Lexemes in Wikidata: 2020 status

Finn Årup Nielsen
DTU Compute, Technical University of Denmark
Richard Petersens Plads, Kongens Lyngby, Denmark
faan@dtu.dk

Abstract

Wikidata now records data about lexemes, senses and lexical forms and exposes them as Linguistic Linked Open Data. Since lexemes in Wikidata was first established in 2018, this data has grown considerable in size. Links between lexemes in different languages can be made, e.g., through a derivation property or senses. We present some descriptive statistics about the lexemes of Wikidata, focusing on the multilingual aspects and show that there are still relatively few multilingual links.

Keywords: Wikidata, lexicographic data, Linguistic Linked Open Data

1. Introduction

Wikidata is the structured data sister of Wikipedia where users can collaboratively edit a knowledge graph (Vrandečić and Krötzsch, 2014). Wikidata does not only support the different language versions of Wikipedia but also the other Wikimedia wikis such Wikisource, Wikimedia Commons, Wikiquote, etc. as well as describe many items without any equivalent article in the other wikis. For instance, Wikidata describes tens of millions of scientific articles (Nielsen et al., 2017). The data in Wikidata is converted to a Semantic Web representation (Erxleben et al., 2014) and a public and continuously updated SPARQL endpoint—*Wikidata Query Service* (WDQS)—is set up at `https://query.wikidata.org`.

Since 2018, Wikidata has included special pages for lexicographic data distinguished from the usual Wikidata "Q-items" with a new namespace for lexemes. Each page represents one lexeme, its sense(s) and its lexical form(s) together with annotation about them and links between them, both within and between lexemes as well as to the Q-items. The lexicographic data is also converted to a Semantic Web representation and available in WDQS. For the RDFication of the lexeme data, Wikidata uses a combination of classical Wikidata URIs and URIs from (Linguistic) Linked Open Data ontologies (Cimiano et al., 2016; McCrae et al., 2017): `ontolex:lexicalForm`, `ontolex:sense`, `ontolex:LexicalEntry`, `ontolex:LexicalSense`, `ontolex:Form` and `dct:language` as well as other URIs, e.g., `dct:language` and `wikibase:lemma`.

We have described the lexicographic information on Wikidata before focusing on the Danish lexemes (Nielsen, 2019a) and also described our SPARQL-based Web application *Ordia* for aggregating and visualizing the Wikidata lexicographic data (Nielsen, 2019b). Here we will make an update of the work on lexemes in Wikidata and focus on the multilingual aspects.

2. Descriptions

In February 2020, Wikidata had more than 77 million Q-items[1] Over 250,000 lexemes are in February 2020 available in Wikidata.[2] This is up from 43,816 we reported in 2019 (Nielsen, 2019a). In February 2020, there were over 3 million lexical forms and over 55,000 senses.

Chains	Count	Between-language count
1	3897	1453
2	1158	333
3	443	127
4	141	33
5	47	9
6	12	3

Table 1: Counts of level of etymological derivations (chains) per 23 February 2020. The last result is available in WDQS from `https://w.wiki/Htz`.

2.1. Languages

Lexemes from 668 languages are recorded in Wikidata.[3] However, many languages have only a single lexeme. The top language with most lexemes is Russian (101,137 lexemes), followed by English (38,122), Hebrew (28,278), Swedish (21,790), Basque (18,519), French (10,520) and Danish (4,565). Russian is also the language with more forms than any other language (1,236,456), followed by Basque (956,473), Hebrew (446,795), Swedish (148,980), Czech (77,747) and English (64,798). For senses, the languages from the top are Basque (20,272), English (12,911), Hebrew (3,845), Russian (2,292) and Danish (2,217).

2.2. Etymology

Etymological information may be described through the *derived from* property (P5191) corresponding to `lemonet:derivedFrom` from (Chiarcos et al., 2016). It has been used over 3,800 times, see Table 1. Apart from tracking derivations between different languages, the property may also be used to record intralanguage derivations. Table 1 shows statistics for the total number of derivations and the cross-language derivations by the derivation chain

