Lexicographical data

Getting Started with Modeling Words in Your Language

Lydia Pintscher
Wikidata Portfolio Lead,
Wikimedia Deutschland

Mohammed Abdulai
Wikidata Community
Communication Manager,
Wikimedia Deutschland

This session is recorded: Please mute your microphone and camera when you're not speaking.

Wikidata Lexicodays
28-30 June 2024

Agenda

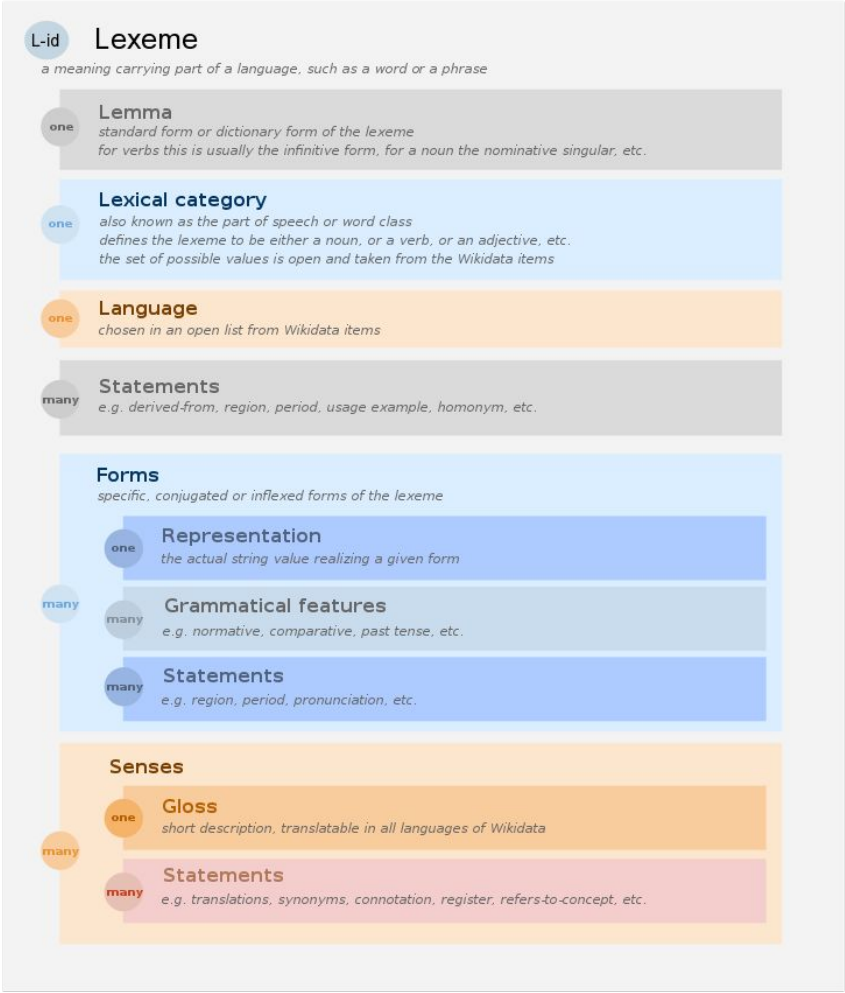
- Lexeme Data Model: A Quick Recap
- Creating Your First Lexeme
- Best Practices For Modeling Words
- Resources And Tools

