David Lindemann and Mikel Alonso

LINKING HISTORICAL CORPUS DATA AND ANNOTATIONS USING WIKIBASE

Abstract This software demonstration presents a data model and a first use case for the representation of text corpus data on a Wikibase instance, including morphosyntactic, semantic and philological annotations as well as links to dictionary entries. Wikibase, an extension of MediaWiki, is the software that underlies Wikidata (Vrandečić & Krötzsch, 2014), an exceptionally large crowdsourced queryable knowledge graph, which includes nodes for ontological concepts, on the one hand, and for lexemes, lexeme senses and lexeme forms, on the other, together with annotations to and relations between them. We argue that the proposed model and the chosen software solutions for the representation of corpus and dictionary data, all free and open source, meet with the requirements of provenance transparency, open access and re-use, and the capability of collaborative work on the data. We also present our own scripts wrapped in a web application that shortcut several workflow steps in a first use case, a 1737 Basque manuscript, transcribed on Wikisource, and represented as an annotated dataset on our Wikibase instance.

Keywords Basque; historical corpus; Wikibase; corpus annotations; Linked Data

1. Introduction

With regard to the digitisation of historical texts in Basque, we have seen many efforts in recent years; various initiatives using different objectives and methodologies. Among these projects, we can find *Klasikoen Gordailua*[1], a repository of classic Basque literature. Another project, *Corpus Historikoa*[2], is a POS-lemma annotated corpus partly based on the former, enhanced with an online interface, allowing lemma-based searches. The corpus contains classical texts from 1545, when the first book written entirely in Basque was published, to 2000. The most recent project is the *BIM-SAHCOBA* corpus (Estarrona et al., 2022)[3], a morphosyntactically annotated historical corpus which comprises the most significant works from the 15th century to the mid-20th century, when present-day standard Basque appeared. An interface allows searches based on the part of speech, on inflection form descriptions, lemmata, and neighbouring tokens and their metadata. An initiative with a totally different scope is the digital edition of the *Lazarraga Manuscript*[4] (Bilbao et al., 2011), a recently found 16th century text, published along with philological annotations. It is our aim to propose a model which would combine contents of the cited versions of the Basque Historical Corpus, in a way that the user could also follow each corpus token back to its original source, e.g., a manuscript or early modern print facsimile image collection.

[1] Accessible at https://klasikoak.armiarma.eus/

[2] Accessible at https://www.ehu.eus/etc/ch/

[3] Accessible at http://bim.ixa.eus/

[4] Accessible at https://www.ehu.eus/monumenta/lazarraga/

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.

We have presented experiments to represent and interrelate historical and standard lexicographic data in Basque as Linked Data (Lindemann & Alonso, 2021), in which we have integrated a lexicographical dataset into the Wikisource and the Wikidata platforms. The resulting connections between a Basque historical source and the Wikidata knowledge graph, more precisely, connections between entries in a classic dictionary (Larramendi, 1745) on Wikisource and lexical entries (entities of type lexeme) describing standard Basque lemmata, can be manually or programmatically queried for, accessed and edited.

This software demonstration presents a data model and a first use case for the representation of contents of text corpus data on a Wikibase instance, including the types of annotations found in the above cited digital versions of historical Basque text, i.e., morphosyntactic, semantic and philological annotations, as well as links to dictionary entries.

2. Technical Considerations

Proposing the model described here, it is our goal to ensure interoperability with widely used standards for corpus annotation and lexical data representation. Token annotation properties pointing to literal values on Wikibase can be straightforwardly declared equivalent to the NIF ontology (Hellmann et al., 2013), and lexical data on Wikibase is modelled following Ontolex-Lemon (McCrae et al., 2017). On the other hand, we are not aware of established standards regarding the interface between the corpus and the lexicon, and, more precisely, regarding properties that link text tokens and token spans to dictionary entries at lexeme (lemma, headword), sense, and form level, that is, in all these cases and unlike according to NIF, linking not to literals but to entities (nodes on the graph).

As for the software solution to deploy for the implementation of the model and subsequent population with data, we choose Wikibase, which is the only free graph database software we are aware of that features user management, editable entity data exhibition pages, reversible edit histories, and graphical and programmatical SPARQL query endpoints.

The use case for which we propose this model is documents that belong to the Basque Historical Corpus, although given the general character of the model we claim that it can serve in other contexts, too[5]. That corpus contains literature written in Basque from before 1900, and today it exists in the different versions referenced above, and stored in separated and incompatible data silos (based on relational databases) and made available through different online user front ends. In this presentation, we focus on the corpus-lexicon interface, while our model also should be as flexible as to allow associating source text tokens with their representation in different editions of the source[6], so that different versions of the text could be displayed, together with all kinds of annotations, using the dataset stored on Wikibase.

[5] A recent experiment following the proposed model and involving Serbian corpus data modelled according to NIF and an Ontolex dictionary is documented at https://serbian.wikibase.cloud.

[6] We think of a property pointing from a token or token span to a string representation of the token, qualified with a property pointing to the corresponding edition.

