Lexemes in Wikidata: 2020 status

Finn Årup Nielsen

DTU Compute
Technical University of Denmark

22 June 2020

Wikidata

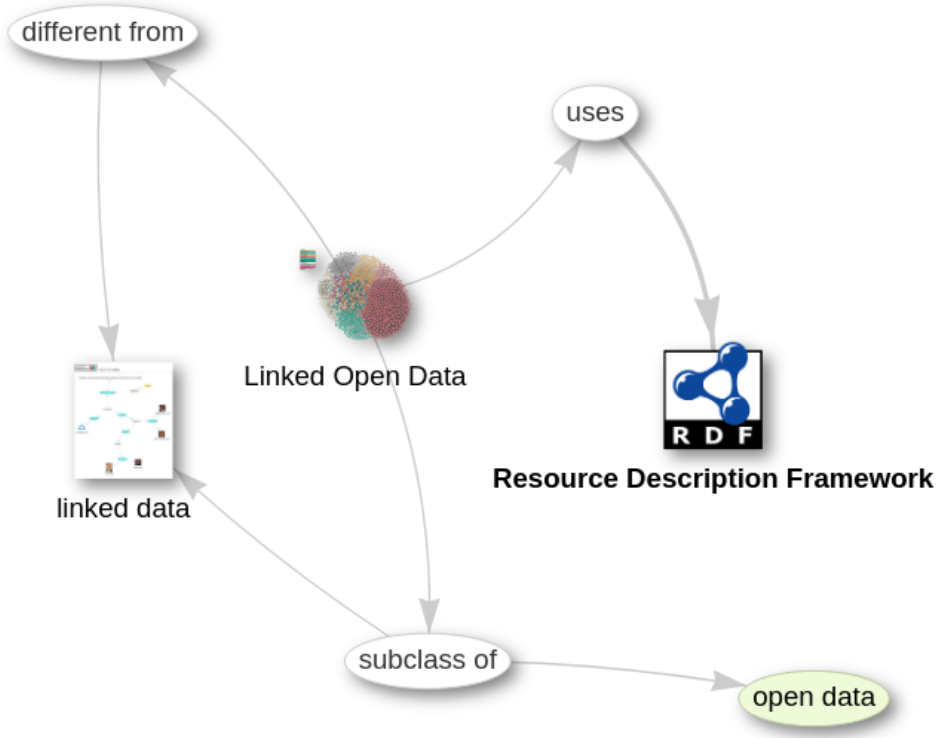


Figure 1: From Scholia at https://scholia.toolforge.org/topic/Q18692990.

Wikidata is the structured data sister of Wikipedia where users can collaboratively edit a knowledge graph (Vrandečić and Krötzsch, 2014).

Describes "Q-items": Wikipedia topics, scientific articles (Nielsen et al., 2017), researchers, artworks, etc.

SPARQL queryable via *Wikidata Query Service* (WDQS) at https://query.wikidata.org after conversion to a Semantic Web representation (Erxleben et al., 2014).

