Ordia: A Web application for Wikidata lexemes

Finn Årup Nielsen

Cognitive Systems, DTU Compute, Technical University of Denmark, Kongens Lyngby

Summary

Ordia is a web application for Wikidata lexemes querying the *Wikidata Query Service* on-the-fly when the user navigates the Ordia interface.

A user may view individual lexeme, their forms and senses as well as its compound and derivation graph.

Other webpages on Ordia aggreates lexemes for language, lexical category, grammatical feature, properties or references.

Ordia has also special features: text-to-lexemes and integration of knowledge graph embedding similarity results.

Wikidata lexemes

Wikidata has as of May 2019 over 46,000 lexemes. Ordia has a dedicated page that shows live statistics:

Statistics

Count	Description	Query
236224	Number of grammatical feature links	[] wikibase:grammaticalFeature []
127789	Number of forms	[] a ontolex:Form
46193	Number of lexical category links	[] wikibase:lexicalCategory []
46176	Number of language links	[] dct:language []
46168	Number of lexemes	[] a ontolex:LexicalEntry
12265	Number of sense links	[] ontolex:sense []
12172	Number of senses	[] a ontolex:LexicalSense
6775	Number of sense to item links	[] wdt:P5137 []

Ordia's Stack

Python, Flask and Jinja: Standard Python-based web stack.

Wikidata Query Service: Extended SPARQL endpoint at https://query.wikidata.org/ with visualization capabilities.

Javascript, JQuery, DataTables: Formatting the result from the Wikidata Query Service, e.g., for better clickable tables.

GitHub: Repository at https://github.com/fnielsen/ordia.

Wikimedia Toolforge: The cloud infrastructure provided by the Wikimedia Foundation hosts the canonical website of Ordia at https://tools.wmflabs.org/ordia/.

Ordia is inspired by Scholia.[1]

Conceptual choices

The user should easily be able to perform powerful SPARQL queries by navigating the Ordia Interface

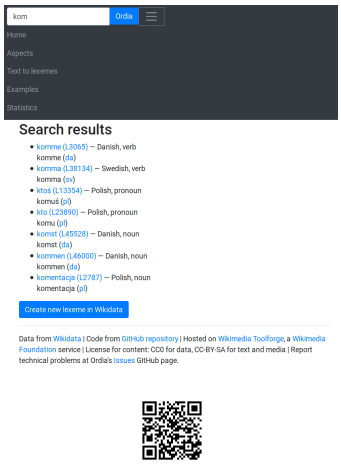
URLs should be predicable, e.g., /lexical-category/Q24905 shows the Wikidata item for verb (Q24905) as a lexical category; /property/P31/value/Q639512 is the aspect for instance of (P31) cranberry morpheme (Q639512).

Each page should link to other pages and let the user discover new lexicographic relations.

The interface should use graphics whenever possible, e.g., graphs for lexeme and concept relations and for displaying images associated with senses of lexemes.

Searching

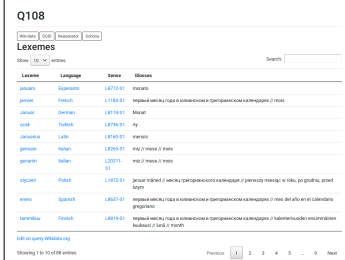
Ordia may be used to search Wikidata for lexemes and their forms. Ordia uses the Wikidata API for this function (this is also possible in Wikidata itself via the "L:" prefix).



The Javascript *DataTable* package allows the user to filter results displayed in tables in Ordia on the client side.

Knowledge graph embedding

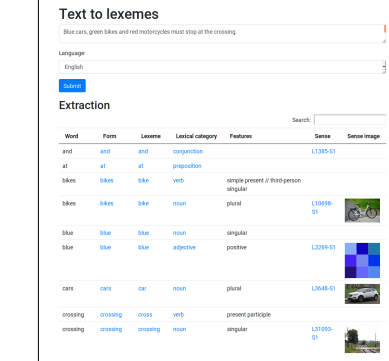
Ordia can shown any of the Q-items of Wikidata. These pages show the linked lexemes and senses as well as a list of related Wikidata items obtained through the RDF2Vec-inspired[2] Wembedder web-service[3] which implement a similarity measure via knowledge graph embedding. A table is constructured by interpolating the Wembedder API into a SPARQL query sent to Wikidata Query Service. Here it is for January (Q108), where June and March are the most similar concepts.