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. Data Model and Implementation

Heavily inspired by the latest trends in the field of Linguistic Linked Open Data[7], we model a corpus token as node in a knowledge graph, and link it to the following data objects (cf. Figure 1):

1. the respective paragraph in the source document (that is, for example, a transcription carried out and stored on the Wikisource platform);
2. a lexeme node, which is annotated with the standard Basque lemma (we call this property *has standard lexeme*);
3. a lexical form associated to that lexeme, which is annotated with the grammatical features describing the form (*has standard form*);

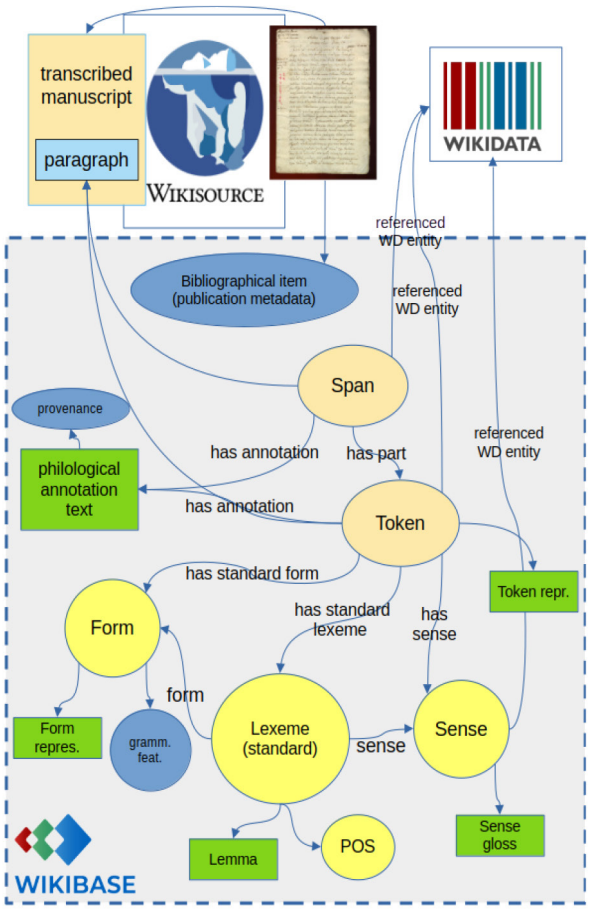

Fig. 1: Proposed model for the corpus-lexicon interface

4. a lexical sense associated to that lexeme, which is annotated with a sense gloss (we name it *has sense*, since here the distinction is not standard vs. non-standard but whether a sense is to be flagged as historical (i.e., in use in some past period);

[7] For recent related work, see Stanković et al. (2023).

5. an ontology concept that matches to the word sense. Note that the redundancy in the properties named *referenced WD entity*, with both token and dictionary sense as subject, is intended: There will be cases where a dictionary sense is already aligned to an ontology concept[8], and there will be others, where automatic or human annotators link a text token to a named entity, which is referred to using a Wikidata item ID. Values of this property attached to either the token or to the dictionary sense can be inferred from each other.

6. to a text chain containing philological annotations, such as, in our use case: *instead of something crossed out, as side note, hardly decipherable,* etc.

Furthermore, we represent token spans as separate nodes; these are linked to the contained tokens, and to annotations that apply to the whole span.

We implement and populate the model on our own Wikibase instance[9] hosted on Wikibase Cloud[10]. We chose that software for a range of reasons:

1. The software itself and the chosen hosting solution are free and open-source software, with an active community dealing with its development and user support.

2. Uploaded contents are instantly exhibited in a graphical user interface which allows for collaborative editing of every single claim made on the corpus token, including reversible edit histories.

3. The created dataset is publicly queryable using SPARQL.

4. We are able to export the dataset as an RDF/OWL document, as far as classes and properties on the Wikibase are aligned with external RDF vocabularies[11].

5. The dataset is compatible with Wikidata, in the sense that lexemes, forms, senses, their descriptions, and ontology concepts can be aligned between our own Wikibase and Wikidata, which in the Wikibase Ecosystem is called *federation*[12], so that it is possible to launch queries that deliver relations from both graphs at the same time[13].

6. Core classes and properties used on a Wikibase by default for describing lexemes deploy Ontolex-Lemon, the W3C-recommended model for lexical data, so that the created datasets are compatible with Wikidata and other resources deploying that model; a relation between entities of the same classes in two different datasets or Wikibase instances can be straightforwardly set.

[8] According to Ontolex-Lemon, that link is named *ontolex:reference*. On Wikidata, it is http://www.wikidata.org/entity/P5137.

[9] Accessible at https://monumenta.wikibase.cloud.

[10] Wikibase Cloud is a service provided by Wikimedia Deutschland, see https://wikibase.cloud.

[11] See a Wikibase property declared equivalent to a NIF property at https://monumenta.wikibase.cloud/wiki/Property:P147#P179.

[12] See https://meta.wikimedia.org/wiki/LinkedOpenData/Strategy2021/Wikibase.

