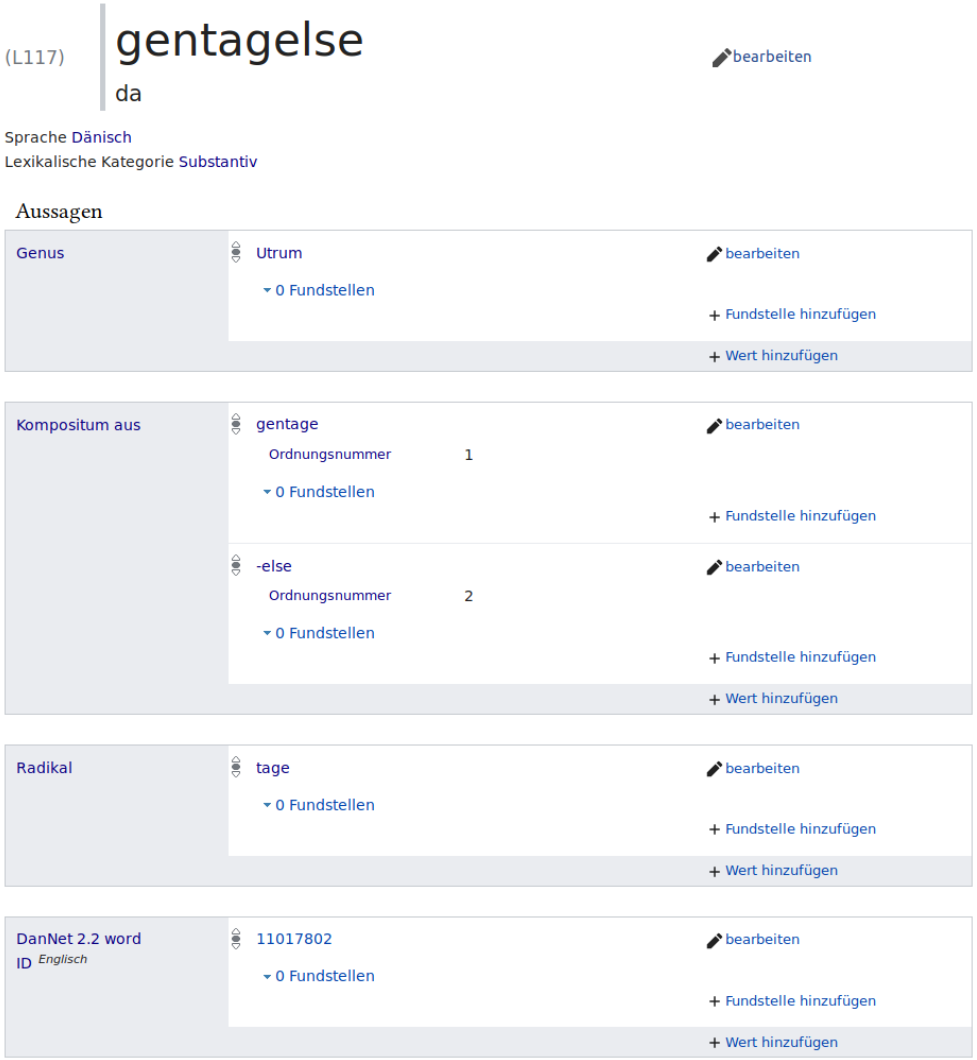
Wikidata lexemes

Finn Årup Nielsen

(Wikimedia Denmark, Technical University of Denmark)

Wikimania, Saturday 17 August 2019 15:00-15:30, Menchú

Wikidata Lexemes



In 2018, Wikidata introduced a new type of entities for lexemes, their form(s) and sense(s).

Lexemes are prefixed with ‘L’, e.g., L117 for the Danish word *gentagelse* (repetition).

On the same page: the sense(s) and form(s) of the lexeme

The lexeme, form and sense (L-Wikidata) may be described by links to the ordinary Wikidata (Q-Wikidata).

