Enhancing Lexicography by Means of the Linked Data Paradigm: LexO for CLARIN

Andrea Bellandi, Fahad Khan, Monica Monachini
Institute For Computational Linguistics "A. Zampolli"
CNR Pisa, Italy
name.surname@ilc.cnr.it

Abstract

This paper presents a collaborative web editor for easily building and managing lexical and terminological resources based on the OntoLex-Lemon model. The tool allows information to be easily manually curated by humans. Our primary objective is to enable lexicographers, scholars and humanists, especially those who do not have technical skills and expertise in the Semantic Web and Linked Data technologies, to create lexical resources *ex novo* even if they are not familiar with the underlying technical details. This is fundamental for collecting reliable, fine-grained, and explicit information, thus allowing the adoption of new technological advances in the Semantic Web by the Digital Humanities.

1 Introduction and Motivation

Lexicography is traditionally recognised as that branch of applied linguistics which is concerned with the design and construction of resources that describe the lexicon of a language. In the digital era, it is very important that language resources can be represented in such a way that machines can process them and that they can be queried and shared across different communities. From this perspective it is possible to imagine a large-scale interconnected ecosystem of open, queryable and standardised lexicographic datasets and technologies. The Semantic Web, in particular the Resource Description Framework[1] (RDF), the Ontology Web Language[2] (OWL) and the Linked Data (LD) paradigm it is based on, makes this possible. Ontologies, in particular, have become an increasingly important method for formally modelling domains, and sharing them through the web.

In 2006, Tim Berners-Lee stated the four guiding principles for publishing data as LD[3]: i) use (Unique Resource Identifier) URIs as names for things; ii) use HTTP URIs so that people can look up those names; iii) when someone looks up a URI, provide useful information, using the standards (RDF, SPARQL[4]); iv) include links to other URIs, so that they can discover more things. These principles encourage both the maximum of interoperability between datasets and facilitate a more explicit encoding of meaning within and between datasets. The benefits of representing lexicographic content as LD are reusability, accessibility, interoperability and visibility at a Web scale. The content can be seamlessly integrated with content from external lexical resources. Lexical entries and their components are uniquely identified and become reusable thanks to URIs. The linguistic resource becomes a graph structure where each node is an entry point to navigate the whole graph, and each relation between two elements is typed and defined in a vocabulary.

In this context, the lemon model (McCrae et al., 2012), now called OntoLex-Lemon, was developed for creating lexicons that describe the lexicalization of ontological concepts. The number of users potentially

[1] See https://www.w3.org/RDF/ for RDF, and https://www.w3.org/TR/rdf-schema/ for RDFS (last access: 19/04/2019).

[2] https://www.w3.org/OWL/ (last access: 19/04/2019).

[3] If, in addition to the four prerequisites, data are made available under an open license, then it is classified as Linked Open Data (LOD).

[4] SPARQL query language for RDF https://www.w3.org/TR/rdf-sparql-query/ (last access: 19/04/2019).

interested in editing or consuming OntoLex-Lemon data is thus very large (McCrae et al., 2017), and also the rapid development of the Linguistic Linked Open Data (LLOD) cloud[5], is a success story influenced by the development of the OntoLex-Lemon model.

Our primary objective is to enable lexicographers, scholars and humanists to create lexical resources *ex novo* even if they are not familiar with the paradigm and the languages underlying their representation. Our tool allows information to be easily manually curated by humans. This is fundamental for collecting reliable, fine-grained, and explicit information. The tool we propose is among the first attempts to make the OntoLex-Lemon model accessible to all, especially to those who do not have technical skills in Semantic Web and Linked Data technologies, allowing the adoption of new technological advances in the Semantic Web by Digital Humanities.

2 Related works

Today, some tools for editing lexicons in different formats exist. Just to mention a few, we cite Lexus (Ringersma and Kemps-Snijders, 2007) and ColdicIn (Bel et al., 2008) for the Lexical Markup Framework (LMF) encoding, (Szymanski, 2009) for Wordnets, and CoBaLT (Kenter et al., 2012) for the management of lexicons in TEI P5.

In the context of the Semantic Web, editing tools for lexical or terminological resources are not so widespread, and in many cases, scholars are forced to adopt ontology editors, to formalize their lexical or terminological resources. Concerning the OntoLex-Lemon model, to the best of our knowledge, only two tools exist. The first one is lemon source, a Wiki-like site for manipulating and publishing lemon data aimed at the collaborative development of lexical resources. It makes it possible to upload a lexicon and share it with others. lemon source is an open source project, based on the lemon API, and it is freely available online for use. However, it deals with older versions of the OntoLex-Lemon model, and seems not to be not updated anymore. The most relevant tool to ours is VocBench a web-based, multilingual, collaborative development platform for managing OWL ontologies, SKOS(/XL) thesauri, Ontolex-lemon lexicons and generic RDF datasets. In (Fiorelli et al., 2017) the authors present their work on extending VocBench with facilities tailored to the OntoLex-Lemon model. However, LexO is more oriented on the needs of digital humanities. Firstly, we are working to link lexical senses to portions of text (e.g., attestations). Additionally, the editor is meant for formalizing peculiar features of linguistic resources such as etymology, representing aspects related to where words come from and how they originated, and diachrony for handling historical and ancient lexica and terminologies as well (Khan et al., 2016). It is worth emphasising that the process of extension in LexO is facilitated by the fact that OntoLex-Lemon, our lexical model of reference, is designed to be modular and to integrate new components easily.

