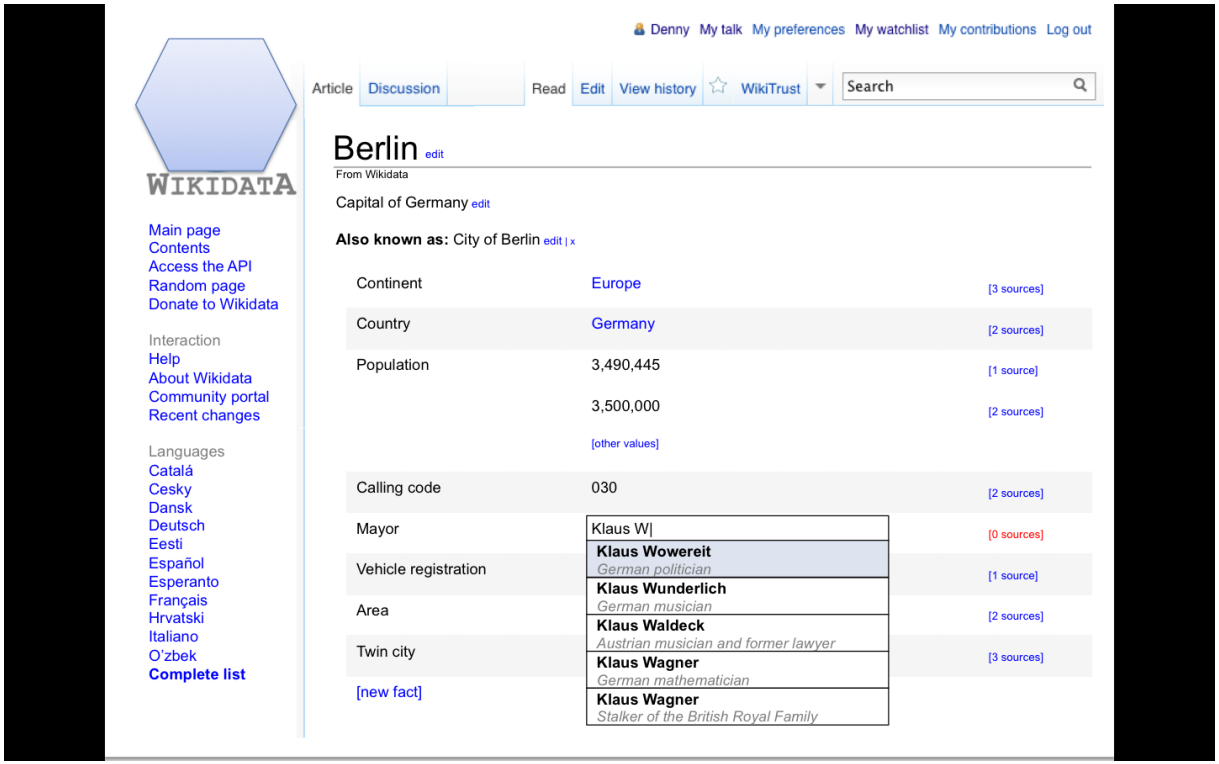
Danish in Wikidata Lexemes

Finn Årup Nielsen

Cognitive Systems, DTU Compute, Technical University of Denmark

27 July 2019

Wikidata



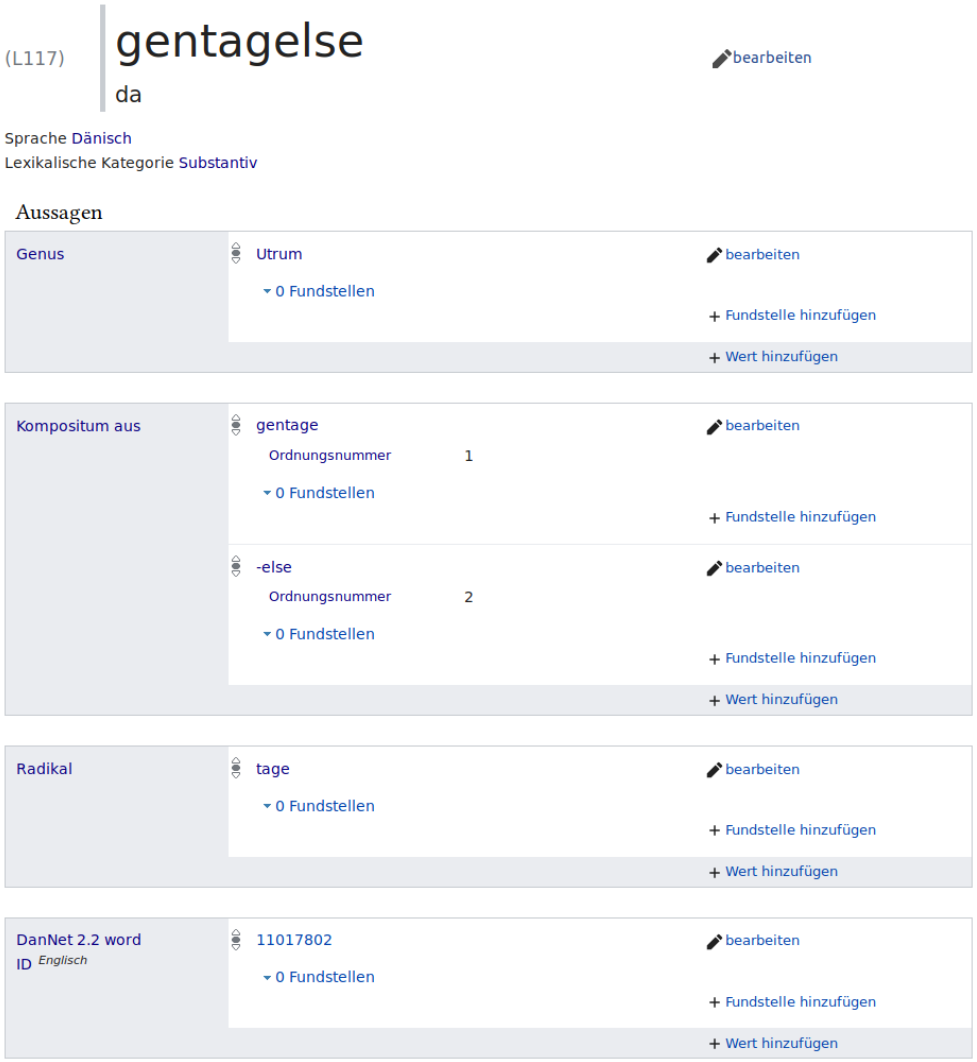
"Wikidata: Verifiable, Linked Open Knowledge That Anyone Can edit" (Dario Taraborelli)

Open data accessible from website, API, dump files and SPARQL endpoint.

Every page is an "item" with label, aliases, properties, property values, and Wikipedia et al. links (Vrandečić and Krötzsch, 2014).

Wikidata-site mockup from 2012 for Berlin (Q64): https://www.wikidata.org/wiki/Q64.

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-56b2-2511-95f1877d046

wds:L117-F3-83a6b790-4e1d-56b2-2511-95f1877d046e a wikibase:Sta
                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 []

