Danish in Wikidata lexemes

Finn Årup Nielsen
Cognitive Systems, DTU Compute, Technical University of Denmark
Kongens Lyngby, Denmark

Abstract

Wikidata introduced support for lexicographic data in 2018. Here we describe the lexicographic part of Wikidata as well as experiences with setting up lexemes for the Danish language. We note various possible annotations for lexemes as well as discuss various choices made.

1 Introduction

Wikipedia's structured sister Wikidata (Vrandečić and Krötzsch, 2014) at https://www.wikidata.org/ supports interlinking different language versions of Wikipedia as well as several other Wikimedia sites, such as Wikibooks and Wikimedia Commons. One wiki that has been missing from the list is Wiktionary, — the dictionary wiki. Wiktionary has a structure different from the other wikis as multiple different words and concepts might be described on the same page, only connected through the same orthographic representation.

In 2018, Wikidata enabled support for lexicographic data via special lexeme wiki pages. Compared to Wiktionary, Wikidata lexemes offer a solution with directly machine-readable data: it is not necessary to write parsers to obtain the lexeme data in a structured format. Wikidata lexemes also reduce the amount of redundant input: In Wiktionary, each language edition sets up its own dictionary, and a word described in one Wiktionary is not directly available in another Wiktionary. Further issues with Wiktionary are the linkage to the Wikidata concept ontology and the linkage to external resources such as WordNet (Miller, 1995). Neither of these links is non-trivial to set up, though matching lexical entries between Wiktionary and WordNet may be done with good accuracy (McCrae et al., 2012).

Below we will describe how lexemes are supported on Wikidata,[1] and list some of the Danish resources relevant for Wikidata lexemes. Then we will detail how Danish lexemes have been annotated and discuss some of the choices made.

[1]There is an introduction to Wikidata lexemes on Wikidata itself at https://www.wikidata.org/wiki/Wikidata_talk:Lexicographical_data.

2 Wikidata lexemes

Wikidata stores lexeme data on a new type of pages prefixed with the letter 'L' and further identified with an integer, e.g., the Danish lexeme *gentagelse* (repetition) has the identifier "L117" and available for view and edit at https://www.wikidata.org/wiki/Lexeme:L117. On the same page, multiple senses and forms for the lexeme may be defined. They are identified by suffixes to the lexeme identifier, e.g., the plural indefinite form *gentagelser* would be identified as "L117-F3", while the first sense—if it was defined—would have been identified as "L117-S1". Forms and senses are defined separately, so it is currently difficult to define a specific sense for a specific form. Lexemes, forms and senses may be associated with properties, and these properties are identified with integer prefixed with the letter 'P'.

The Wikidata lexeme data maps to an RDF representation,[2] and the RDF data is queryable via the *Wikidata Query Service* SPARQL endpoint at https://query.wikidata.org/. The mapping uses part of the *Lexicon Model for Ontologies* (LEMON) ontology (Q59, 2016; McCrae et al., 2012). The central OWL concepts for the lexeme data are ontolex:LexicalEntry, ontolex:Form and ontolex:LexicalSense for lexeme, form and sense respectively with the prefix http://www.w3.org/ns/lemon/ontolex#. Each of these three OWL concepts has associated basic data in Wikidata. Apart from the identifier, the lexeme has the lemma, language and lexical category, the form has its orthographic representation and grammatical features while the sense may have multiple glosses. This basic data cannot be associated with qualifiers and references like normal Wikidata properties.

Links from Wikidata lexemes (L-pages) to Q-items (i.e., the ordinary Wikidata items) are of two kinds: Either for the description of the lexical and grammatical "metadata" for the lexeme or for the description of the meaning of a sense. In the latter case, the Q-items function as the wordnet notion of *synsets*. Wikidata has specific properties to link

[2]https://www.mediawiki.org/wiki/Extension:WikibaseLexeme/RDF_mapping

Danish	Total	Description	SPARQL query fragment
1268	43816	Number of lexemes	`[] a ontolex:LexicalEntry`
4826	118742	Number of forms	`[] a ontolex:Form`
617	11194	Number of sense	`[] a ontolex:LexicalSense`
8594	218803	Number of grammatical feature links	`[] wikibase:grammaticalFeature []`

Table 1: Statistics for lexeme data in Wikidata. See also the statistics displayed on the Ordia website at https://tools.wmflabs.org/ordia/statistics/.