3 LexO

Here, we present a first version of LexO[6] (Bellandi et al., 2018), called LexO-lite[7], that is a collaborative web editor for easily building and managing lexical and terminological resources for the Semantic Web and based on the OntoLex-Lemon model. The features of LexO were defined on the basis of our experience gained in the creation of lexical and terminological resources in the framework of several projects in the field of Digital Humanities. In particular:

- DiTMAO[8], a digital-born multilingual medico-botanical terminology focused on Old Occitan and developed by philologists;
- FdS, a multilingual diachronic lexicon of Saussurean terminology in the framework of a lexicographic project[9];

[5] `http://linguistic-lod.org/llod-cloud` (last access: 19/04/2019).

[6] The source code is available at `https://github.com/cnr-ilc/LexO-lite`. A simple demo of LexO adapted to a subset of Italian wordnet adjectives is available at `https://ilc4clarin.ilc.cnr.it/services/LexO`

[7] The suffix "lite" refers to the limited ability to manage small medium-sized lexica, due to the in-memory persistence we adopted, that is not a scalable strategy in case the resource size increases considerably. Currently, we are planning a "full" version of LexO for managing large resources.

[8] `https://www.uni-goettingen.de/en/487498.html` (last access: 19/04/2019)

[9] Demo available at `http://ditmao-dev.ilc.cnr.it:8082/saussure` (last access: 19/04/2019)

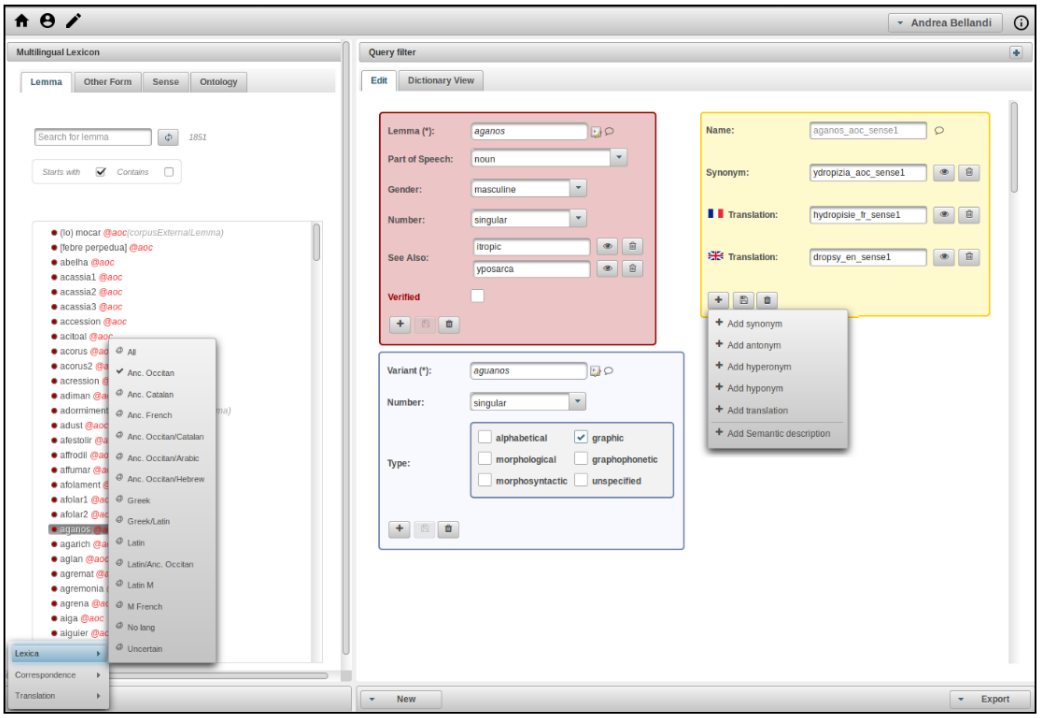

Figure 1: The main LexO's interface. (a) multilingual lexicon panel - (b) lexical entry editor.

- Totus Mundus, a bilingual Chinese-Italian resource dealing with Matteo Ricci's Atlas. LexO has been used by historians to build the linguistic resources related to the Map[10].