Lexeme data model: A quick recap

Wikidata Lexicodays
28-30 June 2024

Glossary:
Wikidata:Lexicographical_data/Glossary

Data model:
MediaWiki:Extension:WikibaseLexeme/Data Model



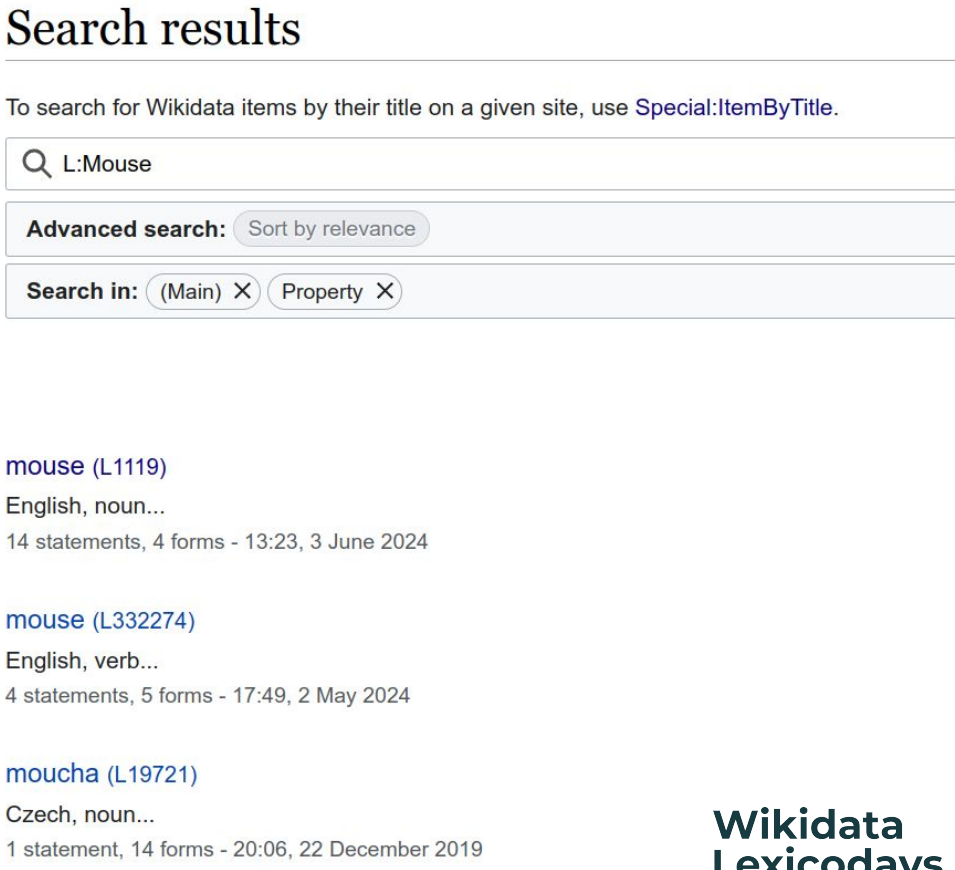
Wikidata Lexicodays
28-30 June 2024