Wikidata lexeme basic model

- lemma (wikibase:lemma)
- language (dct:language)
- lexical category (wikibase:lexicalCategory)
- Zero or more statements with property and property values
- Zero or more Senses (ontolex:sense)
  - gloss (skos:definition)
  - Zero or more statements with property and property values
- Zero or more Form (ontolex:lexicalForm)
  - representation (ontolex:representation)
  - Zero or more grammatical features (wikibase:grammaticalFeature)
  - Zero or more statements with property and property values

Wikidata Lexemes RDF

```
wd:L117-F3 a wikibase:Form,
                ontolex:Form ;
        rdfs:label "gentagelser"@da ;
        ontolex:representation "gentagelser"@da ;
        wikibase:grammaticalFeature wd:Q146786,
                wd:Q53997857 ;
        wdt:P5279 "gen·ta·gel·ser" ;
        wdt:P2859 "\"gEn$%ta:?$@l$s6" ;
        p:P5279 wds:L117-F3-83a6b790-4e1d-56b2-2511-95...

wds:L117-F3-83a6b790-4e1d-56b2-2511-95f1877d046e a wik...
                wikibase:BestRank ;
        wikibase:rank wikibase:NormalRank ;
        ps:P5279 "gen·ta·gel·ser" .
```

Wikidata lexeme statistics

Count	Description	Query
815724	Number of grammatical feature links	[] wikibase:grammaticalFeature []
356681	Number of forms	[] a ontolex:Form
56518	Number of lexemes	[] a ontolex:LexicalEntry
56518	Number of language links	[] dct:language []
56518	Number of lexical category links	[] wikibase:lexicalCategory []
14196	Number of senses	[] a ontolex:LexicalSense
14196	Number of sense links	[] ontolex:sense []
7586	Number of sense to item links	[] wdt:P5137 []

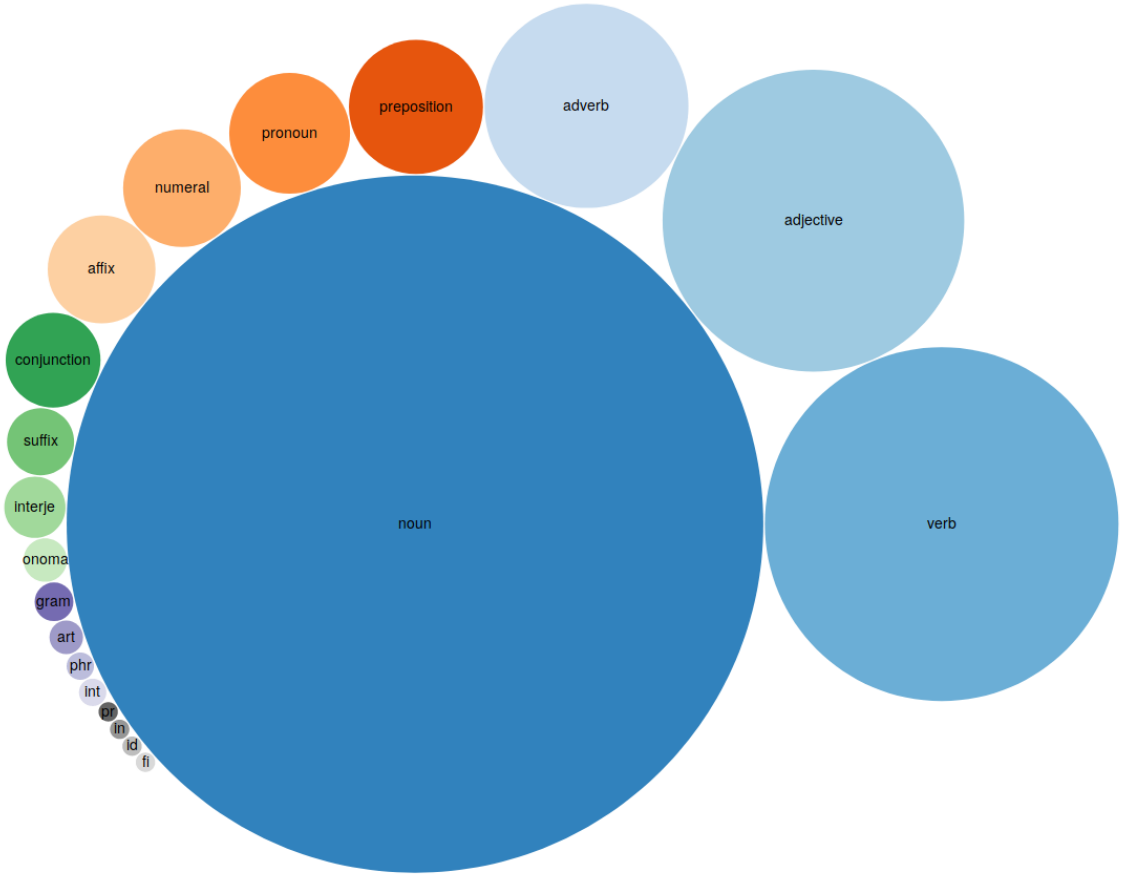
Ordia's statistics: https://tools.wmflabs.org/ordia/statistics/

