitting each lexicographic component with its ontology counterpart and retaining the original hierarchy and order of the lexical data where necessary.

Progressing incrementally has not only enabled constant validation and error management, but also allowed for an adaptation period, during which the process of writing queries for validation shed light on the model and methods of improvement. Taking into account input from partners and collaborators who have been experimenting with the RDF data, particularly as part of our work in the H2020 Lynx project[3], we were able to improve the queries and iteratively modify the model so that the results optimally represent the needs of the users. This working method has proved efficient, not only in the sense of illuminating problems that would have otherwise remained unknown, but also due to the involvement of practical users who make use of the RDF data, resulting in a model that is both theoretically and empirically sound.

Advancements in modelling

The ultimate goal of the latest iteration was to retain generalizability and universality of the model, while still representing the richness and complexity of the Global series. To that end, the iteration involved numerous updates and improvements to the 2016 model which

2 https://www.w3.org/TR/turtle/
3 http://lynx-project.eu/

Julia Bosque-Gil is a postdoctoral researcher at the Distributed Information Systems Group at University of Zaragoza. She has recently obtained her PhD at the Ontology Engineering Group (Universidad Politécnica de Madrid) for her thesis investigating the use of linguistic linked data for lexicography, and collaborated with K Dictionaries and Semantic Web Company in the representation of multilingual lexicographic data as RDF in the LD4HELTA project. She is currently working in the representation, transformation and linking of multilingual resources as linguistic linked data as part of the Prêt-à-LLOD project and the NexusLinguarum COST Action.
jbosque@unizar.es

was proposed in the framework of the round-tripping condition (Bosque-Gil et al. 2016). One facet of improvement was a thorough revision of the mappings proposed for the different XML paths, the KD ad-hoc vocabulary that was developed for internal use to bridge the gaps between the OntoLex and LexInfo RDF vocabularies, and the KD XML Schema. In addition to applying the *lexicog* module illustrated in Bosque-Gil et al. (2019), this revision saw an update of the KD XML Schema (DTD) in terms of its collection of predetermined semantic or syntactic cues and their values. In the 2016 model, the KD vocabulary included individuals, classes and properties that could not be directly mapped to the LexInfo vocabulary, primarily for two reasons: (a) mismatches between the DTD values of a tag and LexInfo classes, and (b) a different level of granularity in the predefined values in the DTD and the individuals in the linguistic category registry of LexInfo[4]. Given that in these cases a one-to-one mapping from KD into LexInfo was not viable, new elements had to be created under the KD namespace, for example, *kd:prepositionalCase*. The 2019 revision attempted to align the KD DTD values with LexInfo's most recent version[5] as much as possible, to avoid the less desired solution of adding ad hoc ontology elements to represent elements unique to KD, and thus limiting the possibility of linking to external resources. In general, the conversion strives to be as universal as possible, to allow more extensive cross-linking to different resources and consequently expanding the graph. The 2019 conversion has extended the outreach of LexInfo elements, covering significantly more data in KD versus the previous iterations. However, the source data annotation does not only pertain to lists of DTD tags and predefined values; part of the lexicographic workflow takes into account the free values that editors suggest for a given tag, especially in cases in which the predefined list of attribute values does not offer an adequate annotation in the editor's eyes and hence a nuance or further detail is provided. Since this is valuable content for both the data description as well as the day-to-day operations of KD aimed at schema improvement, the 2019 revision is systematically treating free values provided by the editors as individuals in the KD namespace. By dynamically adding these values to the namespace every time the pipeline runs, we allow for future inference of their types thanks to restrictions on properties range, as well as future careful revision of them and even consideration as a new (predefined) value in the Schema, reflecting semantic and pragmatic shifts in the language, or as a potential replacement for a predefined value of which the usage is gradually in decline.

4 http://www.lexinfo.net/ontology/2.0/lexinfo.owl
5 http://www.lexinfo.net/ontology/3.0/lexinfo

```
PREFIX lexicog: <http://www.w3.org/ns/lemon/lexicog#>
PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>
PREFIX skos: <http://www.w3.org/2004/02/skos#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX vartrans: <http://www.w3.org/ns/lemon/vartrans#>
PREFIX lime: <http://www.w3.org/ns/lemon/lime#>
PREFIX lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?mainEntry ?nestedEntry ?lexicalEntry ?originalResource
WHERE{
  ?mainEntry rdfs:member ?nestedEntry .
  ?nestedEntry lexicog:describes  ?lexicalEntry .
  ?originalResource lexicog:entry ?nestedEntry .
} LIMIT 500
```