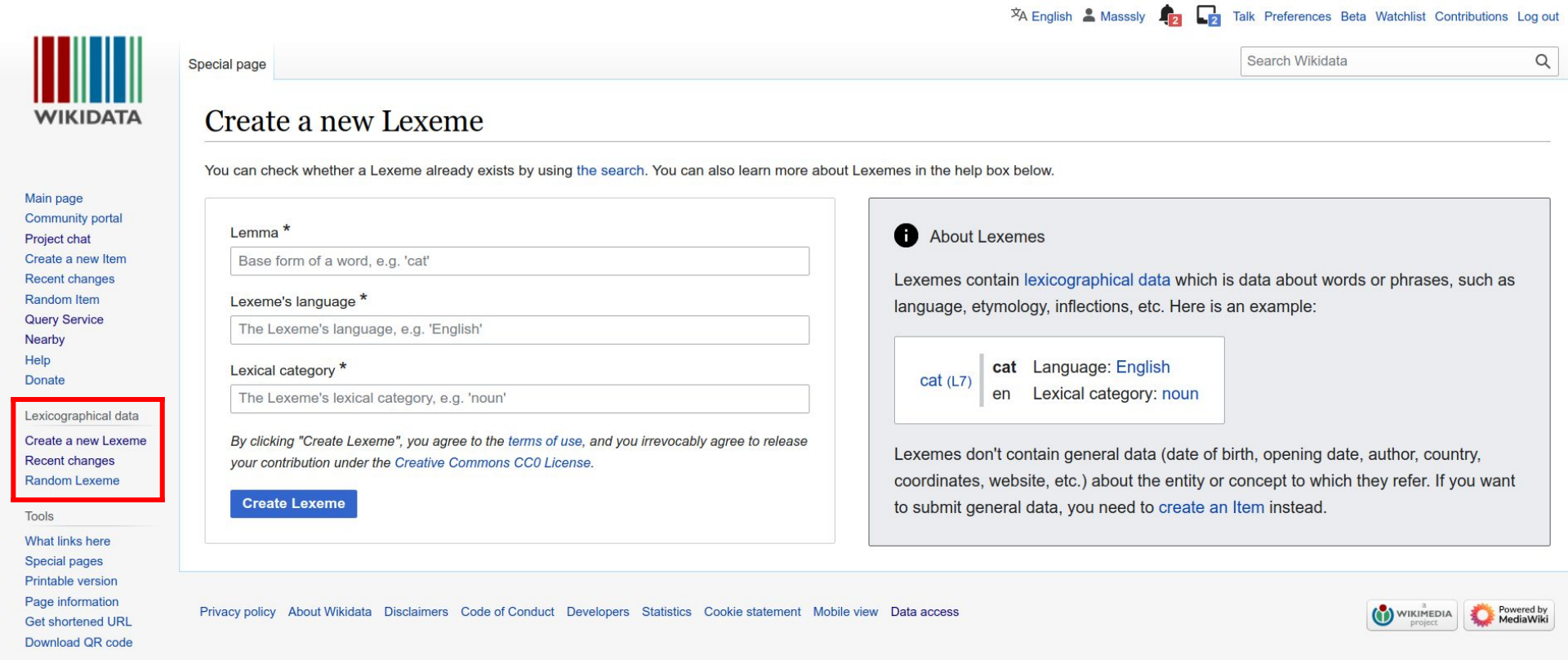
Creating Your First Lexeme

Wikidata Lexicodays
28-30 June 2024

Search for Lexemes

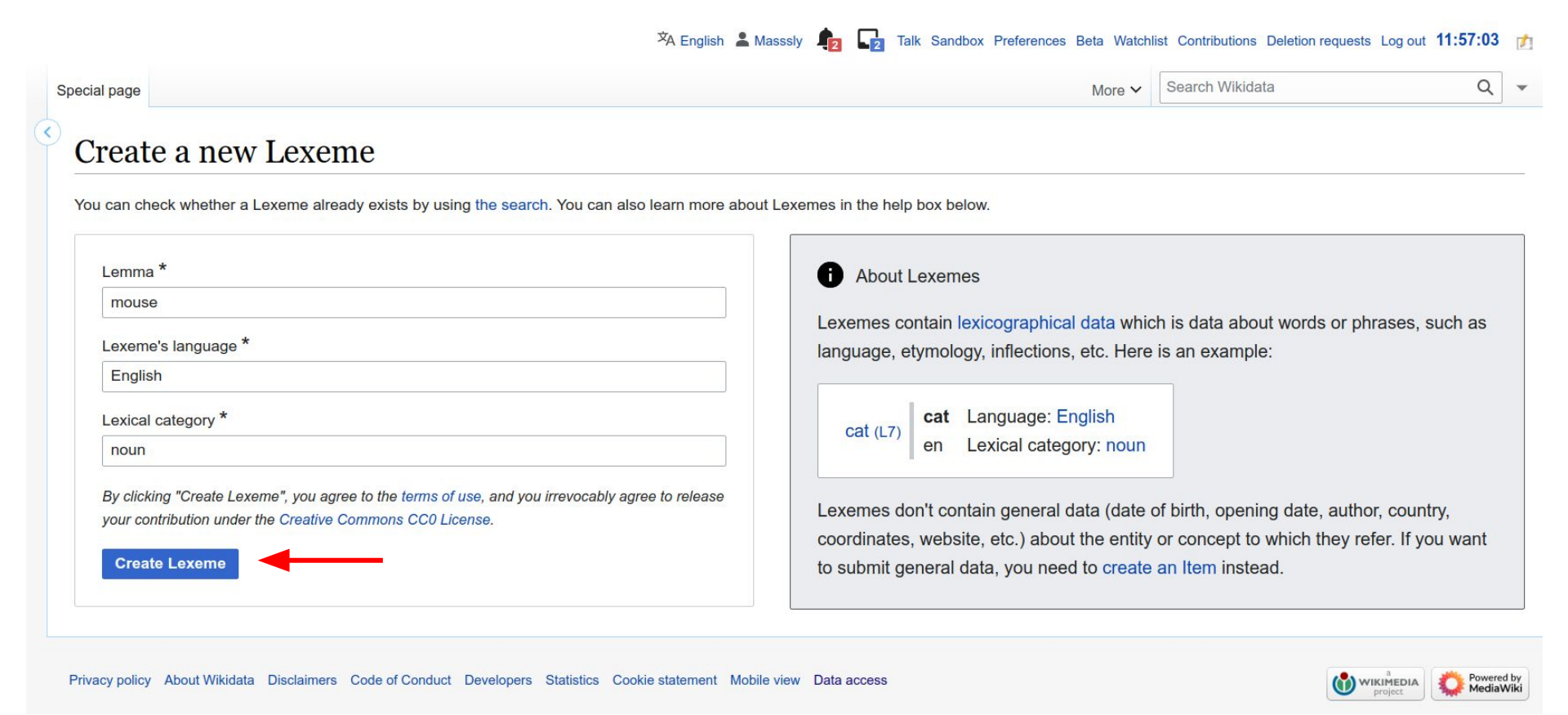
- Type “L:word” in the search box and press Enter
- Works with lemmas and Forms
- No suggestion appears, don’t worry