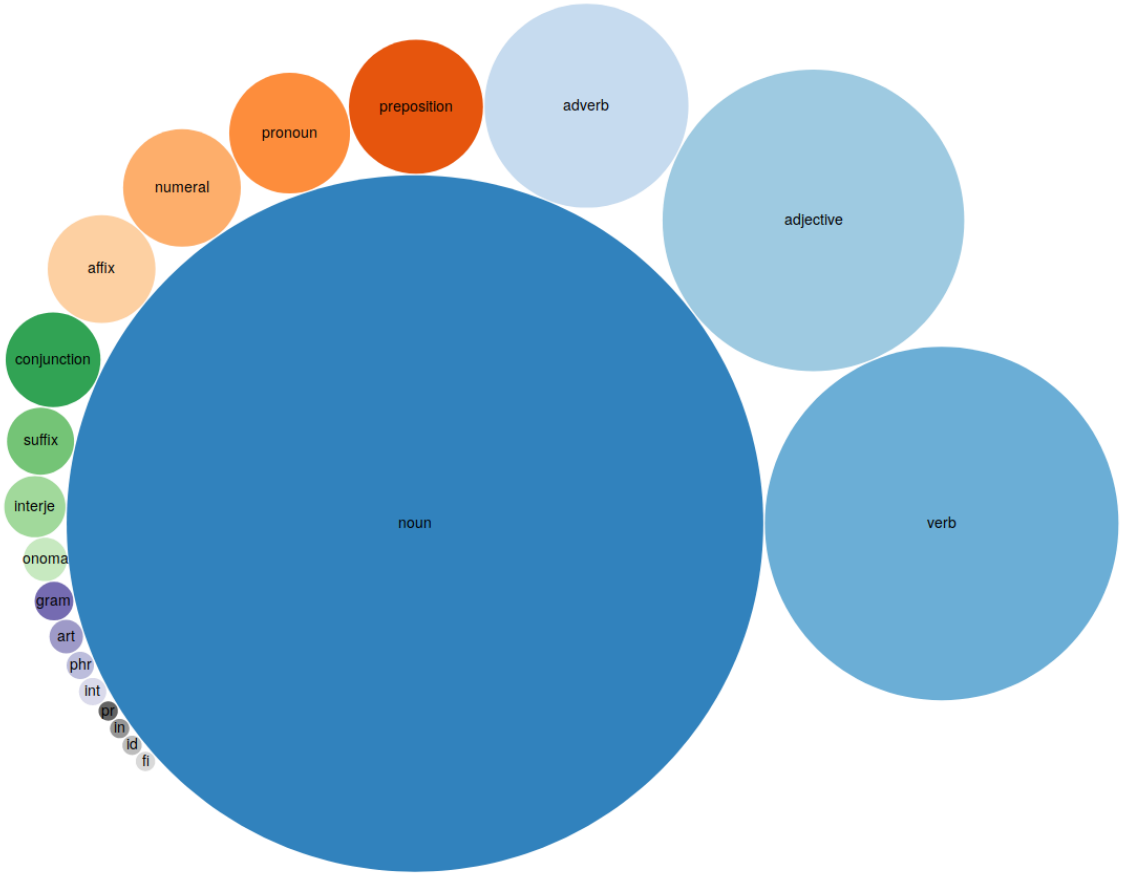
Ordia's statistics: https://tools.wmflabs.org/ordia/statistics/

