Danish in Wikidata Lexemes

Finn Ārup Nielsen

Cognitive Systems, DTU Compute, Technical University of Denmark

27 July 2019
“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.
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-56b2-2511-95f1877d046e a wikibase:Statement, 
        wikibase:BestRank ; 
    wikibase:rank wikibase:NormalRank ; 
    ps:P5279 "gen·ta·gel·ser" .
## Wikidata lexeme statistics

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

**Ordia’s statistics:** [https://tools.wmflabs.org/ordia/statistics/](https://tools.wmflabs.org/ordia/statistics/)
### Wikidata lexeme language statistics

<table>
<thead>
<tr>
<th>Number of lexemes</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>15980</td>
<td>English</td>
</tr>
<tr>
<td>10448</td>
<td>French</td>
</tr>
<tr>
<td>7620</td>
<td>Swedish</td>
</tr>
<tr>
<td>3021</td>
<td>Basque</td>
</tr>
<tr>
<td>2807</td>
<td>Nynorsk</td>
</tr>
<tr>
<td>2614</td>
<td>Czech</td>
</tr>
<tr>
<td>2453</td>
<td>Polish</td>
</tr>
<tr>
<td>2279</td>
<td>German</td>
</tr>
<tr>
<td>2207</td>
<td>Danish</td>
</tr>
<tr>
<td>721</td>
<td>Japanese</td>
</tr>
</tbody>
</table>

Ordia’s language statistics [https://tools.wmflabs.org/ordia/language/](https://tools.wmflabs.org/ordia/language/)
Wikidata lexeme Danish statistics

https://tools.wmflabs.org/ordia/language/Q9035
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/
Wikidata lexeme properties

Lexemes: “instance of” (values, e.g., countable noun, mass noun, compound word, unadapted loan word), auxiliary 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: pronunciation 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.

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Count</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILI</td>
<td>27</td>
<td>P5063</td>
</tr>
<tr>
<td>BabelNet</td>
<td>60,478</td>
<td>P2581</td>
</tr>
<tr>
<td>DanNet word</td>
<td>1,514</td>
<td>P6140</td>
</tr>
</tbody>
</table>

For DanNet we have specified 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 Wiki-data Query Service can generate graph visualization on the fly and include associated images from Wikimedia Commons.

https://w.wiki/6My or rød in Ordia
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:

```xml
<dannet-statement> EXTRA rdf:type {
  # DanNet identifier should either be no value or a string
}
```

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 let Wikidata include the data.
References