Go to https://www.wikidata.org/wiki/Special:NewLexeme

Wikidata Lexicodays
28-30 June 2024



Hit the “Create Lexeme” button

Wikidata Lexicodays
28-30 June 2024

(L4223)

mouse

en

edit

Language English

Lexical category noun

Statements

+ add statement

Senses

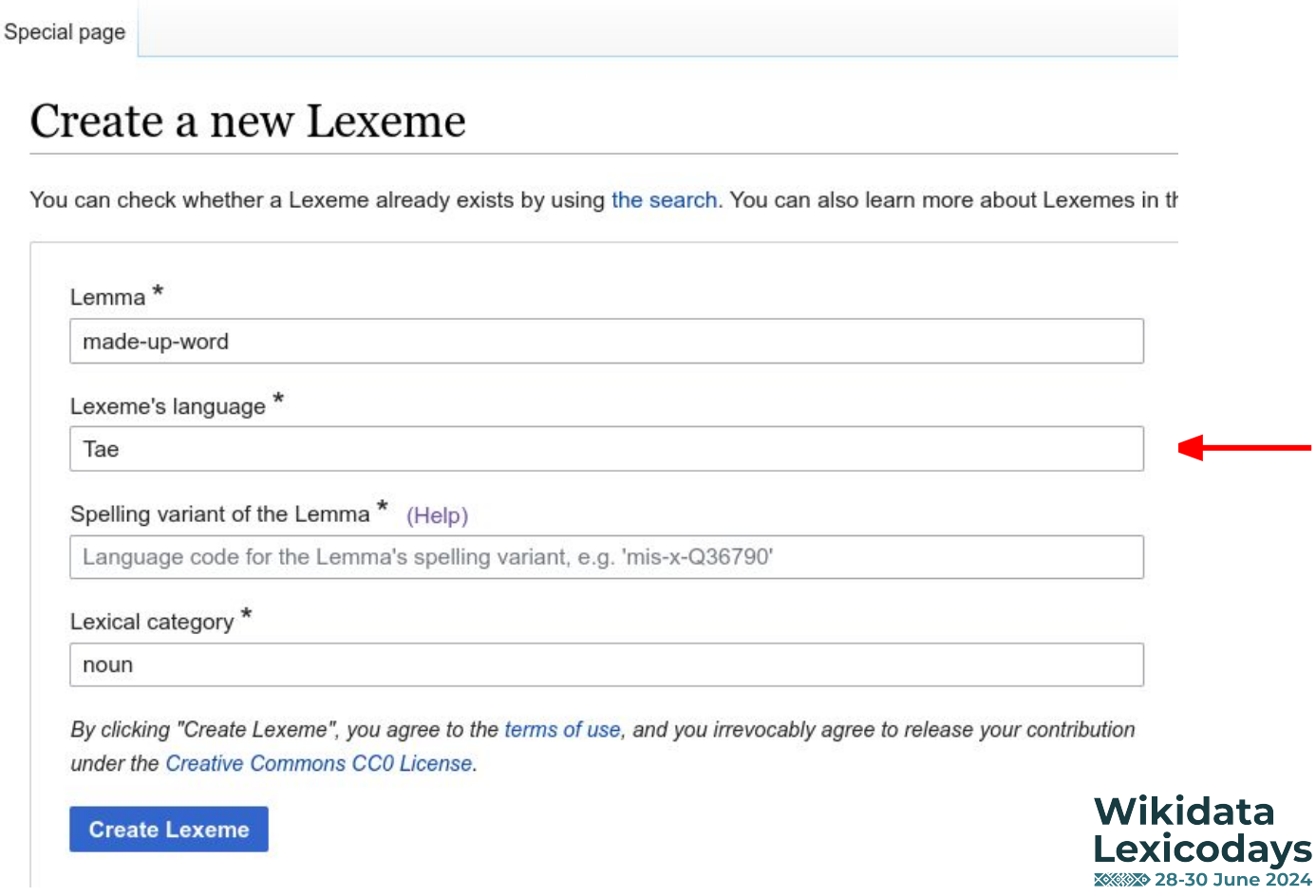
+ add Sense

Forms

+ add Form

Take note:

Not all languages are supported for Lexemes yet!



Wikidata Lexicodays
28-30 June 2024

Special page

Create a new Lexeme

You can check whether a Lexeme already exists by using the search. You can also learn more about Lexemes in th

Lemma *

made-up-word

Lexeme's language *

Tae

Spelling variant of the Lemma * (Help)

Tae

Lexical category *

noun

By clicking "Create Lexeme", you agree to the terms of use, and you irrevocably agree to release your contribution under the Creative Commons CC0 License.