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

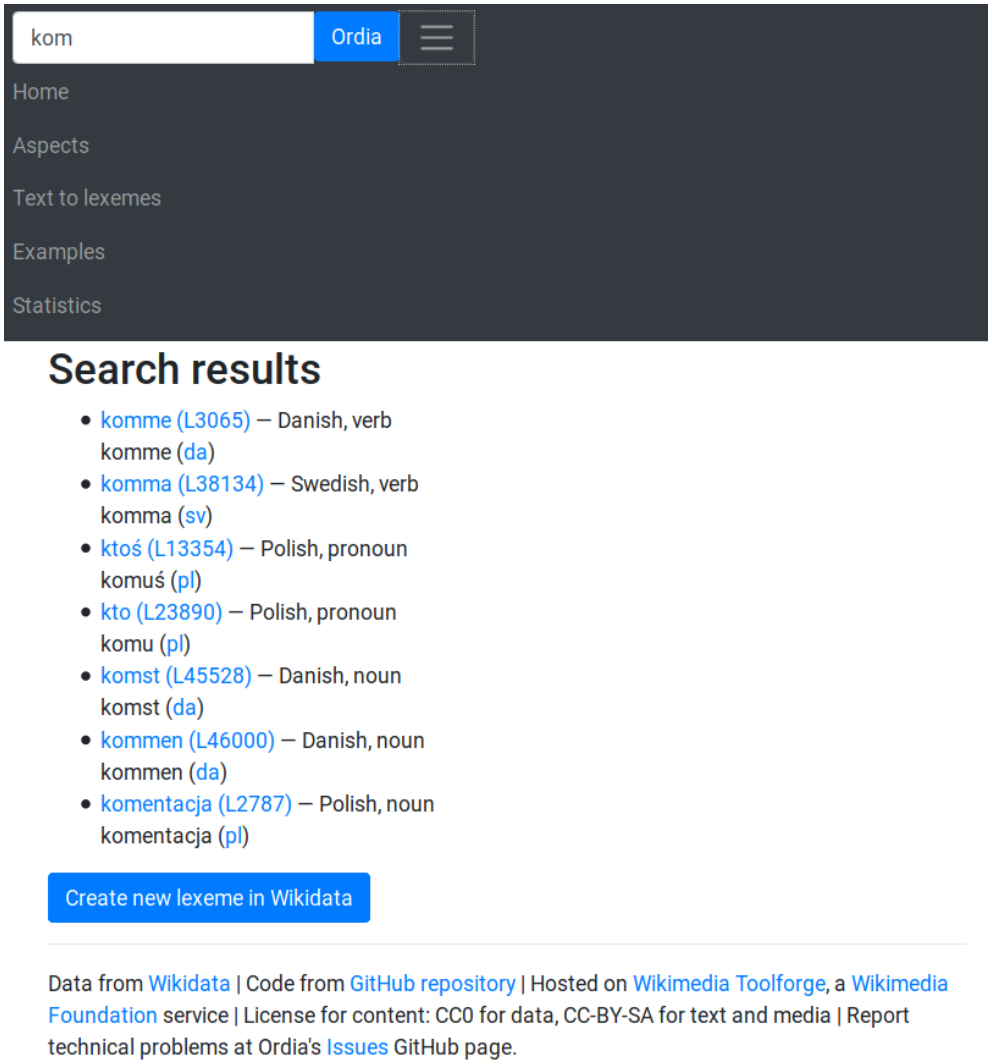
Ordia's language statistics https://tools.wmflabs.org/ordia/language/

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 Haiki https://tools.wmflabs.org/hauki/

Wikidata lexeme properties

Lexemes: "instance of" (values, e.g., countable noun, mass noun, compound word, unadapted loan word), auxilliary verb, DanNet 2.2 word-ID, "derived from", word stem, image file link, usage example.

Senses: "item for this sense" (very important because it links to the rest of the Wikidata graph with, e.g., hypernym information), image file link, audio file link, translation.

Forms: pronounciation audio file link, X-SAMPA, hyphenation (https://tools.wmflabs.org/ordia/hyphenation/).