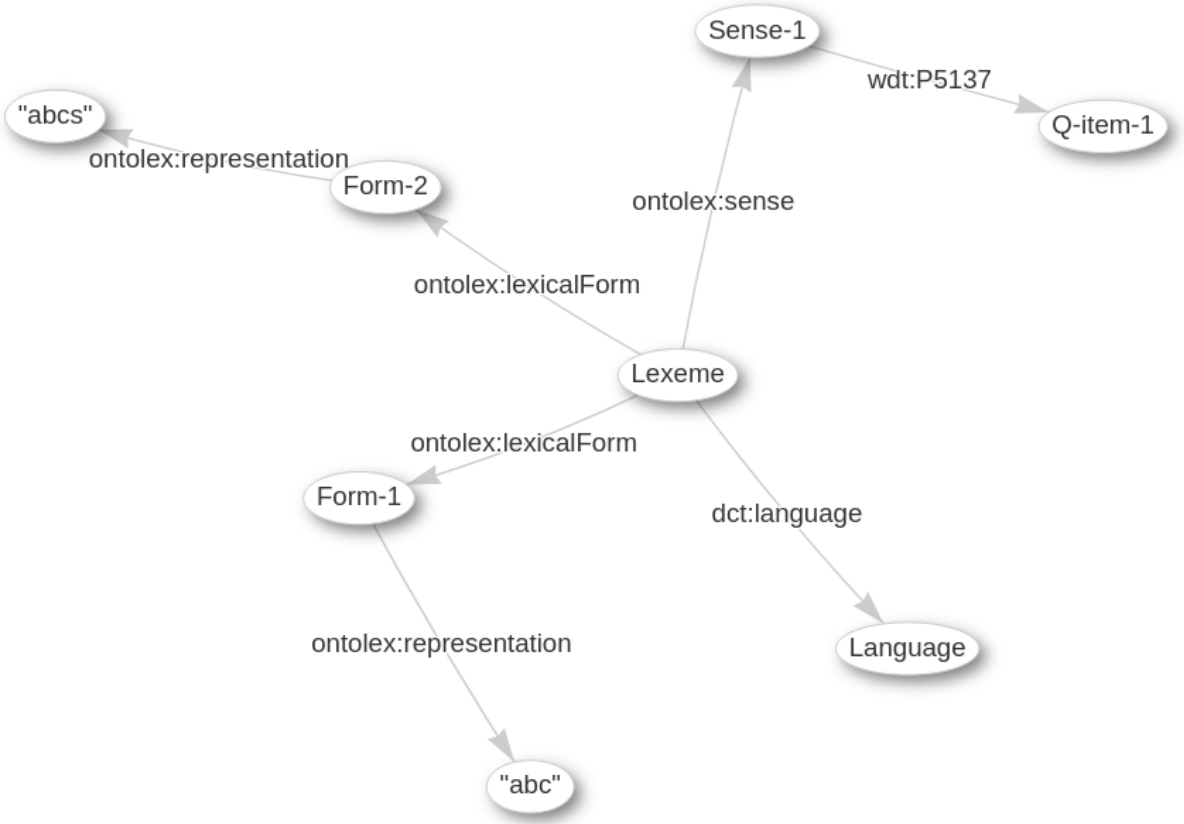
Wikidata lexemes

Since 2018, Wikidata has included special pages for lexicographic data: The lexeme pages prefixed with 'L' (cf. 'Q'-pages).

Here the Danish noun lexemes 'sengetøj' (L252976, bedding) that combines 'seng', '-e-' and 'tøj'.

Also queryable in WDQS
https://w.wiki/UW5

Wikidata lexemes



Lexemes in Wikidata are linked to their senses and forms.

Uses Wikidata and Linguistic Linked Open Data URIs, e.g., ontolex:sense and ontolex:lexicalForm.

Wikidata lexeme statistics

February	June	Description
77 mio	87 mio	Q-items
250,000+	298,888	Lexemes
3 mio+	4.9 mio	Forms
55,000+	69,361	Senses
668	704	Languages

Updated lexeme statistics available in Ordia (Nielsen, 2019) at https://ordia.toolforge.org/statistics/.

Wikidata lexeme language statistics

February	June	Description
101,137	101,128	Russian
38,122	40,754	English
28,278	28,282	Hebrew
21,790	22,413	Swedish
18,519	22,899	Basque
10,520	10,592	French
4,565	5,476	Danish
?	32,099	Latin
?	7,271	Czech
?	5,883	German

Number of lexemes in Wikidata per language in February 2020 and June 2020. Updated language statistics in Ordia at https://ordia.toolforge.org/language/.

Note the sudden rise of Latin from February to June.

Connection between language

How can Wikidata lexemes connect lexemes between two language so we can create bilingual dictionaries?

Via *derived from* property (P5191)

Via sense, the *item for this sense* (P5137) and Q-items

Via sense and the *translation* property (P5972)

Via sense and the *demonym* property (P6271)