Text-to-lexemes

The text-to-lexemes facility in Ordia enables a user to enter a text in the web interface and let Ordia extracts sentences and words.

Extracted words are sent for matching to the Wikidata Query Service and Ordia displays the result of the matching in a table.



Here the sentence for text-to-lexemes is "Blue cars, green bikes and red motorcycles must stop at the crossing." and the first part of the returned table is shown.

In this case all words are matched to lexemes in Wikidata. If some words were not matched then Ordia creates convenient links for entry of the lexemes in Wikidata.

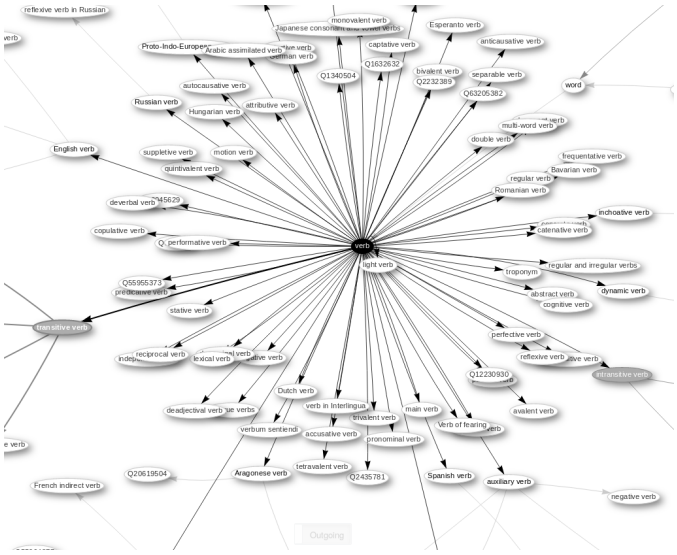
https://tools.wmflabs.org/ordia/text-to-lexemes?text=Blue+cars,+green+bikes+and+red+motorcycles+must+stop+at+the+crossing.&text-language=en

Aspects

A number of so-called *aspects* show specific information related to: languages, lexical categories, grammatical features, properties or references.

Each of these aspects has an dynamically updated index page (e.g., /language for language) and dynamically constructed pages for individual Wikidata items (e.g., /language/Q9035 for the Danish language).

Here is the ontology graph that is part of the *lexical category* aspect for *verbs* (Q24905)



The black node is the concept of interest (verb) while arrows point to subconcepts (hyponyms).

"Transitive verb" and "intransitive verb" concepts are greyed to indicate that they are used as lexical categories for lexemes in Wikidata.

Current use

Copy-and-paste of a text into Ordia's text-to-lexeme facility to see if all words match Wikidata lexemes, and if not enable the user to quickly create the missing lexemes.

Get an overview of the Wikidata linguistic ontology and its usage, from, e.g., the lexical category aspect.

Searching for lexemes

View statistics, e.g., number of lexemes, number of forms, lexemes per language and lexemes per lexical category.

Future

Current annotation for Wikidata lexemes is too sparse for Ordia to make sense as a cross-language dictionary, etymological dictionary or synonym dictionary.

Possible future extensions could include internationalization of the user interface and Wikidata entry directly in Ordia.

Acknowledgment

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).

References

[1] Nielsen FÅ, Mietchen D, and Willighagen E. Scholia, Scientometrics and Wikidata. *The Semantic Web: ESWC 2017 Satellite Events*, 2017;.

[2] Ristoski P and Paulheim H. RDF2Vec: RDF Graph Embeddings for Data Mining. *The Semantic Web – ISWC 2016*, 2016;pages 498–514.

[3] Nielsen FÅ. Wembedder: Wikidata entity embedding web service. 2017;.