Wikidata and Linguistic Linked Open Data

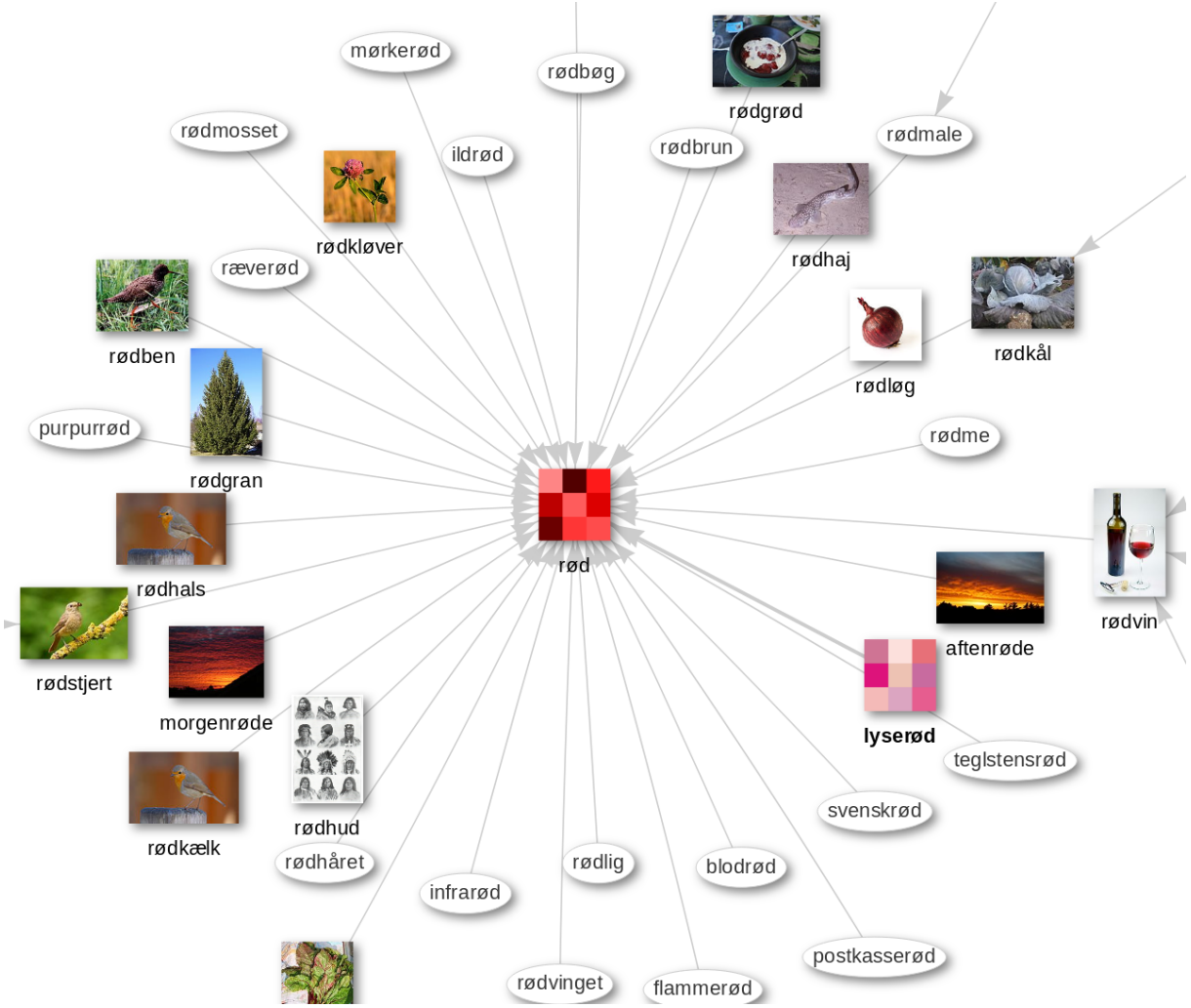
Wikidata has deep links to items in the Linguistic Linked Open Data cloud.

Identifier	Count	Property
ILI	27	P5063
BabelNet	60'478	P2581
DanNet word	1'514	P6140

For DanNet we have speficified when a Wikidata lexeme does *not* have and associated DanNet 2.2 word identifier with *no value*, see, for instance, *investeringsforvaltningsholdingvirksomhed*.

The generic property *exact match* (P2888) has been used to link to some hundred Princeton WordNet URI identifiers, especially the ImageNet WordNet 3.0 identifiers (Nielsen, 2018).

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.”

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 can 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.

Wikidata license

Wikidata uses *Creative Commons Zero* (public domain)

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

. . . use of the Europarl corpus, NST data, . . .

Summary

Wikidata lexemes is a very young project.

Expanding from zero lexemes in 2018 to 56528 in July 2019 in multiple language.

Rich annotation scheme for lexemes, forms and senses.

SPARQL queries with visualization possible.

Wikidata Lexeme data validation is possible

CC0 licens for linguistic resources could Wikidata to include the information.

References

Nielsen, F. Å. (2018). Linking ImageNet WordNet Synsets with Wikidata. *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France*, pages 1809–1814. DOI: 10.1145/3184558.3191645.

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.

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