Etymology

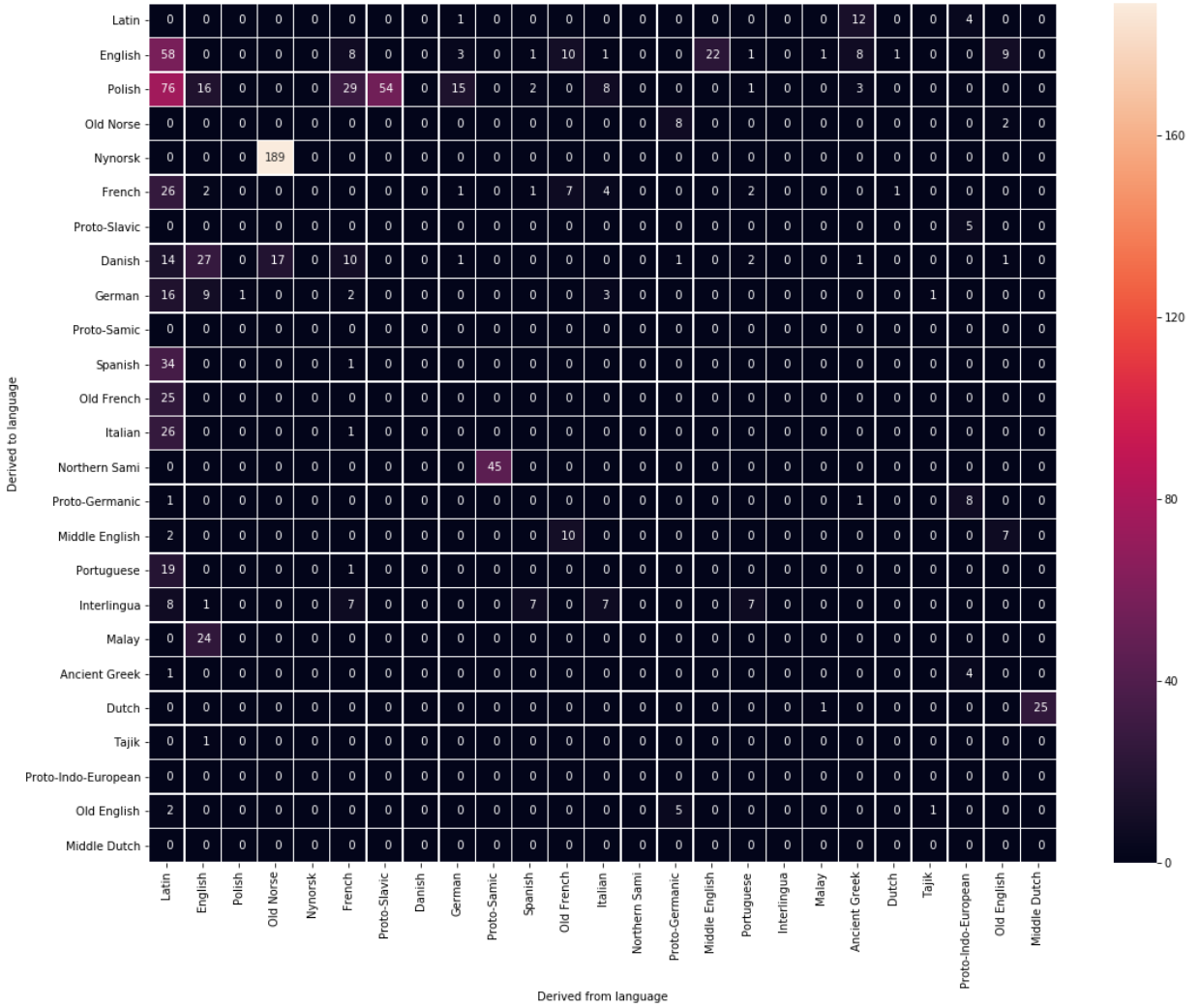
Etymological relations (either within or between languages) can be described with *derived from* (P5191), currently used some thousand times.

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.

brasa (pt) — braise (fr) — bresze (Middle French) — breze (Old French) — *... (Gothic) — *brasō (Proto-Germanic) — *b^{h}res- (Proto-Indo-European)

Etymology between languages



Etymological derivation matrix for 25 languages

A sparse matrix, — much smaller than Wiktionary.

Largest number of recorded links are from Old Norse to Norse.

A PageRank analysis yields Proto-Indo-European on the top.

