Validating Danish Wikidata lexemes

Finn Årup Nielsen[†], Katherine Thornton[*], Jose Emilio Labra Gayo[‡]

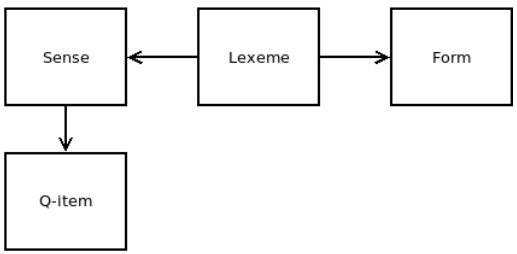
[†]Cognitive Systems, DTU Compute, Technical University of Denmark, Kongens Lyngby, Denmark; [*]Yale University Library, New Haven, CT, USA; [‡]University of Oviedo, Spain

Summary

Two of the newest features of Wikidata are support for lexicographic data (lexemes), and support for Shape Expressions (ShEx). We demonstrate the first application of ShEx for validation of entity data for Wikidata lexemes. Validation of entity data in Wikidata against ShEx schemas allows editors to discover missing or incorrect information. It may also form a basis for discussion of the data models implicitly used in Wikidata. We present a use case and benchmark for ShEx and discuss its current limitations.

Wikidata lexemes

Wikidata[1] has as of September 2019 over 70,000 lexemes, see https://tools.wmflabs.org/ordia/statistics/. There are lexemes from over 300 languages, including Danish lexemes.[2] These lexemes can be described by properties specifying forms, senses, languages, lexical categories, grammatical features, hyphenation, etc.



Currently over 2,500 Danish lexemes are recorded: nouns, verbs, adjectives, numerals, adverbs, etc.

Validating Wikidata

For some time, Wikidata has had the ability to constrain and validate its data via several means:

1. Datatype restrictions, e.g., the wdt:P31 property will only accept other Wikidata items as values, — not literal values.
2. Literal value restrictions via regular expressions, e.g., for the DanNet[7] property: "`((\d{8})(-\d+)?)|(temporary_\d+)`"
3. Property constraints, e.g., "single value constraint" and "distinct values constraint"
4. Formulation of SPARQL queries via *Wikidata Query Service*, e.g., "find every lexeme without any form".

ShEx

ShEx (Shape Expressions) is a concise, formal language for modeling and validating RDF graphs.[3,4]

Since May 2019, Wikidata editors can collaboratively edit pages with ShEx schemas and subsequently use them for validating Wikidata entities.

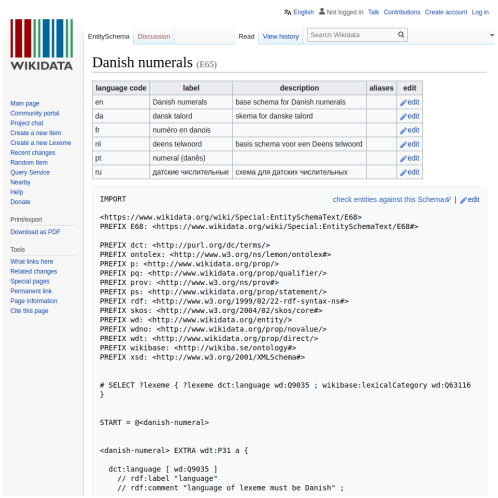
A ShEx schema may, e.g., check that a lexeme is defined with a specific language:

```
START = @<danish-numeral>
<danish-numeral> {
  dct:language [ wd:Q9035 ]
}
```

ShEx for Danish lexemes

We wrote ShEx schemas for Danish lexemes with the identifiers E15 (Danish lexeme), E34 (Danish noun), E54 (lexeme), E56 (Danish verb), E62 (Danish pronoun) and E65 (Danish numeral) as well as a ShEx for Danish hyphenation E68.

Here is the top of Danish numeral schema which is defined at https://www.wikidata.org/wiki/EntitySchema:E65:



With the ShEx `IMPORT` statement part of a ShEx schema may be reused in another schema.

In the current ShEx schema for Danish numerals, the hyphenation rules are imported from another ShEx schema.

