Triplifying a dictionary: Some learnings

Timea Turdean and Shrikant Joshi

Timea Turdean is Technical Consultant at Semantic Web Company, Vienna. In her current position she supports clients and partners integrating semantic technologies and is involved in different research projects dealing with linguistic data, earth observation data and publication data. She holds a MSc. from Vienna University of Technology, and her background is in text mining and sentiment analysis. timea.turdean@semantic-web.com

1. Introduction

The Linked Data Lexicography for High-End Language Technology (LDL4HELTA) project[1] was launched in cooperation between Semantic Web Company (SWC)[2] and K Dictionaries (KD)[3], combining lexicography and computational linguistics with semantic and linked (open) data mechanisms and technologies. One of the implementation steps of the project was to create a language graph from the dictionary data. The input data consists of the Spanish lexicographic resource of KD, which is translated into multiple languages and is available in XML format. The data needed to be triplified (that is, converted to RDF[4]) for several purposes, including enhancing its enrichment with external resources.

Section 2 of this article describes previous work carried out in this domain. Section 3 discusses in detail the actual process of triplification of the dictionary XML into RDF. An interesting experiment was carried out by using and applying the same principles for the translation of a dictionary, as described in Section 4. Although the initial success has ratified the process, some work is still required to explore and enhance it further, which is described as part of the conclusions in Section 5.

2. Previous work

There are different initiatives and efforts that investigate the process and usefulness of triplifying lexicographic data. Terminesp[5] is a well-known database that was transformed into RDF following linked data best practices (cf. Gracia 2015). Our work builds on the findings of Klimek and Brümmer (2015), who have investigated the usage of the Lemon model[6] on KD's German lexicographic XML data, and demonstrated how it can be represented in RDF and noted some missing elements that needed to be reconsidered. Bosque-Gil et al. (2016) also report about combining linked data in lexicography, particularly regarding usage of the Ontolex model[7] on the monolingual part of KD's Spanish dataset mentioned above.

3. Triplification process

The triplification of a dictionary is a process of mapping its data (which in KD's case is in propriatory XML format) to RDF triples. Following the triplication process, the resulting data was stored in a database to facilitate further processing.

In previous works, an RDF lexicographic model was proven to work for KD's lexicographic resources. The present article reports on how this model was applied on the Global Spanish dataset (i.e. the monolingual core and its translations in other languages) and triplified. In the process we ensured that the RDF complied with Semantic Web (SW) standards[8].

3.1. Nature of a dictionary entry

The XML format of KD's Global Spanish dataset consists of a complex structure containing nested components. Each word constitutes an entry, containing information such as: pronunciation; inflections; range of application; sense indicators; compositional phrases; translations (of different components); alternative scripts; register; geographical usage; sense qualifier; version; synonyms; lexical sense; examples of usage; homograph information; language information; specific display information; identifiers; and more…

Entries can have predefined values that can recur, but their fields can also have so-called free values, which can vary too, including: Aspect; Tense; Subcategorization; Subject Field; Mood; Grammatical Gender; Geographical Usage; Case; and more…

3.2. Constructing a lexical model

After studying the entry structure, it was necessary to construct a model representing the entries in the SW conceptual form to go from the dictionary's XML format to its triples. The model was designed by Bosque-Gil et al. (2016), and an example representing two Spanish words having senses that relate to each other is presented in Figure 1.

1 https://ldl4.com/
2 https://www.semantic-web.at/
3 http://kdictionaries.com/
4 https://www.w3.org/RDF/
5 http://linguistic.linkeddata.es/terminesp/
6 http://lemon-model.net/
7 https://www.w3.org/community/ontolex/wiki/Final_Model_Specification
8 http://semanticweb.org/wiki/Semantic_Web_standards.html

Usually, when modelling linked data or just RDF it is important to make use of existing models and schemas to enable easier and more efficient use and integration. A well-known lexicon model is Lemon[9], whose core path can cover some of this dictionary's needs (cf. Klimek and Brümmer, 2015), but not all of them. The Ontolex model[10], which is more complex and considered to be the evolution of Lemon, offers more capabilities in this regard. However, also after adapting the KD data to the OntoLex model, some pieces of information were still missing and an additional ontology was needed to be created to cover all such elements and catch the specific details that did not get sufficiently treated (such as the free values). We named this model extension OntolexKD.

The process used to do the mapping from KD's XMLs to RDF consists of several steps. This can be visualised as a processing pipeline which manipulates the XML data. The tool that we used for this mapping was UnifiedViews[11]. This is an ETL (Extract, Transform and Load) tool with which you can configure your own data processing pipeline to generate RDF data. One of its use cases is to triplify different data formats and store the resulting RDF data in a database. Our processing pipeline appears in UnifiedViews as displayed in Figure 2.

