David Lindemann

TEACHING TERMINOLOGY THROUGH WIKIBASE AND WIKIDATA

Abstract This software demonstration presents approaches to employ Wikibase in a university course on Terminology, and results of terminology projects lead by students. Wikibase, an extension of MediaWiki, is the software that underlies Wikidata, a very large crowdsourced queriable knowledge graph. We use an own Wikibase instance for cloud-based collaborative student projects on Terminology in Basque, a European minority language spoken in Spain and France, and co-official in some regions of Spain. We show that the course work with and about this software covers a great deal of what a course in Terminology for students in the translator and interpreters training should cover; it embraces relevant facets of knowledge engineering, and the acquisition of the features of a graph database and its user interface. The datasets created in student projects and partially shared through Wikidata are examples for a low-barrier crowdsourcing terminology workflow, eventually useful in other contexts as well.

Keywords terminology; teaching; knowledge graphs; Wikibase; Wikidata

1. Introduction

Understanding structures of terminological databases is regarded a fundamental skill in Translation Studies (see, for example, Cabré, 2000), and knowing to implement workflows for the creation of digital terminological resources might be necessary for developing a terminological Bachelor Final Project, while it certainly is indispensable for Master Final Projects in Terminology.[1] At the same time, we believe that challenging students with problems of practical terminology work motivates them to dig into multiple theoretical aspects of Terminology like conceptual relations, word formation, neology, and linguistic purism, and into the representation of knowledge in formal models in general. Also, in the attempt to write about their own endeavours, students need to acquire adequate meta-terminology, which we believe they will retain better when it has immediate application in an own project.

The Terminology course we offer for Translation Studies Bachelor students is introduced by providing fundamental concepts and problems of Basque and general Terminology Science,[2] but then deals with basic lexicographical and terminological resource structures, including formal representations of lexical-semantic relations, showing examples for the different kinds of lemma-focused

[1] See, for example, the module description of the Online Master in Terminology at UPF, Barcelona: https://www.upf.edu/web/terminologiaonline/memoria-de-master.

[2] We do not have space here for listing a comprehensive bibliography. Basque Terminology Science is heavily inspired by Catalan publications by Cabré et al.; some of them have been translated.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

and concept-focused[3] terminological resources: LSP dictionaries, thesauruses, taxonomies, term banks and wordnets, in human-readable distribution formats, and as structured datasets in machine-readable representation formats. It is shown how the models applied in structured datasets find their counterparts in the layout given to the representation shown to the human end user. This relation is looked at from two perspectives: How a resource entry structure is translated into entry layout features, and how layout features could be used to define a schema when trying to structure a resource entry text. In fact, the latter is what students will be asked to do in their own projects: To look out for an existing resource containing Basque terminology for the domain of their choice, and to work on that, with the aim of integrating the Basque terms, and, eventually, their descriptions, in the Wikidata Knowledge Graph. They receive advice for defining an appropriate workflow. We also offer to employ an own script that extracts concepts appearing in a custom set of Wikipedia articles. This simple approach has turned out to be valuable for a range of projects.[4] Alternatively, students might choose existing resources that contain terms in other languages as starting point; here, the mission is to define Basque equivalents for the described concepts, by examining specialized text written in Basque, or Basque terminological resources formerly not connected to the object of study.

All students know and heavily use Wikipedia in their homework and for documentation in their translation tasks, and practical translation courses at UPV/EHU that involve the Basque language address Wikipedia content creation by translation (most often from Spanish to Basque), which forces them to get along with requirements and editing procedures given on that platform,[5] produces the satisfaction of having contributed to a globally known resource, and lowers the barrier for subsequent contributions.[6] This is one strong reason for also looking at Wikidata (Vrandečić, 2012), Wikipedia's sister project at the Wikimedia Foundation, which started more than a decade ago as the structured, machine-readable counterpart of the human-centered and only roughly structured Wikipedia, but soon included also content from other sources as well, and which is now regarded one of the largest free knowledge graphs on the web, with significant conceptual coverage of important domains of knowledge (see e.g., Waagmeester et al., 2020). At the same time, the set of lexical-semantic relations represented in the Wikidata Knowledge Graph include all those found in the other lexical-conceptual resources presented, and Wikidata entities very often also include "external id" statements