[13] For example, a standard Basque lexeme on Wikidata is linked to additional sources where it is described; Wikidata also contains further knowledge about named entities such as birth dates for persons or geographic coordinates for locations.

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.

4. Population of the Model: A First Use Case

We are currently populating the proposed model with tokens from a 1737 Basque manuscript, the transcription of which has been carried out on Wikisource[14], and inserting annotations of the above described types including philological annotations by Lakarra (1985), and own annotations. The Wikibase items created to describe each token link back to the corresponding paragraph in the manuscript transcription on Wikisource, which is on that platform accessible together with digital images of the source manuscript.

All annotation operations, including those pointing to dictionary entries, forms, or senses, can be set manually in the Wikibase interface. In order to shortcut the operations, we have defined functions in a script that responds to the user via a web application[15]. Tasks addressed by these functions are token entity creation, token span definition, dictionary linking, and text annotation display.

5. Summary

We have proposed a model for a corpus-lexicon interface, implemented it on a Wikibase instance, and started to populate it with a Wikisource transcription as the first textual source. We also have produced scripts that ease the different workflow steps, more precisely, a web application that allows a user to define token spans, and to add annotations the model allows for on token or span level. We want to point out the advantages of having chosen Wikibase as working platform: All tokens and token spans have their own entity data page online, where annotations can be reviewed and edited. All edits are recorded in the edit history of each token or token span item, on the one hand, and in the user contribution records, on the other. That means the platform is ready for collaborative work on the data, e.g., in a larger project where different human editors are involved. Through the Wikibase API, annotations can also be set massively; this is useful when results stemming for example from a lemmatization or Named Entity Recognition tool shall be uploaded and then validated by human annotators.

Through SPARQL, advanced queries are enabled; these may involve all token or span annotations, their context, the source document metadata, and everything that can be inferred from the Wikidata alignments at different levels[16].

References

Bilbao, G., Gómez, R., Lakarra, J. A., Manterola, J., Monoule, C., & Urgell, B. (2011). *Lazarraga eskuizkribuaren edizioa eta azterketa.* Lazarraga Eskuizkribuaren Edizioa Eta Azterketa, Vitoria-Gasteiz: UPV-EHU. https://www.ehu.eus/monumenta/lazarraga/

[14] Accessible at https://eu.wikisource.org/wiki/Azkoitiko_Sermoia.

[15] We are using a python Flask webapp, which calls functions involving the wikibaseintegrator python module for communication with Wikibase. See https://monumenta.wikibase.cloud/wiki/Project:About for reference of the code repositories.

[16] See https://monumenta.wikibase.cloud/wiki/Larramendi,_Azkoitiko_Sermoia.

Estarrona, A., Etxeberria, I., Soraluze, A., Etxepare, R., & Padilla-Moyano, M. (2022). The first annotated corpus of historical Basque. *Digital Scholarship in the Humanities*, *37*(2), 391–404. https://doi.org/10.1093/llc/fqab066

Hellmann, S., Lehmann, J., Auer, S., & Brümmer, M. (2013). Integrating NLP Using Linked Data. In H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty, & K. Janowicz (Eds.), *The Semantic Web – ISWC 2013* (pp. 98–113). Springer. https://doi.org/10.1007/978-3-642-41338-4_7

Lakarra, J. A. (1985). Literatur gipuzkerarantz: Larramendiren Azkoitiko sermoia (1737). *Anuario del Seminario de Filología Vasca 'Julio de Urquijo'*, *19*(1), 235–281. https://doi.org/10.1387/asju.7685

Larramendi, M. (1745). *Diccionario trilingüe castellano, bascuence y latin dedicado a la M.N. y M.L. provincia de Guipuzcoa.* Bartholomé Riesgo y Montero. https://www.kmliburutegia.eus/Record/203133

Lindemann, D., & Alonso, M. (2021). A workflow for historical dictionary digitisation: Larramendi's Trilingual Dictionary. *Electronic Lexicography in the 21st Century: Proceedings of the eLex 2021 Conference*, (pp. 598–614). https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_39_pp598-614.pdf

McCrae, J., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (Eds.), *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017* (pp. 587–597). Lexical Computing CZ s.r.o. https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf

Stanković, R., Chiarcos, C., Utvić, M., & Kitanović, O. (2023). Towards ELTeC-LLOD: European Literary Text Collection Linguistic Linked Open Data. In S. Carvalho, A. F. Khan, A. O. Anić, B. Spahiu, J. Gracia, J. P. McCrae, D. Gromann, B. Heinisch, & A. Salgado (Eds.), *Proceedings of the 4th Conference on Language, Data and Knowledge* (pp. 180–191). NOVA CLUNL, Portugal. https://aclanthology.org/2023.ldk-1.16

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A Free Collaborative Knowledge Base. *Communications of the ACM*, *57*, 78–85. https://doi.org/10.1145/2629489

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

Mikel Alonso Arrospide
UPV/EHU University of the Basque Country
mikel.alonsoa@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.