
Lexicographical Data
is coming
on Wikidata
Léa Lacroix
Leszek Manicki

What’s the need?

A lot of information is duplicated on Wiktionaries

+

A lot of people ask how to query and reuse information from Wiktionaries

We will provide

Data about words and phrases
Freely reusable
Structured to be machine-readable
Improved by a multilingual community

that can be used by:

Wiktionaries

Wikidata

Students

NGOs

Wikisource

Scientists

Public institutions

Start-ups

Data-journalists

What can we do once we have it?

Workload sharing and new ways to contribute to Wiktionary

- Working together on the same data (if wanted!)
- New tools to make contributing easier and open it up to new contributor groups

Potential users: Wiktionary, Wikidata Game

Dictionary applications and more

- Looking up definitions and translations
- Special purpose dictionaries (rhyme, specific topics)
- Thesauri and synonym dictionaries
- Build translation tools
  (especially for underserved languages that don't have any yet)

Potential users: Leo, Apertium

Language learning tools

- Creating word lists and lessons
- Illustrating words
- Creating games and exercises

Potential users: Parley, Duolingo

Research

- How do languages evolve over time, social class and more?
- Do classes of words change their meaning over time?
- Localizing words on maps

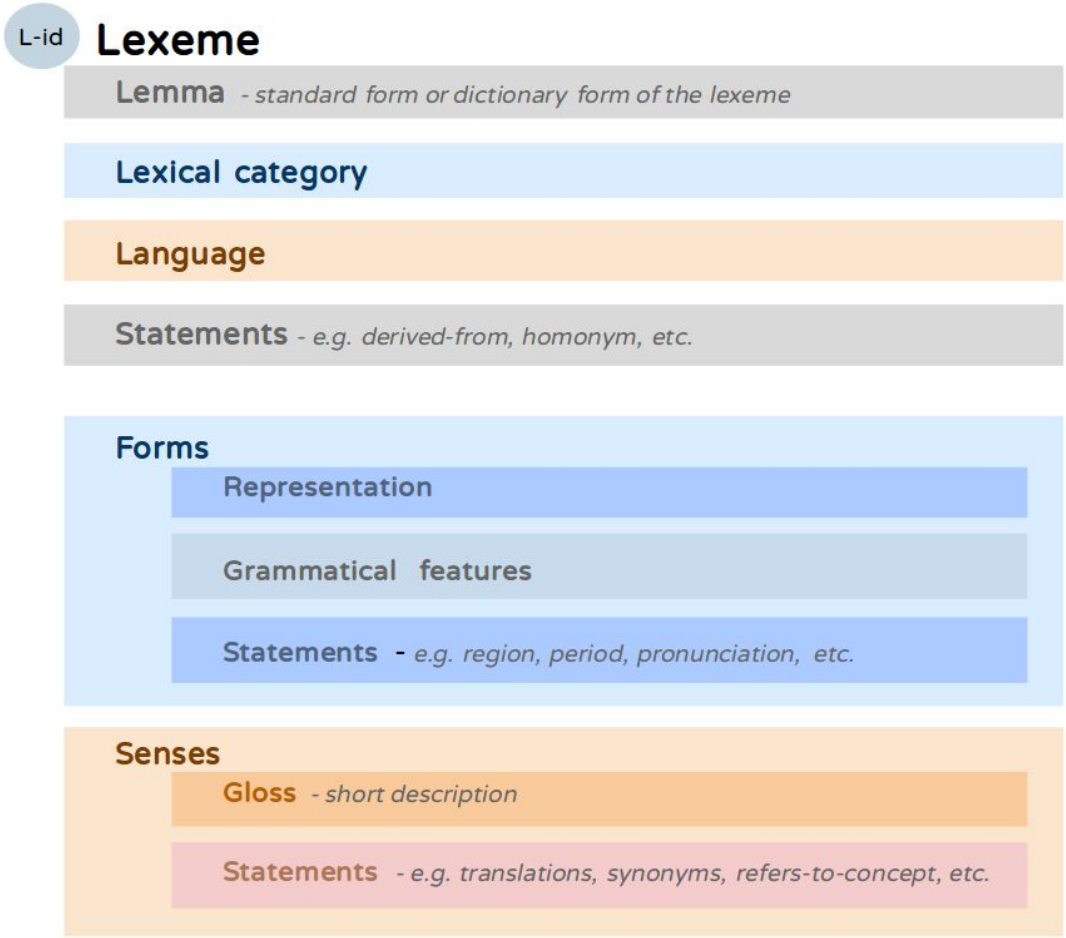
Potential users: The Rosetta Project

Text analysis

- Sentiment analysis
- Part of speech tagging
- Named entity recognition

Potential users: TextRazor, Wikisource

How is it going to work?



More info: mw:Extension:WikibaseLexeme/Data Model

(L15) # Leiter
de

edit

Language German
Lexical Category noun

Statements

morphology	German declination W1 0 references + add reference	edit
	+ add value	
gendered form	Leiterin 0 references + add reference	edit
	+ add value	

+ add statement

Forms

L15-F1 Leiter
de

edit

Grammatical features

Statements about L15-F1

IPA pronunciation	ˈlaɪ̯.tɐ 0 references + add reference	edit
	+ add value	
hyphenation	Lei-ter 0 references + add reference	edit
	+ add value	

Questions?

On the technical side

- MediaWiki extension built on top of Wikibase
  - Source code at Gerrit
  - Translate UI at TranslateWiki
- API
- Dumps
- Linked Data Interface
- Wikidata Query Service
- Stable Gadgets interface does NOT exist at this point

API

- MediaWiki-powered web API
- Allows getting, editing, and searching lexicographical data
- Pywikibot friendly
- https://www.wikidata.org/w/api.php
- Actions to start with:
  - wbgetentities
  - wbeditentity
  - wbsearchentities
- Try out in API Sandbox

Dumps

- Lexeme data for offline use, e.g. with tools like Wikidata Toolkit
- Formats: RDF (ttl, nt), JSON
  - https://dumps.wikimedia.org/wikidatawiki/entities/
- XML dumps also exist
  - https://dumps.wikimedia.org/wikidatawiki/
  - https://dumps.wikimedia.org/other/incr/wikidatawiki/
- RDF mapping: Work In Progress

Linked Data Interface

- Formats: RDF, JSON, HTML
- https://www.wikidata.org/entity/..
- RDF mapping: Work In Progress

Wikidata Query Service

- Query Lexicographical Data (and other Wikidata data) using SPARQL
- https://query.wikidata.org/
- Status: Work In Progress (RDF mapping)

More questions?

Next steps

- Now: you can try the demo system!
  Come to the team for a 1:1 test during the hackathon
- Also now: participate in discussions
  d:Wikidata:Lexicographical data
- May 23rd: deployment of the first version
  adding, editing, connecting Lexemes and Forms
- Later: regular deployment with new features
  Senses, better search, SPARQL queries...

Contact

Léa Lacroix
User: Lea Lacroix (WMDE)
lea.lacroix@wikimedia.de

Community discussions:
d:Wikidata:Lexicographical data

Thanks for your attention :)