[3] We use the term *lemma-focused* for resources such as dictionaries, which use a lemma list as macrostructure, in opposition to concept-focused resources, which model concepts as nodes that relate to each other, like in a hierarchized thesaurus or a knowledge graph, and that are annotated with equivalent terms in one or more languages, i.e., the terms that are used in texts when referring to the concept. When comparing viewpoints and workflows in specialized lexicography, on one side, and terminology, on the other, we follow Costa (2013).

[4] The script returns a list of Wikidata entities extracted from the blue-coloured hyperlinks in Wikipedia article text. After deleting concepts irrelevant to the project from that list, these are cloned on the local Wikibase instance, where they can be enriched with missing Basque equivalents, and other annotations.

[5] Translators often have to deal with cloud-based software platforms designed by the client, and clients expect them to be fast and flexible in that sense. This aspect, in our opinion, is still not sufficiently acknowledged in Translation Studies curricula.

[6] The author wants to thank Wikimedia Basque Country for their support in translation courses, providing a dedicated training library, and delivering tutorial seminars.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

to those other resources on entity level.[7] This and the fact that editing Wikidata requires not more than a Wikimedia user account makes it interesting to aim at enriching Wikidata as a student project.

As for the basic meta-terminology to use in the course, students are introduced to terms used in the description of term banks following Fontenelle & Rummel (2014) and of relevant representation formats following Reineke & Romary (2019), that is at its core, to describe *terms* as lexical units of a language that occur in specialized text, and that denote *concepts* as part of a specialized domain; if seen the other way round, that is, looking at a terminological resource, terms appear as lexicalizations or *labels* of concepts, which are connected to each other through typed relations. This is the way concepts are modelled on a Wikibase.[8]

2. Student Projects

In most cases, the sources chosen by the students will be unstructured data, i.e., a PDF publication, such as one of the specialized Basque dictionaries of the UZEI series,[9] or, for example, a list of recipes and a glossary listed on a Basque cooking website. On the other side, if the chosen source is an established structured terminological resource, that resource will most probably be well covered with concept labels and textual definitions in English and other major languages, while the coverage of Basque equivalents to these will present gaps. The students' aim, accordingly, will be to fill these with data obtained in their own terminological project.

All source datasets are converted to Wikibase format, that is, concepts are represented as Wikibase *items*, i.e., as nodes in a graph, annotated with preferred and alternative labels and definitions (in multiple languages), and with conceptual and other relations as edges – the Wikibase *properties*. If the source is Wikidata, the entity subset to work on is cloned on the local Wikibase. This conversion is done by the teacher and made transparent to the involved students.

Students are familiarized with the Wikibase online interface, and with the Open Refine software application,[10] which they install locally. The first step is to try to match the concepts of the students' concept schemes to Wikidata, using the application. For example, if the source to work with is a Basque-Spanish specialized dictionary, matching concepts can be looked for on Wikidata using the Basque and the Spanish term in a label-based search, eventually restricted to instances of a class (e.g., "instance of color", for color names). After that, regardless of whether projects have started with an external resource uploaded to Wikibase, or with a subset of Wikidata, all

[7] For example, alignments to Princeton WordNet are represented using the property http://www.wikidata.org/entity/P8814, and those to the EuroVoc thesaurus using http://www.wikidata.org/entity/P5437.

[8] In addition, on a Wikibase, terms may be represented as *lexemes* (lexical entries), which allows for a richer linguistic or lexicographic description, following Ontolex-Lemon. See https://eneoli.wikibase.cloud for a recent example of describing terms as Wikibase lexemes.