Query 1. Retrieving a list of nested dictionary entries, along with their containers and the lexical entries they refer to, according to the OntoLex lexicog module

mainEntry	nestedEntry	lexicalEntry	originalResource
:ES-00000024-00000025-nested	:ES_DE00000024	:LexiconES/abajar-vb	:mlds-ES3
:ES-00000024-00000025-nested	:ES_DE00000025	:LexiconES/abajar-vb	:mlds-ES3
:ES-00000026-00000027-nested	:ES_DE00000026	:LexiconES/abajo-adv	:mlds-ES3
:ES-00000026-00000027-nested	:ES_DE00000027	:LexiconES/abajo-interj	:mlds-ES3
:ES-00000028-00000029-nested	:ES_DE00000028	:LexiconES/abalanzar-vb	:mlds-ES3
:ES-00000028-00000029-nested	:ES_DE00000029	:LexiconES/abalanzarse-vb	:mlds-ES3

Table 1. Extract from the results of Query 1

Validation and querying

Following the development process described in section 3, the validation step consists initially of a JSON Schema (detailed in Bosque-Gil et al. 2019), followed by performing a series of queries on the KD SPARQL endpoint, with specific queries designed for each step in the conversion. These predefined queries permit inspecting the modelling tag by tag in the DTD, as its OntoLex-*lemon* counterpart. For example, Query 1 allows to validate the representation of containers of nested entries and their links to the lexical entries they describe.

An extract of Query 1 results is shown in Table 1. The same dictionary entry container (e.g. 024-025-nested) contains two nested entries (24 and 25), which both describe the Spanish verb *abajar* [to lower, decrease], having two different forms (transitive vs intransitive) hence originally separated into two entries. The container 026-027 groups together two dictionary entries, *abajo* [down, downstairs] (adverb) and *abajo* [down!] (interjection), and 028-029 represents the container that in the original resource gathered the transitive and reflexive uses of *abalanzar* [to leap on, jump, throw].

```
SELECT DISTINCT ?entry
{
  ?entry ontolex:lexicalForm [ontolex:writtenRep "bow"@en ] .
}
```

Query 2. (same prefixes as in Query 1 apply) Retrieving all lexical entries with the lemma *bow* in English

```
SELECT DISTINCT ?sense
{
  :LexiconEN/bow-n ontolex:sense ?sense .
}
```

Query 3. (same prefixes as in Query 1 apply) Retrieving all senses linked to the artificial entry :LexiconEN/bow-n, created to act as a "container" of senses and to allow linkage with other resources if the homograph number is unknown

Queries 2 and 3 delve into the modelling of homographs. Query 2 retrieves the list of lexical entries with the lemma *bow* (noun) in English, which will include separate lexical entries for homographs.

Query 3 retrieves all lexical senses linked to the artificial entry :LexiconEN/bow-n. The artificial entry *bow* enables gathering the information originating from the different homographs (i.e. from :LexiconEN/bow-n-1, and :LexiconEN/bow-n-3), as well as from other dictionaries in which *bow* is given as a translation (without specifying to which homograph it applies). Thanks to this method of clustering, the query currently results in 56 possible senses in different

sense
:LexiconEN/bow-n-arco-n-ES_SE00006877-sense
:LexiconEN/bow-n-arco-n-ES_SE00006878-sense
:LexiconEN/bow-n-inclinaci%C3%B3n-n-ES_SE00041443-sense
:LexiconEN/bow-n-lazo-n-ES_SE00045188-sense
:LexiconEN/bow-n-mo%C3%B1o-n-ES_SE00050680-sense
:LexiconEN/bow-n-proa-n-ES_SE00060090-sense
:LexiconEN/bow-n-reverencia-n-ES_SE00065273-sense
:LexiconEN/bow-n-arco-n-IT_SE00002942-sense
:LexiconEN/bow-n-arco-n-IT_SE00002945-sense
:LexiconEN/bow-n-fiocco-n-IT_SE00016220-sense
:LexiconEN/bow-n-gala-n-2-IT_SE00017474-sense
:LexiconEN/bow-n-prora-n-IT_SE00033071-sense
:LexiconEN/bow-n-riverenza-n-IT_SE00036412-sense
:LexiconEN/bow-n-Bogen-n-DE_SE00006470-sense
:LexiconEN/bow-n-Bogen-n-DE_SE00006471-sense

Table 2. Extract of the results from Query 3

languages of the word *bow* as a noun in English across the Global series.

An extract of the list of results for Query 3 is shown in Table 2. In this way, the artificial entry :LexiconEN/bow-n, which was absent in the original resource, serves now as a linking point with other multilingual resources in the Global series as well as an entry point for the different senses of the homographs in the English dataset (cf. Image 1).