Wikidata lexeme language statistics

Number of lexemes	Language
15980	English
10448	French
7620	Swedish
3021	Basque
2807	Nynorsk
2614	Czech
2453	Polish
2279	German
2207	Danish
721	Japanese

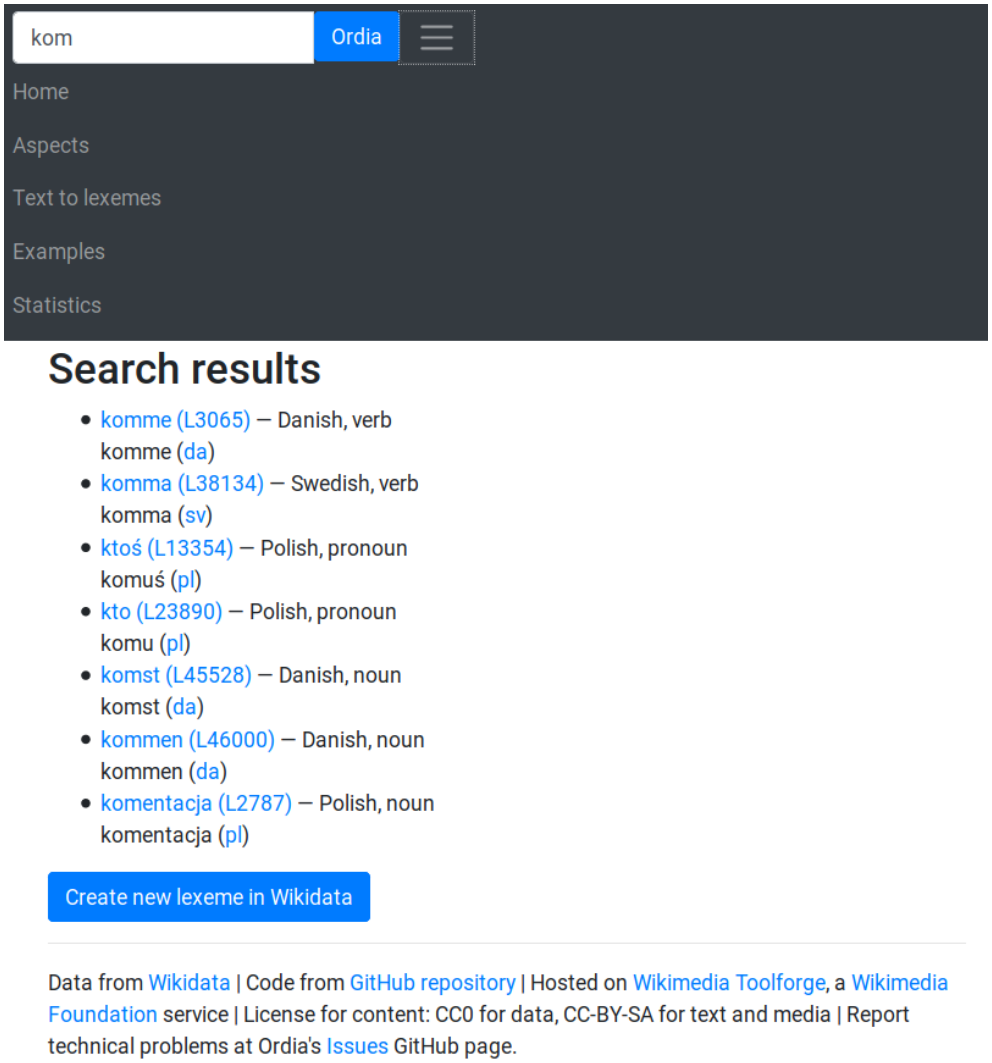
Ordia's language statistics https://tools.wmflabs.org/ordia/language/