The pipeline is composed of data processing units (DPUs) which communicate with each other iteratively. In a left-to-right order, the process outlined in Figure 2 represents:

- A DPU used to upload the XML files into UnifiedViews for further processing;
- A DPU which transforms XML data to RDF using XSLT[12]. The style sheet is part of the configuration of the unit;
- The .rdf generated files are stored on the filesystem;
- Finally, the .rdf generated files are uploaded into a triple store, such as Virtuoso Universal Server[13].

3.3. URIs

Complexity increases also through the URIs (Uniform Resource Identifier) that are needed for mapping the information in the dictionary since linked data requires every resource to have a clearly identified and persistent identifier. The start was to represent a single word (headword) under a desired namespace and build on it to associate it with its part of speech, grammatical gender and number, definition and translation.

The base URIs follow the best practices recommended in the ISA study on persistent URIs[14] following the pattern: http://{domain}/{type}/{concept}/{reference}.

An example of such URIs for the forms of a headword is:

- http://kdictionaries.com/id/lexiconES/entendedor-n-m-sg-form
- http://kdictionaries.com/id/lexiconES/entendedor-n-f-sg-form

9 http://lemon-model.net/
10 https://www.w3.org/community/ontolex/wiki/Final_Model_Specification
11 https://unifiedviews.eu/
12 https://www.w3.org/Style/XSL/
13 https://virtuoso.openlinksw.com/
14 http://philarcher.org/diary/2013/uripersistence/

Shrikant Joshi holds a PhD in Linguistics from Université de Lausanne, with a focus on the semantics of affixation, its formalisation and subsequent computational processing, and a BE in Electronics Engineering and an MA in French from University of Pune. He has been teaching courses in NLP, French and German Linguistics at the University of Pune as a visiting lecturer. Currently he is working as Technical Consultant and Researcher at Semantic Web Company. shrikant.joshi@semantic-web.com

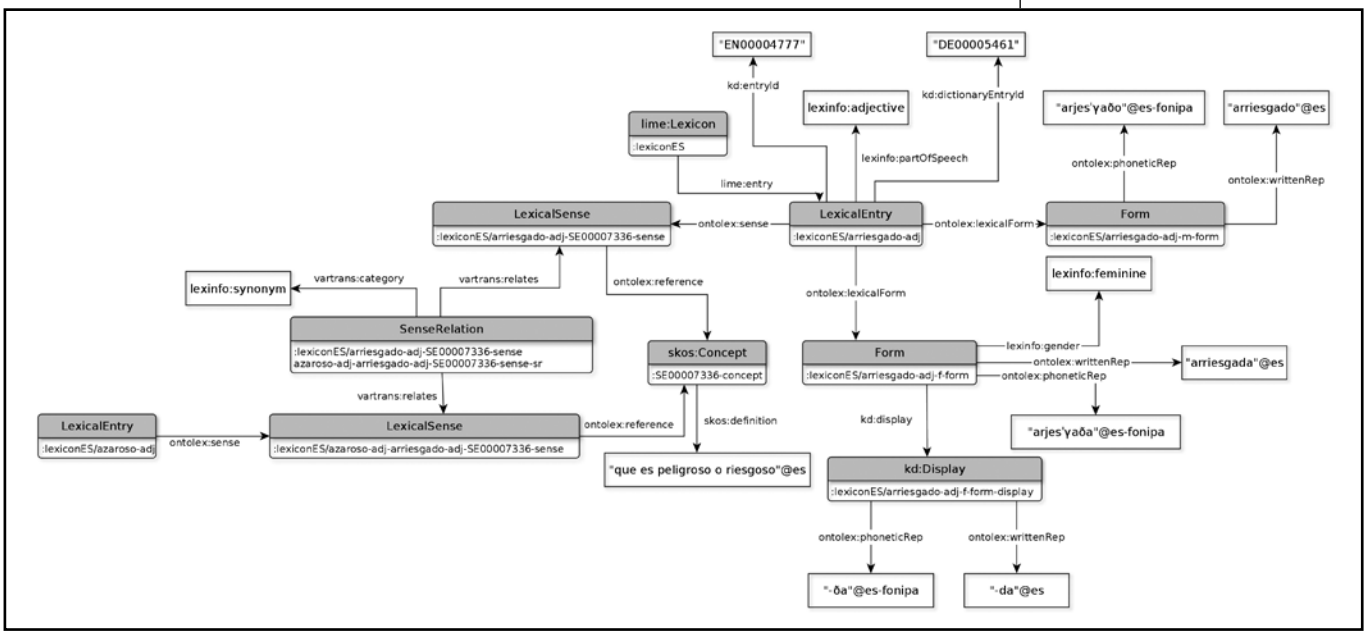

Figure 1: Language model example

TIAD shared task 2017 – Translation Inference Across Dictionaries

The first shared task on Translation Inference Across Dictionaries was aimed to explore best methods and techniques for automatic generation of new bilingual dictionaries based on existing resources. It relied on extracts from 15 bilingual dictionaries of K Dictionaries (KD) for developing three new language pairs that were validated against existing KD data and by human translators.