[1] `https://www.wikidata.org/wiki/Special:Statistics`

[2] `https://tools.wmflabs.org/ordia/statistics/`

[3] `https://tools.wmflabs.org/ordia/language/`

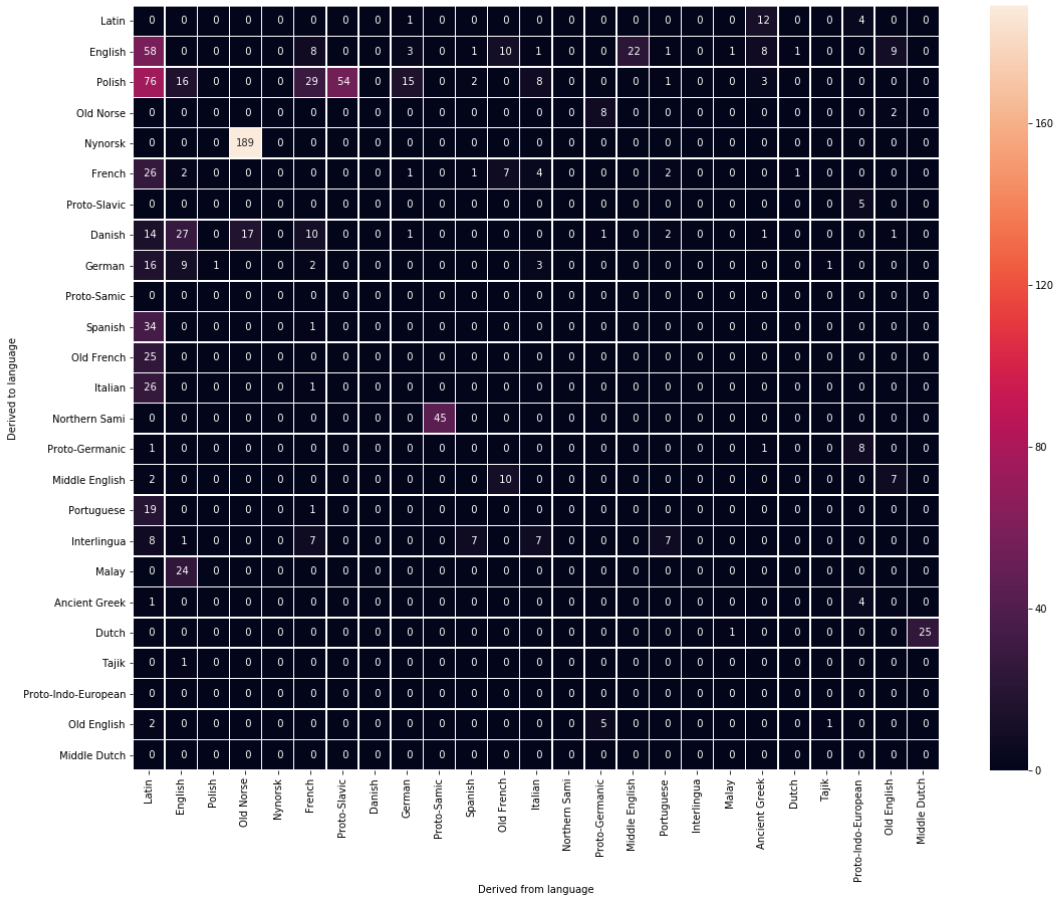

Figure 1: Derivation matrix: Count of the number of derived lexemes between languages as recorded in Wikidata and the Wikidata Query Service as of 28 February 2020.

length. The currently longest derivation chain is 6, and an example of a long between-language derivation is from the Afrikaans word *hond* (dog) through Dutch, Middle Dutch, Old Dutch, Proto-Germanic to Proto-Indo-European.

The etymological derivation matrix in Figure 1 shows the yet sparse between-language derivation data among the 25 languages with the most derivations. Most of the languages are Indo-European, though among the 25 are also Sámi languages, Malay and Interlingua. The largest number of recorded (direct) derivations is from Old Norse to Nynorsk, — but with just 189 links. Latin is the source language with the most derivations. A PageRank analysis in NetworkX of the directed and count-weighted derivation graph with $\alpha = 0.9$ presents Proto-Indo-European on the top, followed by Latin, Ancient Greek and English.

Derivations and compounding may also be described by the compound property (P5238). As of 25 February 2020, Danish (1,735), French (320), Polish (245) and English (197) are the languages which have used the property the most.[4]

[4]Counting distinct lexemes with WDQS with the SPARQL `?lexeme dct:language ?language ; wdt:P5238 [] .` with the result at https://w.wiki/J5x.

The etymological data in Wikidata is dwarfed by the amount that can be extracted from Wiktionary (de Melo, 2014).