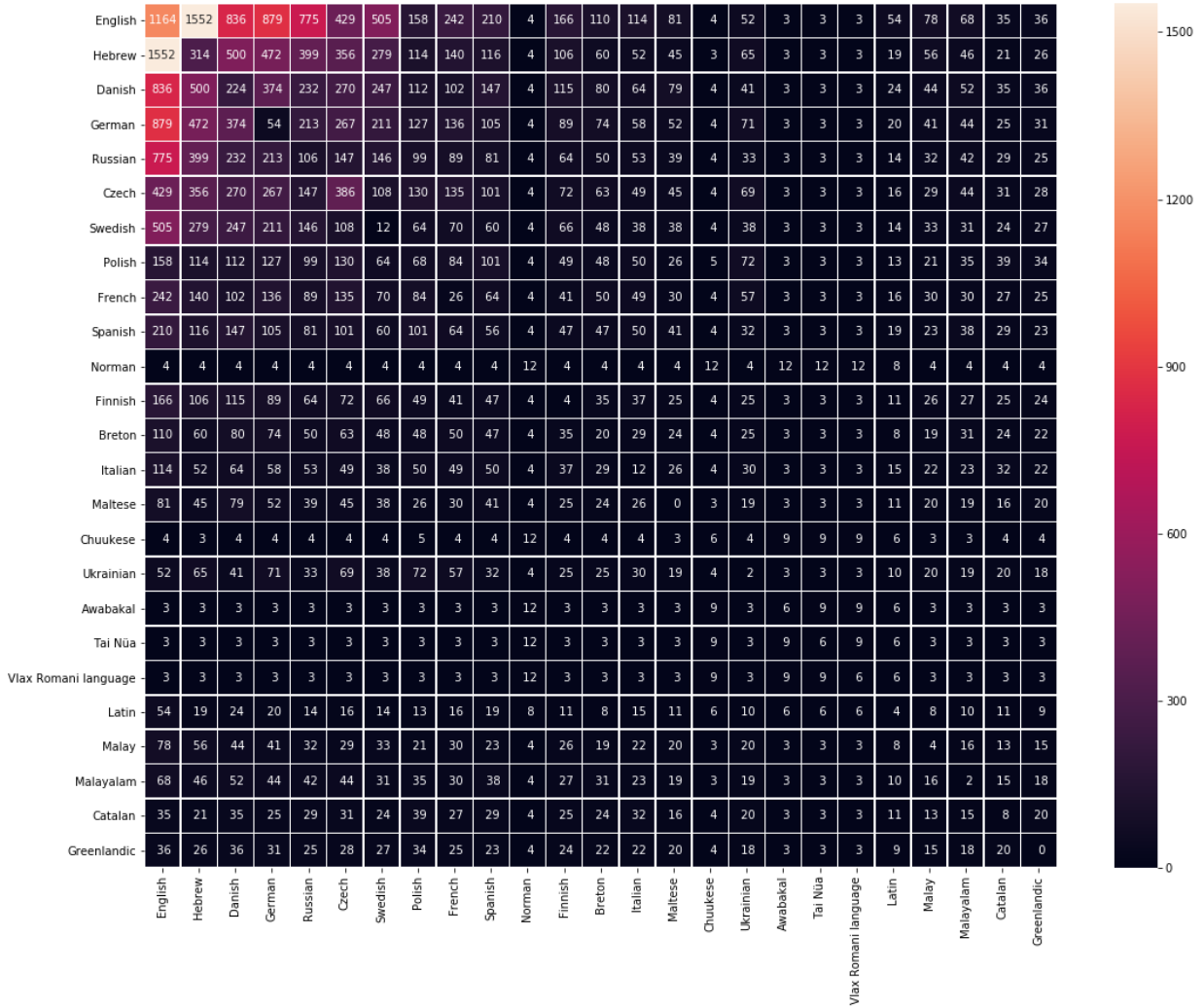
Linking lexemes via senses

Via sense and Q-items, e.g., the Danish noun “bil” has the sense bil (car) that may be linked to Wikidata’s Q-items Q1420, which is also linked from the English noun “car” and car-sense.

```
bil                        bil-sense                car-concept
wd:L36385 ontolex:sense wd:L36385-S1 wdt:P5137 wd:Q1420

car                      car-sense                car-concept
wd:L3648 ontolex:sense wd:L3648-S1 wdt:P5137 wd:Q1420
```