[9] The UZEI dictionaries are available as PDF at https://www.euskadi.eus/hiztegiak-banku-terminologikoak-eta-entziklopediak/web01-a2eutres/eu/#9588.

[10] See https://openrefine.org.

students are at a similar point: They have a set of Wikibase entities, some or all of them aligned to Wikidata entities, and the labels, definitions, and, eventually, the relations contained as statements in Wikidata. Now, missing labels, descriptions, and relations can be added or updated. Basque concept lexicalizations proposed by students are to be furnished with usage attestations, which are referenced as Wikibase *reference* using an external URL or a Wikibase entity describing the cited resource. Concept relations to be worked on are most typically "instance of", "kind of" a.k.a. "subclass of", and "part of",[11] but use cases may involve others (such as "used by"[12] for a scheme of social media concepts, or "has part" for a cooking concept scheme).[13]

Manual and batch edits to Wikibase entity data are tracked in the reversible Wikibase edit history, and collaborative work may involve the use of entity discussion pages and/or user talk pages. Furthermore, students use self-created Wikitext pages to describe their project; the teacher supports the process providing SPARQL queries tailored for each use case. Most typically, such query will list existing (Wikidata) item labels and descriptions, and their counterparts as they are described on the local Wikibase in that point in time. After a collaborative review process, project results can be uploaded to Wikidata. The uploaded datasets in all cases contain Basque concept lexicalizations and descriptions, and in some, also concept relations and entities to be created as new on Wikidata.

2.1 Example Project on Climate Change

The documentation of student projects carried out so far are accessible from the home page of our Wikibase instance, in Basque language.[14] As example for a student project, we present here a recent piece of work presented as Bachelor Final Project in Basque, and briefly described in English:[15] Paula Garay's concept scheme on Climate Change terminology. The workflow can be described as follows:

1. Relevant authority sources are identified, in this case, glossaries released by government authorities, international organizations, and NGOs.[16]

2. The English terms described in these resources are uploaded to Wikibase as English labels for newly created concept ("Q") items, together with English definitions given in these, referencing each term and each concept definition to its source. Since referencing labels in Wikibase is not possible, the terms and the definitions are represented as separate statements as well;[17] this allows further descriptions of the lexical item considered an equivalent term in a language, using Wikibase qualifiers and references.[18]

[11] That is, the three fundamental concept relations according to ISO 25964-2, a norm dealing with the interoperability of terminological resources; see https://www.iso.org/standard/53658.html.

[12] On the own Wikibase, represented as https://eusterm.wikibase.cloud/entity/P50.

[13] See the ingredients of a lasagne on Wikidata at https://www.wikidata.org/wiki/Q20034#P527.

[14] See https://eusterm.wikibase.cloud.

[15] See https://eusterm.wikibase.cloud/wiki/GRAL_Paula_Garay.

[16] See exact references on the project page.

[17] For an example, see https://eusterm.wikibase.cloud/wiki/Item:Q6981#P94.

[18] See https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

3. In Open Refine, the English terms are used for reconciliation against Wikidata entities, and the Wikidata entity ID of matching candidates are added to the Wikibase entities.

4. Through a federated SPARQL query, a table is produced, which contains the source term and definition, and the definition of the matching candidate on Wikidata. In this step, the student validates the alignment of Wikibase and Wikidata entity, and updates it if, looking at the English concept description found on Wikidata, the alignment is not regarded an exact match.

5. Wikibase entities that have been linked to one and the same Wikidata entity are merged. These are those entities that describe terms stemming from different sources but that describe the same concept. Consequently, these are then described on the same Wikibase entity page.[19]

6. The resulting alignment contains 376 concepts, from which 200 had a Basque label on Wikidata (62 had Basque definitions). These have been imported to Wikibase, together with labels in four other languages that the student understands (German, Spanish, Catalan, and French), the links to access Wikipedia articles describing the concept in these languages, and statements (property-value pairs) for the following properties, if present on Wikidata: *instance of, part of, subclass of*.