Q-items to synsets in external lexical resources, including BabelNet (Navigli and Ponzetto, 2010) (P2581) and the Collaborative Interlingual Index (P5063) (Bond et al., 2016). Alternatively, the more generic property for Linked Open data URIs (P2888) can be used. Wikidata has linked some WordNet synsets used in ImageNet. The correspondence between the resources is not necessarily straightforward to establish (Nielsen, 2018).

Some statistics for the lexeme data in Wikidata are displayed in Table 1. It displays, e.g., that the number of forms is close to 120'000. In comparison, the English Wiktionary has currently around 5.9 million content pages, while the Danish Wiktionary has around 38 thousand.[3] The numbers are not directly comparable as multiple forms may be listed on one Wiktionary content page. A count on the distinct number of (monolingual) form representations in Wikidata gives 89'728 on 27 March 2019 based on the following SPARQL query:

```
SELECT
  (COUNT(DISTINCT(?representation))
  AS ?count)
{ [] ontolex:representation
  ?representation . }
```

3 Danish resources

There are some Danish resources relevant for Wikidata lexemes, e.g., corpora for language usage examples. As Wikidata is distributed under the Creative Commons Zero (CC0) license, the resources incorporated into Wikidata need to be compatible with that license.

Old out-of-copyright Danish works are typically with an antiquated spelling, e.g., where the first letter of nouns has a capital letter. Wikipedia and Wiktionary may not be used because their content is under an attribution and share-alike license, not compatible with the CC0 license. Modern Danish sentences can be retrieved from, e.g., Danish law texts at https://www.retsinformation.dk/, Danish translations of international treaties and conventions, such as the Treaty of Lisbon, and the Danish part of the Europarl corpus (Koehn, 2005). Fairy tales by Hans Christian Andersen can be found with modern spelling.

Of the lexicographical resources, the standard Danish dictionary, *Retskrivningsordbogen*, has a restrictive license. Another large Danish dictionary with over 300'000 entries and used, e.g., with the computer program *aspell*, is under the GNU General Public License and is not compatible with Wikidata's CC0. DanNet (Pedersen et al., 2009) has a WordNet-derived open license and a Wikidata property (P6140) for the DanNet words — corresponding to Wikidata lexemes—has been created in November 2018. DanNet is distributed as OWL, so should fit well with Wikidata lexemes.

NST Lexical database for Danish[4] has Speech Assessment Methods Phonetic Alphabet (SAMPA) pronunciation specification for over 235'000 Danish words and stated to have the CC0 license.

As of June 2019, we have used 160 sentences from the Danish part of the Europarl corpus,[5] and linked to 1258 DanNet 2.2 word identifiers,[6] while the NST phonetic data has hardly been used.

4 Annotating lexemes, forms and senses

Wikidata has a continuously growing number of properties that can be used to annotate lexemes, forms and senses. General properties—that are relevant for lexemes of most word classes—are *usage example* (P5831), *word stem* (P5187) and *derived from* (P5191), where the latter may indicate etymological origin or origin of derivations. Compound parts may be linked with a property (P5238) and the order of the parts may be specified with a property used as a qualifier (P1545). The Wikidata property for DanNet words (P6140) are linked to version 2.2 of the resource. As of March 2019, 844 lexemes with associated information about DanNet words are linked.[7] The data model of Wikidata allows for the specification of "no value", thus it is

[3] https://en.wiktionary.org/wiki/Special:Statistics and https://da.wiktionary.org/wiki/Speciel:Statistik

[4] https://www.nb.no/sprakbanken/show?serial=sbr-26

[5] See Ordia's statistics at https://tools.wmflabs.org/ordia/reference/Q5412081

[6] The SPARQL query `SELECT ?dannet { ?lexeme wdt:P6140 ?dannet }` on the Wikidata Query Service.

[7] https://www.wikidata.org/wiki/Property_talk:P6140 displays the DanNet property statistics.

possible to specify that a lexeme cannot be found in the DanNet 2.2 resource. For instance, adverbs and rare nouns, such as *lommevogn* (L40687), are not in DanNet and indicated as such. The *usage example* property (P5831) can store a short free-form text and the qualifier *stated in* (P248) can point to a Q-item with metadata about a work where the text appears. A related property is *attested in* (P5323) which also can point to a Q-item.