TIAD 2017 was organized by Noam Ordan, Morris Alper and Ilan Kererman (KD) and Jorge Gracia (OEG, Madrid Politechnic University). The results were presented in a workshop co-located with the Language, Data and Knowledge conference at NUI Galway on June 18, 2017 by four teams:

- Kathrin Donandt, Christian Chiarcos and Maxim Ionov; Goethe University, Frankfurt
- Tom Knorr; Neurocollective, San Francisco CA
- Thomas Proisl, Philipp Heinrich, Stefan Evert and Besim Kabashi; Erlangen University
- Uliana Sentsova; National Research University Higher School of Economics, Moscow

The papers are published as part of the LDK 2017 Workshop Proceedings http://ceur-ws.org.

Noam Ordan

https://tiad2017.wordpress.com/

These two URIs represent the singular masculine and singular feminine forms of the Spanish word *entendedor*.

- http://kdictionaries.com/id/lexiconES/entendedor-adj-form-1
- http://kdictionaries.com/id/lexiconES/entendedor-adj-form-2

If the dictionary contains two different adjectival endings, as with *entendedor* which has different endings for the feminine and masculine forms (*entendedora* and *entendedor*), and they are not explicitly mentioned, then we use numbers in the URI to describe them. If the gender is explicitly mentioned, then the URIs would be:

- http://kdictionaries.com/id/lexiconES/entendedor-adj-form
- http://kdictionaries.com/id/lexiconES/entendedora-adj-form

In addition, it should be considered that the aim of triplifying the XML was for all these headwords with senses, forms and translations, to connect and be identified and linked following the SW standards.

One of the last steps of complexity was to develop a generic XSLT which can triplify all the different languages of this dictionary series and store the complete data in a triple store. The question remains whether the design of such a universal XSLT is possible while taking into account the differences in languages or the differences in dictionaries.

4. Application and exploration

We tried to investigate also whether the automated resource linking could help with the translation of one dictionary into another the language. Two bilingual dictionaries were considered - English(en)-German(de) and German(de)-English(en).

For the word *bank* the following translations are found:

Bank (de) – *bank* (en) – German to English
bank (en) – *Bank* (de) – English to German

The URI of the translation from German to English was designed to look like:

- http://kdictionaries.com/id/tranSetDE-EN/Bank-n-SE00006116-sense-bank-n-Bank-n-SE00006116-sense-TC00014378-trans

And the one for the translation from English to German would be:

- http://kdictionaries.com/id/tranSetEN-DE/bank-n-SE00006110-sense-Bank-n-bank-n-SE00006110-sense-TC00014370-trans

In this case, both represent the same translation but have different URIs because they were generated from different dictionaries (in accordance with the translation order) that need to be mapped to each other so as to represent the same concept.

The word *Bank* in German can mean either a bench or a bank in English. When either of these English senses is translated back into German the result is the German word *Bank*. It is, however, not possible to determine which sense out of the two was translated unless the URI that contains the sense ID is included. It is also important to maintain the order of translation (source-target) but later map both translations to the same sense and same concept. This is difficult to establish automatically.

5. Future work

The actual overlap and automatic linking of the dictionary resources remains to be tested. There are also some lexicographic elements which were not covered by the new OntolexKD model and need to be added.

There is also the necessity to verify and check for differences between KD's XML dataset and the derived KD's triplified dataset. For this, SPARQL queries need to be created that validate and verify the resulting RDF.

References

Bosque-Gil, J., J. Gracia, and A. Gómez-Pérez. 2016. Linked data in lexicography. *Kernerman Dictionary News* 24, 19-24.

Bosque-Gil, J., J. Gracia, E. Montiel-Ponsoda, and G. Aguado-de Cea. 2016. Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case. In *Proceedings of GLOBALEX 2016 Workshop at the 10th Language Resources and Evaluation Conference (LREC 2016), Portorož, (Slovenia).*

Gracia, J. 2015. Multilingual dictionaries and the Web of Data. *Kernerman Dictionary News* 23, 1-4.

Klimek, B., and M. Brümmer. 2015. Enhancing lexicography with semantic language databases. *Kernerman Dictionary News* 23, 5-10.

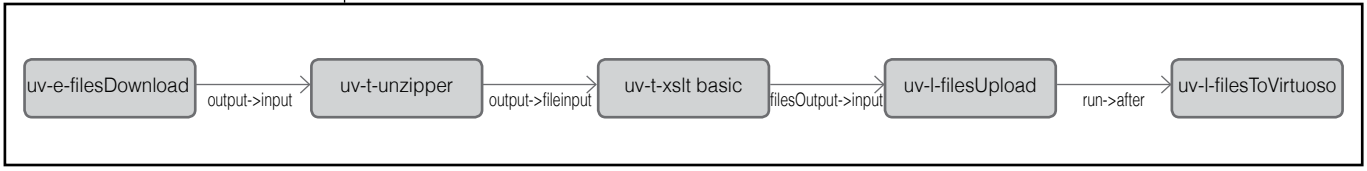

Figure 2: UnifiedViews pipeline used to triplify XML