2.3. Senses

Lexemes link to senses and a sense can link to senses in other languages. The two primary means are through the *translated to* property (P5972) that links to other senses or by the *item for this sense* (P5137) that links to a Q-item. As of 26 February 2020, the former property has been used 3,633 times, while the latter property has been used 25,891 times. Figure 2 shows of the number of translations via the *item for this sense* property for the 25 languages with the most translation links. The diagonal shows twice the number of synonym combinations for lexemes within each language. The current number of translations is much lower than what can be extracted from Wiktionary, see, e.g., (Sérasset, 2014, Table 4). Only the combination English-Hebrew has over 1,000 translations. While Basque is the language with the most senses defined, the senses of the language does not in a sufficient degree link further on to the Q-items to get among the 25 most linked languages that is shown in Figure 2.

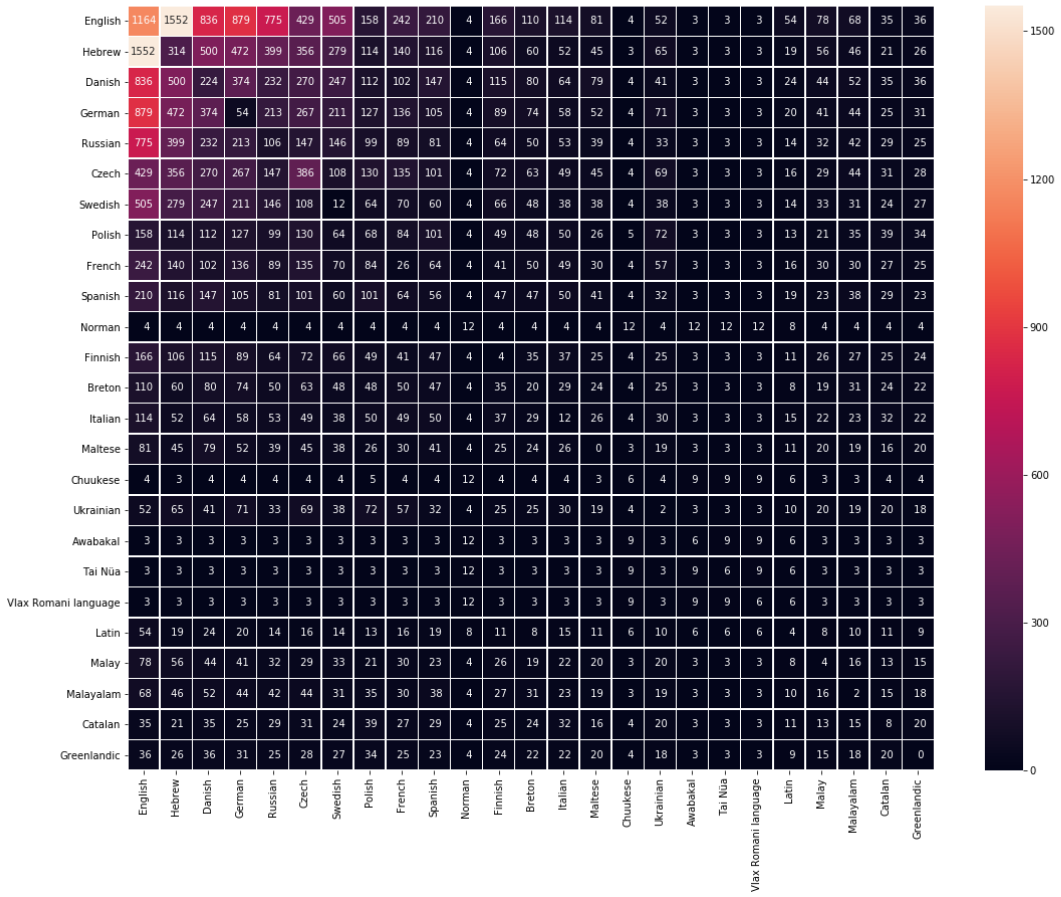

Figure 2: Sense-Q-item links between languages among lexemes in Wikidata. The diagonal shows twice the number of synonym combinations for lexemes within each languages. The data has been extracted with a SPARQL query that contains the following fragment: `?lexeme1 dct:language ?language1 ; ontolex:sense / wdt:P5137 ?item . ?lexeme2 dct:language ?language2 ; ontolex:sense / wdt:P5137 ?item.`

A different way to link senses to Q-items is by the *demonym of* (P6271) property that is only relevant to use for demonyms. It does not link to the sense of the demonym, but rather to the sense of the region associated with demonym, e.g., from the French lexeme *parisienne* (L25620) to the Q-item for Paris (Q90). Figure 3 shows the demonym matrix where the Spanish-Danish language pair has the largest number of links between demonyms.