During the development of LexO we have been influenced by the some of the latest developments taking place in the European H2020 project ELEXIS (in which the Institute for Computational Linguistics of Pisa is involved as a partner). In particular our work has been closely informed by the survey of lexicographic users' needs conducted as part of the project and recently published as a deliverable[11]. This is to ensure that the tool can be potentially used in as wide a range of lexicographic contexts as possible. Furthermore, LexO hides all the technical complexities related to markup languages, language formalities and other technology issues, facilitating access to the Semantic web technologies to non expert users. It provides possibility for a team of users to work on the same resource collaboratively each one according his/her own role(s) (lexicographers, domain experts, scholars, etc.). Finally, it provides a set of services implemented by means of the RESTful protocols that give software agents access to resources managed by means of LexO. The main interface of LexO, as shown in Figure 1, concerns the editing of a multilingual lexicon. It is mainly composed of 2 parts. The leftmost column allows scholars to browse lemmas, forms and senses, according to the OntoLex-Lemon core model, as Figure 1(a) shows. If the resource is multilingual, then users have the possibility of filtering lemmas, forms and senses by language. Information related to the selected entry is shown in the central panel where the system shows the lexical entry of reference, alongside the lemma (red box), its forms (blue boxes) and the relative lexical senses (yellow boxes), as shown in Figure 1(b). It is also possible to list the concepts belonging to an ontology of reference, and link lexical senses to them.

[10]Demo available at `http://lexo-dev.ilc.cnr.it:8080/TMLexicon` (last access: 19/04/2019)

[11]See deliverable D1.1 "Lexicographic practices in Europe: a survey of user needs" at `https://elex.is/deliverables/` (last access: 19/04/2019).

4 The Potential Use of LexO in CLARIN and other Infrastructures

LexO fits in very well with a number of current CLARIN (as well as CLARIAH) initiatives and could prove itself a very useful tool within such an infrastructural context. For instance in the Netherlands in the context of CLARIAH, Linked Open Data has overtaken metadata publication in CMDI as one of the most important means of making data available[12]. Recent Dutch initiatives have included the hosting of a workshop Linked Data for Linguistic Research [13]. In addition the Dutch Cornetto database[14] and Open Dutch WordNet (ODWN)[15] are both available as RDF. The Linked Open Data paradigm is also important for current CLARIN discussions respecting so-called Resource Families[16] where it could become a core means of publishing lexicons. Linked Data has been also adopted in the ELEXIS e-lexicography infrastructure[17] for facilitating linking and publishing of lexicons on the Web. Given the importance of Linked Data for lexicons then, it is clear that a tool like LexO could play a vital role in the editing and visualisation of such resources.

Acknowledgements

This work has been conducted in the context of the cooperation agreement between Guido Mensching, director of the DiTMAO project at the Seminar für Romanische Philologie of the Georg-August-Universität Göttingen, and the Istituto di Linguistica Computazionale "A. Zampolli" of the Italian National Research Council. The authors have also been supported by the EU H2020 programme under grant agreements 731015 (ELEXIS - European Lexical Infrastructure).

References

Bel, N., Espeja Sergio, M. M. and Villegas, M. 2008. Coldic, a Lexicographic Platform for LMF Compliant Lexica. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

Bellandi, A., Giovannetti, E., and Weingart, A. 2018. Multilingual and Multiword Phenomena in a *lemon* Old Occitan Medico-Botanical Lexicon. *Information*, 9(3):52.

Fiorelli, M., Lorenzetti, T., Pazienza, M. T., and Stellato, A. 2017. Assessing VocBench Custom Forms in Supporting Editing of Lemon Datasets In International Conference on Language, Data and Knowledge (pp. 237-252). Springer, Cham.

Kenter, T., Erjavec, T., Dulmin, M. V., and Fier, D. 2012. Lexicon Construction and Corpus Annotation of Historical Language with the Cobalt Editor. In Proceedings of the 6^{th} Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH '12, pp. 16.

Khan, F., Bellandi, A., and Monachini, M. 2016. Tools and Instruments for Building and Querying Diachronic Computational Lexica. In Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH).

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. 2017. The OntoLex-Lemon Model: Development and Applications. In Proceedings of eLex 2017 conference, September (pp. 19-21).

McCrae, J. P., de Cea, G. A., Buitelaar, P., Cimiano, P., Declerck, T., Gmez-Prez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. 2012. Interchanging lexical resources on the Semantic Web. In Proceedings of Language Resources and Evaluation, 46(6), pp. 701709.

Ringersma, J. and Kemps-Snijders, M. 2007. Creating Multimedia Dictionaries of Endangered Languages Using LEXUS. In Proceedings of Interspeech 2007. Baixas, France: ISCA-Int.Speech Communication Assoc., pp. 6568.

Szymanski, J. 2009. *Wordventure developing wordnet in wikipedia-like style*.

[12] Jan Odijk, personal communication

[13] `https://www.clariah.nl/en/new/blogs/linked-data-for-linguistic-research#tuesday-7-february-2017`

[14] Cornetto: `https://portal.clarin.nl/node/1944`

[15] `http://wordpress.let.vupr.nl/odwn/`

[16] `https://www.clarin.eu/resource-families`

[17] `https://elex.is/`