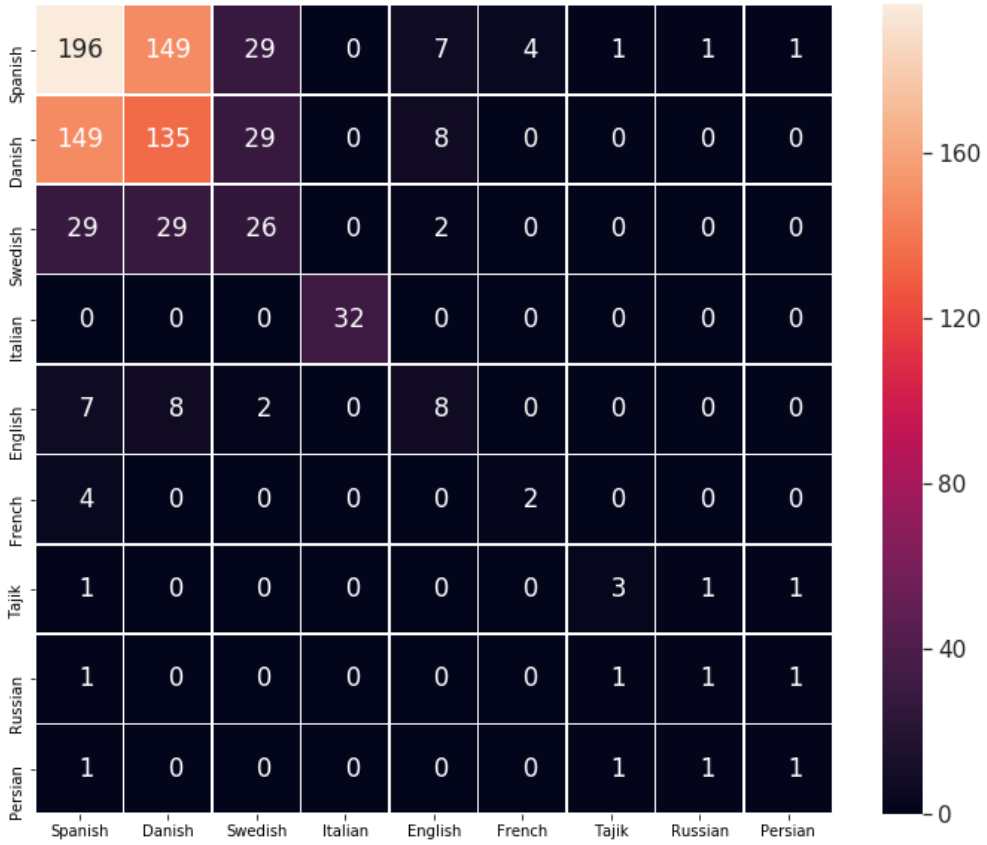
Senses



Sense-Q-item matrix for the top 25 languages.

In February, only the combination English-Hebrew had more than 1,000 translation

Demonyms



Lexemes may (also) be linked to Q-items by the *demonym of* (P6271) property that is only relevant to use for demonyms.

Spanish-Danish language pair had the largest number of links between demonyms in February 2020.

Identifier statistics

February	June	Identifier	Language(s)
14440	18779	Elhuyar	Basque
2878	3610	DanNet word	Danish
1688	1688	WSO Online	Polish
1353	1352	SJP Online	Polish
1288	1288	Doroszewski	Polish
1027	1026	Dobry słownik	Polish
1009	1008	WSJP	Polish
388	463	Oqaasileriffik	Greenlandic, Danish, English
216	241	Vocabolario Treccani	Italian
212	212	OED Online	English
160	160	Kopaliński	Polish

Table 2: External identifiers in Wikidata sorted according to usage per 22 February 2020 and 18 June 2020. Updated statistics is available at https://ordia.toolforge.org/statistics/

Discussion

Wikidata's requirement for Creative Commons Zero (CC0) license creates a problem for the inclusion of non-public domain resourses, such as Wiktionaries and wordnets.

Releasing lexical resources under CC0 will probably led to expansion of Wikidata's lexicographic data.

"Wikimedia Foundation's preliminary perspective on a legal issue" available on https://meta.wikimedia.org/wiki/Wikilegal/Lexicographical_Data: Russian Wikidata lexemes has setup grammatical forms.

Few semantic links of non-noun lexemes: Should *little* be linked to (a hypothetical) *smallness* Q-item, or should adjective Q-items be established or should there be specialized properties, e.g., *pertainym*.

Summing up

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. With Basque and Hebrew as exceptions.

Links between lexemes of different languages can be established by an etymological property as well as through senses and the Q-items of Wikidata.

The intralanguage linkage was as of February 2020 still quite sparse.

Links to external lexicographic resources can be established by several external identifier properties in Wikidata. But except for Basque, still sparse.

Thanks

References

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. DOI: 10.1007/978-3-319-11964-9_4.

Nielsen, F. Å. (2019). Ordia: A Web application for Wikidata lexemes. *The Semantic Web: ESWC 2019 Satellite Events*, pages 141–146. DOI: 10.1007/978-3-030-32327-1_28.

Nielsen, F. Å., Mietchen, D., and Willighagen, E. (2017). Scholia, Scientometrics and Wikidata. *The Semantic Web: ESWC 2017 Satellite Events*, pages 237–259. DOI: 10.1007/978-3-319-70407-4_36.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57:78–85. DOI: 10.1145/2629489.