These query samples exemplify the way in which uploading the data onto a triple store and querying it for results enable validation of the conversion. By counting instances in the original data and matching with the triple store results, we can make sure that every XML component underwent proper conversion to an RDF counterpart. By examining the results of goal-oriented queries, such as obtaining nested entries or all related instances of a noun (as in the case of *bow*), we can see clearly how the cross-linking operates de facto. By determining an incremental conversion and querying of the results as our *modus operandi*, we were able to obtain a clear visual and structural representation of the model. The opposite also applies: by querying a particular instance and retrieving an unexpected result, we were able to handle errors and fix them in the origin.

Summary

Following recent publications regarding the theoretical aspects of KD's initiative and collaboration with partners to convert its data

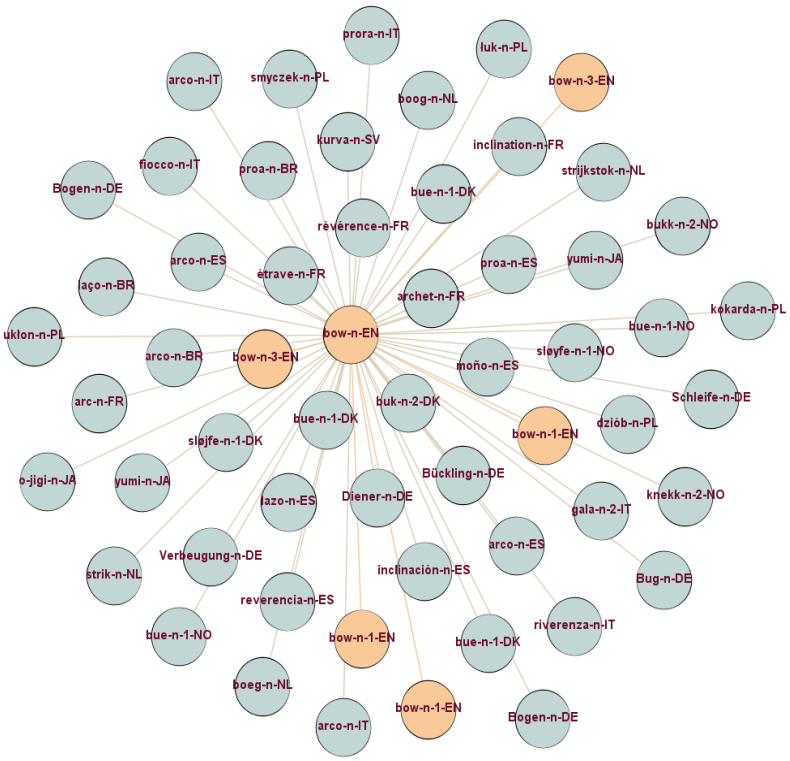

Image 1. The artificial entry :bow-n and its links to senses; senses stemming from translations are in green/blue and those from English in orange

to an LD format, this short paper demonstrates the application of the modelling in terms of pipeline, practical advancements to the model, and the validation and querying process. This endeavor features a real-world example of RDF representation of lexicographic data, demonstrating how a model should account for the structural constraints of a lexical resource, as well as the linguistics shifts and changes to the semantic and syntactic information that is represented therein. The dynamic vocabulary is just one example, proving that in a constantly changing environment, the theoretic representation should be able to develop accordingly. We exemplified how a good model that takes into account such constraints while retaining as much information as possible and remaining flexible, will yield impressive and expansive results. Such is the case of the artificial entry *bow* (noun), which gathers information across all lexical resources of the Global series, creating a de facto graph of cross-linked information within one larger context. Finally, through mutual consultation and exchange we were able to design the queries to match our partners' needs and obtain the best results for real-word applications.

References

Bizer, C., Heath, T., and Berners-Lee, T. 2011. Linked data: The story so far. *Semantic services, interoperability and web applications: emerging concepts*: 205-227. IGI Global.

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., and Aguado-de-Cea, G. 2016. Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case. *Proceedings of GLOBALEX 2016: Lexicographic Resources for Human Language Technology*: 65-72. http://www.lrec-conf.org/proceedings/lrec2016/workshops.html

Bosque-Gil, J., Lonke, D., Gracia, J., and Kernerman, I. 2019. Validating the OntoLexlemon Lexicography Module with K Dictionaries' Multilingual Data. *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*: 726-746. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_41.pdf

Klimek, B., and Brümmer, M. 2015. Enhancing lexicography with semantic language databases. *Kernerman Dictionary News*, 23, 5-10. https://www.kdictionaries.com/kdn/kdn23_2015.pdf

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. 2017. The Ontolex-Lemon model: Development and applications. *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*: 19-21. http://john.mccr.ae/papers/mccrae2017ontolex.pdf

This work has been supported by the European Union's Horizon 2020 research and innovation programme through the Lynx project (grant agreement No 780602). It has been also partially supported by the Spanish projects TIN2016-78011-C4-3-R (AEI/FEDER, UE) and DGA/FEDER 2014-2020 "Construyendo Europa desde Aragón". http://lynx-project.eu/