7. The concept labels imported from Wikidata are reviewed. In this context, the function of Wikidata labels is discussed, which is not fully aligned to terminology criteria; more precisely, "alternative labels" on Wikidata are often not synonyms of the "preferred label", that is, they are not interchangeable terms. Wikidata labels are chosen for the retrieval of items through text-based entity searches; "alternative" labels are, for example, found to be hyperonyms of the "preferred" label (Martín-Chozas et al., 2020).

8. Basque term gaps that the student has found candidates for, or term labels stemming from Wikidata found valid, are annotated with an authoritative source, and a context snippet. This is then represented as a statement to the Wikibase entity, and can be used as source for an upload to Wikidata, where the Basque term will henceforth serve as preferred or alternative concept label.

3. Summary

With the described exercises, students have been trained in using a specialized software for knowledge representation, dealing with problems on the lexical level (reviewing or assigning Basque lexicalizations to concepts), and on semantic level (reviewing or assigning semantic relations such as hyponymy or metonomy), having to deliver supporting evidence for every terminological decision from Basque scientific articles or other sources. They have also documented their project starting point, methodology, workflow, and results on a public page, which can be subject to

[19] See https://eusterm.wikibase.cloud/wiki/Item:Q9635 for an example entry, with English terms and definitions from different sources.

a collective review process. Last but not least, they have seen themselves empowered as contributors to a large open knowledge graph, where they have added data in their working language, for the benefit of all users.

References

Cabré, M. T. (2000). La enseñanza de la terminología en España: Problemas y propuestas. *Hermeneus: Revista de la Facultad de Traducción e Interpretación de Soria*, *2*, 41–94.

Costa, R. (2013). Terminology and Specialised Lexicography: Two complementary domains. *Lexicographica*, *29*(1), 29–42. https://doi.org/10.1515/lexi-2013-0004

Fontenelle, T., & Rummel, D. (2014). Term banks. In P. Hanks, & G.-M. de Schryver (Eds.), *International Handbook of Modern Lexis and Lexicography* (pp. 1–12). Springer. https://doi.org/10.1007/978-3-642-45369-4_21-1

Martín-Chozas, P., Ahmadi, S., & Montiel-Ponsoda, E. (2020). Defying Wikidata: Validation of Terminological Relations in the Web of Data. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 5654–5659). European Language Resources Association. https://aclanthology.org/2020.lrec-1.694

Reineke, D., & Romary, L. (2019). Bridging the gap between SKOS and TBX. *Edition – Die Fachzeitschrift* Für *Terminologie*, *2–2019*, 19–27.

Vrandečić, D. (2012). Wikidata: A new platform for collaborative data collection. *Proceedings of the 21st International Conference on World Wide Web* (pp. 1063–1064). https://doi.org/10.1145/2187980.2188242

Waagmeester, A., Stupp, G., Burgstaller-Muehlbacher, S., Good, B. M., Griffith, M., Griffith, O. L., Hanspers, K., Hermjakob, H., Hudson, T. S., Hybiske, K., Keating, S. M., Manske, M., Mayers, M., Mietchen, D., Mitraka, E., Pico, A. R., Putman, T., Riutta, A., Queralt-Rosinach, N., & Su, A. I. (2020). Wikidata as a knowledge graph for the life sciences. *eLife*, *9*, e52614. https://doi.org/10.7554/eLife.52614

Acknowledgements

The research leading to the presented results has been supported by DLTB research group (Basque Government, IT1534-2) and Monumenta Linguae Vasconum research project (MINECO, PID2020-118445GB-I00, Govt. of Spain).

Contact information

David Lindemann
UPV/EHU University of the Basque Country
david.lindemann@ehu.eus

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.