There are a number of other properties that link sense-to-sense within language, e.g., *hypernym*, *troponym of* and a seldom used *periphrastic definition* property.

2.4. External identifiers

Wikidata has numerous deep links to items in external databases through the properties with the *external identifier* datatype. There are currently 4.789 recorded properties of this type.[5] A few of these relates to the lexicographic items, potentially making Wikidata a multilingual hub for lexicographic resources. The statistics page in *Ordia* at `https://tools.wmflabs.org/ordia/statistics/` shows statistics on 19 linguistics external identifiers. Only 11 of these have currently more than 100 links and they are shown in Table 2. The Elhuyar identifier for a Basque online dictionary has by far the most identifiers. The second most frequent identifier is for words within the Danish wordnet DanNet (Pedersen et al., 2009), and then follows several identifiers for the Polish language. The Greenlandic Oqaasileriffik online dictionary records both Greenlandic, Danish and English lexemes.

Apart from these identifiers, Wikidata has identifiers to link its Q-items to BabelNet (P2581) and for the Interlingual Index Identifier (P5063) (Navigli and Ponzetto, 2010; Bond et al., 2016). They receive 61,378 and 31 links, respectively.

2.5. Other linguistic data in Wikidata

Wikidata can describe linguistic resources and use them to annotate lexemes. Datasets, corpora and dictionaries may have entries in Wikidata. *Ordia* shows resources that have been used in the usage examples for the lexemes of Wiki-

[5] `https://www.wikidata.org/wiki/Category:Properties_with_external-id-datatype`

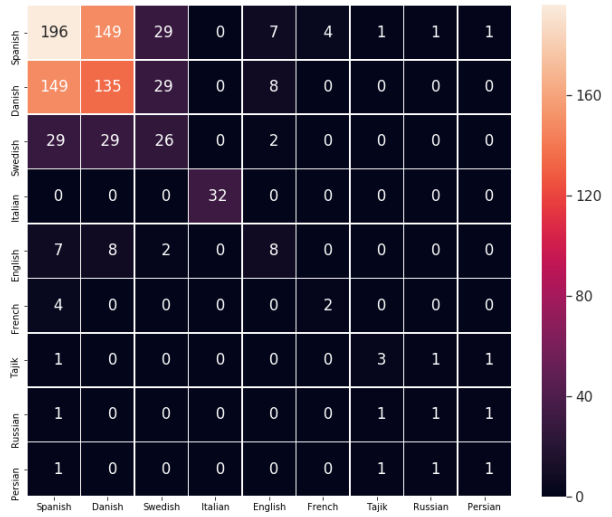

Figure 3: Sense-Q-item links between languages among lexemes in Wikidata with the *demonym for* property. The diagonal counts the number of distinct lexemes per language with demonym senses.

Count	Identifier	Language(s)
14440	Elhuyar	Basque
2878	DanNet word	Danish
1688	WSO Online	Polish
1353	SJP Online	Polish
1288	Doroszewski	Polish
1027	Dobry słownik	Polish
1009	WSJP	Polish
388	Oqaasileriffik	Greenlandic, Danish, English
216	Vocabolario Treccani	Italian
212	OED Online	English
160	Kopaliński	Polish

Table 2: External identifiers in Wikidata sorted according to usage per 22 February 2020. Updated statistics is available at https://tools.wmflabs.org/ordia/statistics/

data.[6] The *National Corpus of Polish* (Q6971865) and the *Europarl* (Q5412081) corpus (Koehn, 2005) are the two resources that have been used the most.

Wikidata's Q-items may link to lexeme items with the *subject lexeme* (P6254). 824 distinct Q-items makes 831 links in total. Most of these Q-items describe Wiktionary pages for French conjugations. A few other items describe scientific papers that focuses on particular lexemes, e.g., the new Swedish pronoun *hen* discussed in (Tavits and Pérez, 2019).

[6] https://tools.wmflabs.org/ordia/reference

3. Discussion

The number of lexeme data in Wikidata continuous to grow, but in many aspects the extent is still low and the annotation for etymology and senses is meager. Russian lexemes and forms are exceptions. They have been automatically set up from the Russian Wiktionary with the Lexicator tool.[7] Wikidata requires the permissive Creative Commons Zero license for its data and this may have prohibited the set up of lexicographic data from other resources, including share-alike-licensed Wiktionary.