Wikidata lexeme Danish statistics



https://tools.wmflabs.org/ordia/language/Q9035

Wikidata lexeme tools



Several tools have been built on top of Wikidata: Wikidata:Tools/Lexicographical_data

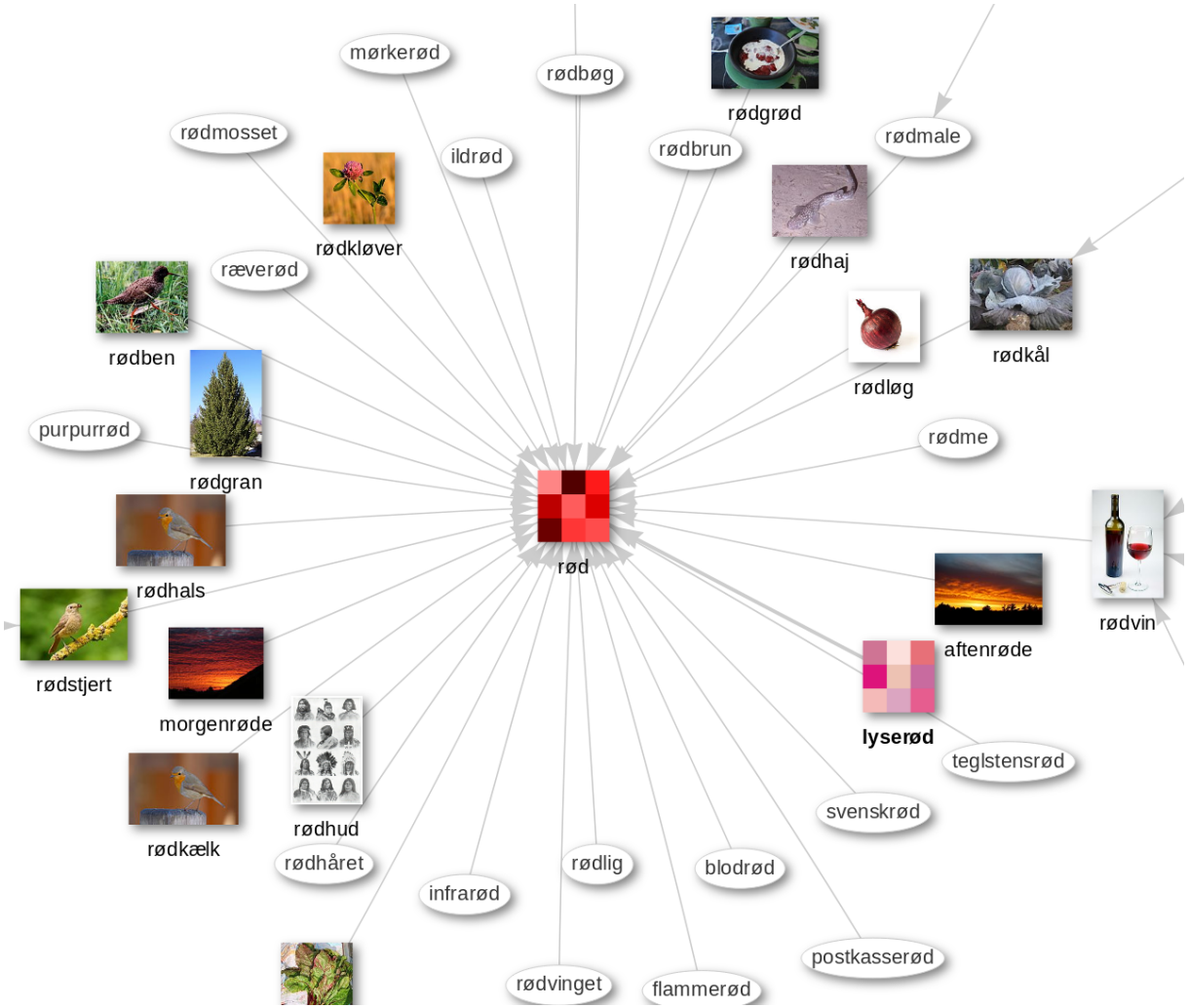
Ordia (Nielsen, 2019) is SPARQL-based webservice at https://tools.wmflabs.org/ordia/

Lea Lacroix' DerDieDas game http://auregann.fr/derdiedas/

Lucas Werkmeister's form input https://tools.wmflabs.org/lexeme-forms/

Alicia Fagerving's Hauki https://tools.wmflabs.org/hauki/

Compound and derivation graph



Wikidata properties for lexemes can specify derivation lexeme (P5191) and compound parts (P5238).