Lexemes may also be associated with classes via the *instance of* property (P31). Properties relevant for lexemes across word classes in Danish are, e.g., whether they loan words and/or compound words.

Forms may be associated with hyphenation and pronunciation specification. Wikidata has properties for X-SAMPA, IPA transcription and Kirshenbaum code. These pronunciation properties have been used on Wikidata's Q-items, but so far not (or very limited) for Danish lexemes.

Senses can be associated with language style (P6191) and perhaps most importantly with *item for this sense* property (P5137) which links the senses of lexemes to the Q-items and thus with the rest of the Wikidata knowledge graph. Synonyms, antonyms, hypernyms and hyponyms may be inferred from the information in that part of the Wikidata knowledge graph.

Links between lexemes in different languages can currently be made with a specific translation property (P5972) applicable for senses, or the connection between lexemes can be made through their senses and the P5137 property linking to Q-item that then binds lexemes from separate languages together.

4.1 Verbs

Verbs can be associated with conjugation class through the P5186 property. We have followed the scheme of (Allan et al., 1995) where there are four main Danish conjugation categories. The auxiliary verb(s) for a verb can be specified with P5401. Some verbs can be assigned to a class, e.g., motion verbs, auxiliary verb, transitive/intransitive verb or deponent verb. The valence (P5526) can also be specified.

4.2 Nouns

Danish nouns may be characterized by grammatical gender and class. Classes of commons nouns may be countable or mass noun, singulare tantum, plurale tantum, collective noun, 'nexual' or 'innexual' noun or nomen agentis. The distinction between nexual and innexual is based on (Hansen and Heltoft, 2019) where the former may refer to "actions and processes, activities and states," and the later "objects or compounds".

4.3 Images and audio

Senses may be associated with images by referencing filenames in the free media archive *Wikimedia Commons*. The link may help language learners and possibly be a resource for training natural language processing machine learning models in the same way that ImageNet has used WordNet, see (Nielsen, 2018; Nielsen and Hansen, 2018) for applications of the use of Wikidata. Typically the senses of nouns may be associated with images, while it may be difficult to identify good images to be associated with, e.g., adverbs. A few Danish verbs have been associated with images, e.g., *gå* (walk) and "visual" adjectives, such as *rød* (red), are also associated with images.

Lexemes can be associated with images. Photos of written signs may exemplify how words are used in the environment, e.g., a photo of a street sign reading "Cyklist vig for gående" is used to illustrate the usages of the lexeme *cyklist* (L43527, cyclist).

Audio files can be associated with the lexicographic data. For forms, the P443 property can link to one of the currently around 130 pronunciation audio files for Danish words, while senses can link to sound files with the P51 property, e.g., the sense for *bi* (L37259, bee) links to a sound recording of bees buzzing and the sense for *bil* (L36385, car) is associated with an audio file of a starting and driving car.

5 Discussion

Wikidata is entirely field-based and especially for lexemes there are very few means to enter free-form information. While exceptions can be noted in standard dictionaries such as Wiktionary, almost every piece of information added for a lexeme in Wikidata must be associated with a property. The explicitness of Wikidata complicates the modeling of language. Below we discuss a few of the issues that have appeared for the Danish language.

5.1 Lexeme splitting

The English lexeme *they* (L371) incorporates the forms *they*, *them*, *their*, *theirs*, *themselves* and *themself*, while French *vous* (plural *you*) and *votre* are separate lexemes (L9289 and L9289). In the Danish online dictionary *Den Danske Ordbog*, the corresponding forms for *they* are split into several dictionary entries, while the German Wikidata lexemes *ich* (L7877, the personal pronoun *I*) has currently no other form than *ich*. The issue was the subject of an inconclusive discussion on Wikidata.[8] As noted by one of the discussants, if the pronouns

[8] https://www.wikidata.org/wiki/Wikidata_talk:Lexicographical_data/Archive/2018/11#How_to_split_or_merge_stedord_(Q36224).

are split, a question is how to link between such different lexemes. A related issue for Danish appears for some adverbs, which could easily be regarded as separate lexemes, such as *hjem*, *hjemad* and *hjemme* (home, homeward, at home). Here the words are distinguished by telicity and a dynamic/static feature, e.g., *hjemad* is atelic and dynamic (Hansen and Heltoft, 2019, p. 216). Possibly new specific properties could describe the relationships.