What might also have held the sense data growth back is the unresolved issue of linking non-noun lexemes. Q-items in Wikidata usually corresponds to common or proper nouns, — at least their labels are usually nouns. The question is how lexemes corresponding to verbs, adjective and adverbs should be linked. Take the example of the English adjective *little*: Should a Q-item for *smallness* be created and the *little* lexeme linked to that item by the P5137 property, should there be a separate Q-item for *little* linked by P5137, or should the *little* lexeme be linked to a *smallness* by some other means? Wordnets may link lexicographic items across part-of-speech classes with, e.g., *derivationally related form* or *pertainym*.

The tool *Wikidata Lexeme Forms*,[8] that works for several languages, helps Wikidata lexeme editors create lexemes and their forms. We have set up several ShEx expression to detect errors of omission and commission or diversions from normal use for Danish lexemes (Nielsen et al., 2019). Such tools help Wikidata editors maintain a form of consistency and comprehensiveness within each language.

4. Conclusion

The lexicographic data in the lexeme part of Wikidata is yet not extensive in most aspects, but continuously grow. The most represented languages are Indo-European, particularly Slavic, Germanic and Romance languages. Links between lexemes of different languages can be established by an etymological property as well as through senses and the Q-items of Wikidata and links to external lexicographic resources can be established by several external identifier properties in Wikidata.

5. Acknowledgments

This work is funded by the Innovation Fund Denmark through the projects DAnish Center for Big Data Analytics driven Innovation (DABAI) and Teaching platform for developing and automatically tracking early stage literacy skills (ATEL).

6. Bibliographical References

Bond, F., Vossen, P., McCrae, J. P., and Fellbaum, C. (2016). CILI: the Collaborative Interlingual Index. *Proceedings of the Eighth Global WordNet Conference*, pages 50–57, January.

[7] https://github.com/nyurik/lexicator. Issues about what is copyrightable lexicographical data in the context of Wikidata has been discussed, see, e.g., https://meta.wikimedia.org/wiki/Wikilegal/Lexicographical_Data.

[8] https://tools.wmflabs.org/lexeme-forms/

Chiarcos, C., Abromeit, F., Fäth, C., and Ionov, M. (2016). Etymology Meets Linked Data. A Case Study In Turkic. *Digital Humanities 2016: Conference Abstracts*, pages 458–460.

Cimiano, P., McCrae, J. P., and Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report, 10 May 2016, May.

de Melo, G. (2014). Etymological Wordnet: Tracing The History of Words. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 1148–1154, May.

Erxleben, F., Günther, M., Mendez, J., Krötzsch, M., and Vrandečić, D. (2014). Introducing Wikidata to the Linked Data Web. *The Semantic Web – ISWC 2014*, pages 50–65.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *The Tenth Machine Translation Summit: Proceedings of Conference*, pages 79–86.

McCrae, J. P., Bosque-Gil, J., del Río, J. G., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Application. *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference.*, pages 587–597.

Navigli, R. and Ponzetto, S. P. (2010). BabelNet: building a very large multilingual semantic network. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, July.

Nielsen, F. Å., Mietchen, D., and Willighagen, E. (2017). Scholia, Scientometrics and Wikidata. *The Semantic Web: ESWC 2017 Satellite Events*, pages 237–259, October.

Nielsen, F. Å., Thornton, K., and Gayo, J. E. L. (2019). Validating Danish Wikidata lexemes. *Proceedings of the Posters and Demo Track of the 15th International Conference on Semantic Systems*, June.

Nielsen, F. Å. (2019a). Danish in Wikidata lexemes. *Proceedings of the Tenth Global Wordnet Conference*, pages 33–38.

Nielsen, F. Å. (2019b). Ordia: A Web application for Wikidata lexemes. *The Semantic Web: ESWC 2019 Satellite Events*, pages 141–146, May.

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L., and Lorentzen, H. (2009). DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43:269–299, August.

Sérasset, G. (2014). DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web: interoperability, usability, applicability*.

Tavits, M. and Pérez, E. O. (2019). Language influences mass opinion toward gender and LGBT equality. *Proceedings of the National Academy of Sciences of the United States of America*, 116:16781–16786, August.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57:78–85, October.