Example rules for Danish lexemes

Many rules exist for Danish lexemes. Some can be gleaned from works on Danish grammar.[5,6] A few examples are:

1. All Danish Wikidata lexemes should have one unique value for DanNet[7] words, — either one unique identifier or no value. Proper Danish nouns, adverbs, pronouns and words from a number of other word classes should *not* have an associated DanNet identifier.
2. A Danish noun should have one single grammatical gender, either common gender or neuter.
3. Each part of a hyphenated representation should contain a vowel.
4. The grammatical gender of a compound should have the same grammatical gender as the final lexeme of the compound, except for compounds suffixed *-fuld*.

Some rules can be specified concisely in ShEx, e.g., the first rule for DanNet can be specified in one line as, e.g., `a [ wdno:P6140 ] | ps:P6140 /^[0-9]{8}$/`

The second rule for grammatical gender may at least initially be formulated on one line `wdt:P5185 [ wd:Q1305037 wd:Q1775461 ]`. Certain words require an exception to this rule.

The fourth rule is possible in ShEx, but our implementation is a verbose schema with enumeration over number of compound parts with a considerable number of AND, OR and NOT to accomodate all individual cases.

ShEx example

Part of the ShEx schema for Danish noun forms, specifying restrictions on which grammatical features should be used, hyphenation (rule defined elsewhere) and the inflection, e.g., Danish plural definite nouns should end with "er?ne":

```
<danish-form> {
  wikibase:grammaticalFeature [
    wd:Q110786  # singular
    wd:Q146786  # plural
    wd:Q53997857  # indefinite
    wd:Q53997851  # definite
    wd:Q146233  # genitive
  ] {1,3} ;

}
AND @<hyphenation>
AND (
    NOT @<singular-definite-not-genitive>
  OR
    @<represensation-ends-with-en-et>
  )
AND (
    NOT @<plural-definite-not-genitive>
  OR
    @<represensation-ends-with-erne>)

# ...

<represensation-ends-with-erne> {
  ontolex:representation /^.+er?ne$/
    // rdf:label "plural definite ending"
    // rdf:comment
     "representation ends with '-e(r)ne'"
    ;
}
```

Here the AND-NOT-OR pattern specifies an IF-THEN pattern, e.g., if the form is plural definite and not genitive then form should end with "er?ne".

The current definition is on the page https://www.wikidata.org/wiki/EntitySchema:E34

Validation with ShEx

When ShEx expressions are defined, they may be used for validating Wikidata items. The default tool is ShEx2 Simple Online Validator at https://tools.wmflabs.org/shex-simple/wikidata/.

The validator requires a list of Wikidata items to be tested. These can be provided by a SPARQL query, e.g., all Danish lexemes can be found with `SELECT ?lexeme { ?lexeme dct:language wd:Q9035 }`

The validator generates a report pinpointing non-conforming Wikidata items.

Here the validator reports missing data about the hyphenation of several forms of the Danish noun *forhindring* (L54102).

Discussion

Why do we write ShEx for Danish lexemes in Wikidata?

1. To identify missing data, e.g., missing DanNet identifier.
2. Identify wrong/inconsistent data, e.g., a form specified to be singular when it is plural
3. Use as basis for concrete discussions among editors about approaches for improving the data.

The validation discovered numerous issues. Most of the non-conformant items we discover are errors of omission (e.g., missing grammatical gender), rather than errors of commission (e.g., wrong grammatical gender).

Acknowledgment

This work is funded by the Innovation Fund Denmark through DABAI and ATEL projects. The third author is partially funded by the Spanish Ministry of Economy and Competitiveness (Society challenges: TIN2017-88877-R). We appreciate feedback from Eric Prud'hommeaux on the ShEx schemas.

References

[1] Vrandečić D and Krötzsch M. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 2014;57:78–85.
[2] Nielsen FÅ. Danish in Wikidata lexemes. 2019;.
[3] Prud'hommeaux EG, Gayo JEL, and Solbrig H. Shape expressions: an RDF validation and transformation language. *SEM '14: Proceedings of the 10th International Conference on Semantic Systems*, 2014;.
[4] Thornton K, et al. Using Shape Expressions (ShEx) to Share RDF Data Models and to Guide Curation with Rigorous Validation. *The Semantic Web*, 2019;pages 606–620.
[5] Allan R, Holmes P, and Lundskær-Nielsen T. Danish. 1995;.
[6] Hansen E and Heltoft L. Grammatik over det Danske Sprog. 2019;.
[7] Pedersen BS, et al. DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 2009; 43:269–299.