The SPARQL-based *Wikidata Query Service* can generate graph visualization on the fly and include associated images from *Wikimedia Commons*.

`https://w.wiki/6My` or *rød* in Ordia

Text-to-lexemes

Text to lexemes

Blue cars, green bikes and red motorcycles must stop at the crossing.

Language:

English

Submit

Extraction

Search:

Word	Form	Lexeme	Lexical category	Features	Sense	Sense image
and	and	and	conjunction		L1385-S1	
at	at	at	preposition			
bikes	bikes	bike	verb	simple present // third-person singular		
bikes	bikes	bike	noun	plural	L10698-S1	
blue	blue	blue	noun	singular		
blue	blue	blue	adjective	positive	L3269-S1	
cars	cars	car	noun	plural	L3648-S1	
crossing	crossing	cross	verb	present participle		
crossing	crossing	crossing	noun	singular	L31093-S1	

Lexeme extraction in Ordia

For a Danish example: “Regeringen spiser grønne æbler om vinteren”.

For “Blue cars, green bikes and red motorcycles must stop at the crossing.”

Collaboration?

Users seem segregated across languages? How can we improve coordination?

Example: How should we specify countable nouns, motion nouns, etc.? For Danish, I have used *instance of* Should *proper noun* be used for lexical category? Should we clean up lexical categories?

Hyphenation (P5279) as an example: should "·" be the hyphenation sign? How should one specify no hyphenation? See Property talk:P5279.

Examples of annotation: *gågade* (Danish, walking street, L57830), *hund* (Danish, dog L31499). Images? Audio? Translation? X-SAMPA?

WordNet?

Identifier	Count	Property
ILI	27	P5063
BabelNet	60'478	P2581
DanNet word	1'577	P6140

Wordnets are professionally curated structured data about words and gives machine readable semantics, senses, hyponym, synonym, etc.?

Global WordNet Association are the central organization.

There has been some contact between Wikidata lexemers and wordnetters.

Showstopper is the license that wordnets are typically use. What can we do?

We have DanNet, ILI and, BabelNet identifiers. Do we need more?

Wikidata license?

Wikidata's use of *Creative Commons Zero* (public domain) means that we (probably?) cannot use Wiktionary and, e.g., Wikipedia for usage example.

The license of other linguistics resources may create limitations on what we can include in Wikidata.

For Danish, I have been using of the Europarl corpus, NST data.

For Danish the language has changed and out-of-copyright works typically have an old spelling.

Where can find good modern CC0 corpora and lexical resources?

Validation?

Wikidata's property constraint system can indicate that, e.g., an identifier is used twice.

Shape Expressions (ShEx) for lexemes (Nielsen et al., 2019) can specify constraints, e.g., specify that Danish noun in definite plural should end with e(r)ne and possible exceptions. ShEx example with DanNet:

```
<dannet-statement> EXTRA rdf:type {
  # DanNet identifier should either be novalue or a string
  ( rdf:type [ wdno:P6140 ] | ps:P6140 xsd:string )
}
```

Available in a special name space in Wikidata: https://www.wikidata.org/wiki/EntitySchema:E15.

Should we extend this approach to more/across languages?

Wish lists for Wikibase/Wiki developers/us

Are we missing fundamental parts in Wikibase?

We have no means of specifying, e.g., frequency of a word (representation regardless of lexeme).

Example: It is not possible to have references for lexical categories and glosses, which is a problem if we want to attribute/source where a gloss came from.

Can we specify grammar in some way?

References

Nielsen, F. Å. (2019). Ordia: A Web application for Wikidata lexemes.

Nielsen, F. Å., Thornton, K., and Gayo, J. E. L. (2019). Validating Danish Wikidata lexemes.

Copyright and licenses