Wikidata's choice of separating form and sense complicates modeling of some words. *vand* (*water*), *øl* (*beer*) and *tøj* (*cloths*) are examples of words that each are regarded as one lexeme but where the specific forms are associated with specific semantics: The common gender version of *øl* relates to a countable noun as in "one beer", while the neuter version relates to a mass noun. Here we could split the lexeme into two Wikidata lexemes, e.g., *vand* with common gender and *vand* with neuter, but that would complicate their relations to other lexemes, e.g., in terms of compounding and etymology. A related issue occurs for deponent verbs. For *finde*/*findes* (active/passive; find/exist) the lexemes have been separated (L39637 and L44601) following the convention of DanNet.

In case of, e.g., the lexemes *mor* and *moder* (mother) their singular forms are different but they have the same meaning and their plural form, *mødre*, is the same. There is no way of merging the separate plural forms when *mor* and *moder* are regarded as separate lexemes as the forms are tied to separate lexeme pages. The creation of a dedicated property could link such forms together.

5.2 Compound splitting

Danish is a language rich in compounds. The compounds and affixes of a lexeme can be specified with the P5238 property where other lexemes can be linked. The currently longest Danish lexeme in Wikidata, *ejendomsadministrationsvirksomhed* (building administration business), could be split as *ejendom-s-administration-s-virksomhed* with two *s*-interfixes and three words with a good semantic relation to the complete lexeme. With a more granular level, the word could be split into *ejen-dom-s-ad-ministr-ation-s-virk-somhed*, where affixes have been split from the roots. Here, *dom* and *virk* has little semantic relationship to the compound lexeme. With the current setup of the P5238 property and the structure of Wikidata, it is difficult to see how the two splits can coexist with the same lexeme. Currently, we typically split on the highest level, e.g., *ejendomsadministration-s-virksomhed*. The lexeme pages for *ejendomsadministration* and *virksomhed* can further split the compounds and derived words.

5.3 Linking compounds to parts

When orthographically similar words with the same etymology are split across multiple lexemes it may be unclear which lexeme a compound derives from. For instance, the compound *vaskemaskine* (L42991, washing machine) could be analyzed as consisting of: 1) a verb stem (*vask*), an interfix (*-e-*) and a noun (*maskine*), or 2) a verb in its infinitive form (*vaske*) and a noun, or 3) a noun (*vask*), an interfix and a noun. During data entry one would need to make an explicit choice. The same choice may appear for affixations, such as *for-be-handle*. While *be-* is arguably a prefix (L44579), *for* may be a prefix or an adverb, — or possibly an preposition.

5.4 Genitive

Danish genitive, where an *-s* suffix is added, has traditionally been regarded as a case, but newer words for Danish grammar challenge that notion and argues that it is a clitic and a derivation making a nominal to non-nominal (Hansen and Heltoft, 2019, p. 255). Originally, we began adding the genitive *-s* forms for the Danish nouns, but has discontinued it after becoming aware of the issue. The Swedish part of Wikidata lexeme continues to add the genitive *-s* forms for nouns. If we were to add the genitive for Danish nouns, then one could argue that genitive versions of other word classes should also be added, — as words from other word classes can be used as nouns, e.g., *de rødes valgsejr* (literally, *the reds' election victory*) where the adjective *røde* has the added genitive *-s*. The advantage of have the *-s* forms is that lookup, e.g., for spellchecking may be more convenient. Other Danish digital dictionaries record the *-s* form.

5.5 Data quality

The structured format of the data and the query tools associated with Wikidata enable us to perform some completeness and internal consistency checks. For instance, we may formulate a SPARQL query that returns Danish lexemes without any usage examples. We have used the Shape Expressions (ShEx) language (Q57, 2017) to formalize such checks, and these ShEx definitions are available on separate pages on Wikidata (Nielsen et al., 2019). As an example, the ShEx definition E65[9] checks Danish numerals regarding data about language, lemma, word stem, word class, DanNet, usage example, sense, form and hyphenation.

[9] https://www.wikidata.org/wiki/EntitySchema:E65

5.6 Applications

What can Wikidata lexemes be used for? Wikidata itself has a dedicated page for application ideas.[10] For spellchecking the current number of lexemes in Wikidata can hardly compete with already established larger word lists, but in the long run using the lexeme forms for spellchecking might be of interest. The advantage is that it is collaboratively extensible, likely able to quickly catch up on neologisms and evolving jargon in comparison to standard dictionaries. It is less clear if Wikidata lexemes can be used for more advanced natural language processing, such as part-of-speech tagging and grammar checking. The ability of the current Wikidata lexeme system has limited means for specifying grammar.

6 Related Research

Among related research, there are several studies reporting the extraction of data from Wiktionary and using the structured data for linguistic tasks or building a resource (Zesch et al., 2008; McCrae et al., 2012; Sérasset, 2014; Pantaleo et al., 2017). For instance, the Java- and database-based system by Zesch et al. (Zesch et al., 2008) for reading, storing and querying lexical semantic knowledge from Wikipedia and Wiktionary enables a user, e.g., to programmatically query for hyponyms of senses. The parser of the described system needs to be adjusted for each language edition of Wiktionary as each edition may use different markup for the lexical semantic information.

The lexicographic part of Wikidata is still comparably small, but contrary to many other online dictionaries with rich semantics, Wikidata users can add and edit the lexicographic information and more or less immediately see it becoming available in the powerful query facility of the SPARQL-based Wikidata Query Service. Our Ordia Web application at http://tools.wmflabs.org/ordia takes advantage of this service (Nielsen, 2019).

7 Acknowledgment

We thank Bolette Sandford Pedersen, Sanni Nimb, Sabine Kirchmeier, Nicolai Hartvig Sørensen and Lars Kai Hansen for discussions and answering questions, and the reviewers for suggestions for improvement of the manuscript. 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).

[10] https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Ideas_of_tools.

References

[Allan et al.1995] Robin Allan, Philip Holmes, and Tom Lundskær-Nielsen. 1995. Danish.

[Bond et al.2016] Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. *Proceedings of the Eighth Global WordNet Conference*, pages 50–57, January.

[Hansen and Heltoft2019] Erik Hansen and Lars Heltoft. 2019. Grammatik over det Danske Sprog. February.

[Koehn2005] Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *The Tenth Machine Translation Summit: Proceedings of Conference*, pages 79–86.

[McCrae et al.2012] John P. McCrae, Elena Montiel-Ponsoda, and Philipp Cimiano. 2012. Integrating WordNet and Wiktionary with lemon. *Linked Data in Linguistics*.

[Miller1995] George Armitage Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38:39–41, November.

[Navigli and Ponzetto2010] Roberto Navigli and Simone Paolo Ponzetto. 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 and Hansen2018] Finn Årup Nielsen and Lars Kai Hansen. 2018. Inferring visual semantic similarity with deep learning and Wikidata: Introducing imagesim-353. *Proceedings of the First Workshop on Deep Learning for Knowledge Graphs and Semantic Technologies*, pages 56–61, April.

[Nielsen et al.2019] Finn Årup Nielsen, Katherine Thornton, and José Emilio Labra Gayo. 2019. Validating Danish Wikidata lexemes. June.

[Nielsen2018] Finn Årup Nielsen. 2018. Linking ImageNet WordNet Synsets with Wikidata. *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France*, pages 1809–1814, April.

[Nielsen2019] Finn Årup Nielsen. 2019. Ordia: A Web application for Wikidata lexemes. May.

[Pantaleo et al.2017] Ester Pantaleo, Vito Walter Anelli, Tommaso Di Noia, and Gilles Sérasset. 2017. Etytree: A Graphical and Interactive Etymology Dictionary based on Wiktionary. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, pages 1635–1640.

[Pedersen et al.2009] Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik

Lorentzen. 2009. DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43:269–299, August.

[Q572017] 2017. Shape Expressions (ShEx) Primer. July.

[Q592016] 2016. Lexicon Model for Ontologies: Community Report, 10 May 2016. May.

[Sérasset2014] Gilles Sérasset. 2014. DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web: interoperability, usability, applicability.*

[Vrandečić and Krötzsch2014] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57:78–85, October.

[Zesch et al.2008] Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1646–1652.