Create Lexeme

Wikidata Lexicodays
28-30 June 2024

For unsupported languages:

Special page

Create a new Lexeme

You can check whether a Lexeme already exists by using the search. You can also learn more about Lexemes in th

Lemma *

made-up-word

Lexeme's language *

Tae

Spelling variant of the Lemma * (Help)

unsupported language (mis)

Lexical category *

noun

By clicking "Create Lexeme", you agree to the terms of use, and you irrevocably agree to release your contribution under the Creative Commons CC0 License.

Create Lexeme

Select the “mis” language variant

(L4222)

made-up-word

mis

edit

Language Tae

Lexical category noun

Statements

+ add statement

Senses

+ add Sense

Forms

+ add Form

Now you can...

- Add statements (eg. auxiliary verb)
- Add a Sense
- Add translation (needs another Sense)
- Add a Form (present, first-person singular, etc)

Lexeme | Discussion | Read

(L1119)

mouse

en

edit

Language English
Lexical category noun

Statements

instance of	count noun 0 references + add reference + add value	edit
described by source	Merriam-Webster online dictionary 0 references + add reference + add value	edit
usage example	When the mouse laughs at the cat, there's a hole nearby. (English) 0 references	edit

Wikidata Lexicodays
28-30 June 2024

Senses

L1119-S1

English	any small rodent of the genus Mus
Spanish	mamifero roedor de pequeño tamaño, del género Mus
Persian	پستانداری کوچک از راستهٔ جوندگان
French	petit rongeur du genre Mus
Hebrew	מכרסם קטן מהסוג Mus
Kannada	ಇಲಿಯು ದಂಶಕಗಳ ಗಣಕ್ಕೆ ಸೇರಿದ
Portuguese	qualquer pequeno roedor do gênero Mus

edit

Statements about L1119-S1

image

Apodemus sylvaticus bosmuis.jpg
2,160 × 1,524; 2.53 MB

0 references

+ add reference

edit

+ add value

item for this sense

mouse

0 references

+ add reference

edit

Senses

L1119-S1

English	any small rodent of the genus Mus
Spanish	mamifero roedor de pequeño tamaño, del género Mus
Persian	پستانداری کوچک از راستهٔ جوندگان
French	petit rongeur du genre Mus
Hebrew	מכרסם קטן מהסוג Mus
Kannada	ಇಲಿಯು ದಂಶಕಗಳ ಗಣಕ್ಕೆ ಸೇರಿದ
Portuguese	qualquer pequeno roedor do gênero Mus

edit

Statements about L1119-S1

image

Apodemus sylvaticus bosmuis.jpg
2,160 × 1,524; 2.53 MB

0 references

+ add reference

edit

+ add value

item for this sense

mouse

0 references

+ add reference

edit

L1119-F2 | mice
en

edit

Grammatical features plural

Statements about L1119-F2

+ add statement

L1119-F3 | mouse's
en

edit

Grammatical features genitive case

Statements about L1119-F3

+ add statement

L1119-F4 | mouses
en

edit

Grammatical features plural

💡 Pro tips!

- Look at what already exists before creating
- Find an example for inspiration
- Look for the most used properties in Ordia https://tools.wmflabs.org/ordia/property/
- Try, don't be afraid of breaking anything :)
- In doubt: ask the other editors for help (LexData talk page, Telegram)

Best Practices for Modeling Words

Wikidata Lexicodays
28-30 June 2024

Should There Be a Lexeme for It?

- Evidence of a Lexeme’s existence in a language is crucial when creating it on Wikidata.
- For well-documented languages (e.g., English, French), this is obligatory.
- For less documented languages, it’s a strong recommendation.
- Evidence can be external identifiers, described-by sources, reference URLs, usage examples, or gloss quotes.

Lemmata

- The lemma of a lexeme should ideally be the representation found in a dictionary.
- This will generally depend on the Lexeme's language and lexical category.
- For languages with multiple scripts, the lemma should represent each script.

Lexical Categories (aka “Parts of Speech”)

- Refer to different types of words in a language
- That help us understand how words function in sentences
- `instance of (P31)` value on a Lexeme should be more specific than the Lexeme's lexical category.

Senses

- Senses describe the different meanings of a Lexeme
- Therefore, each Lexeme should have at least 1 sense
- Qualify each sense with an `item for this sense (P5137)` statement where an Item exists

Forms

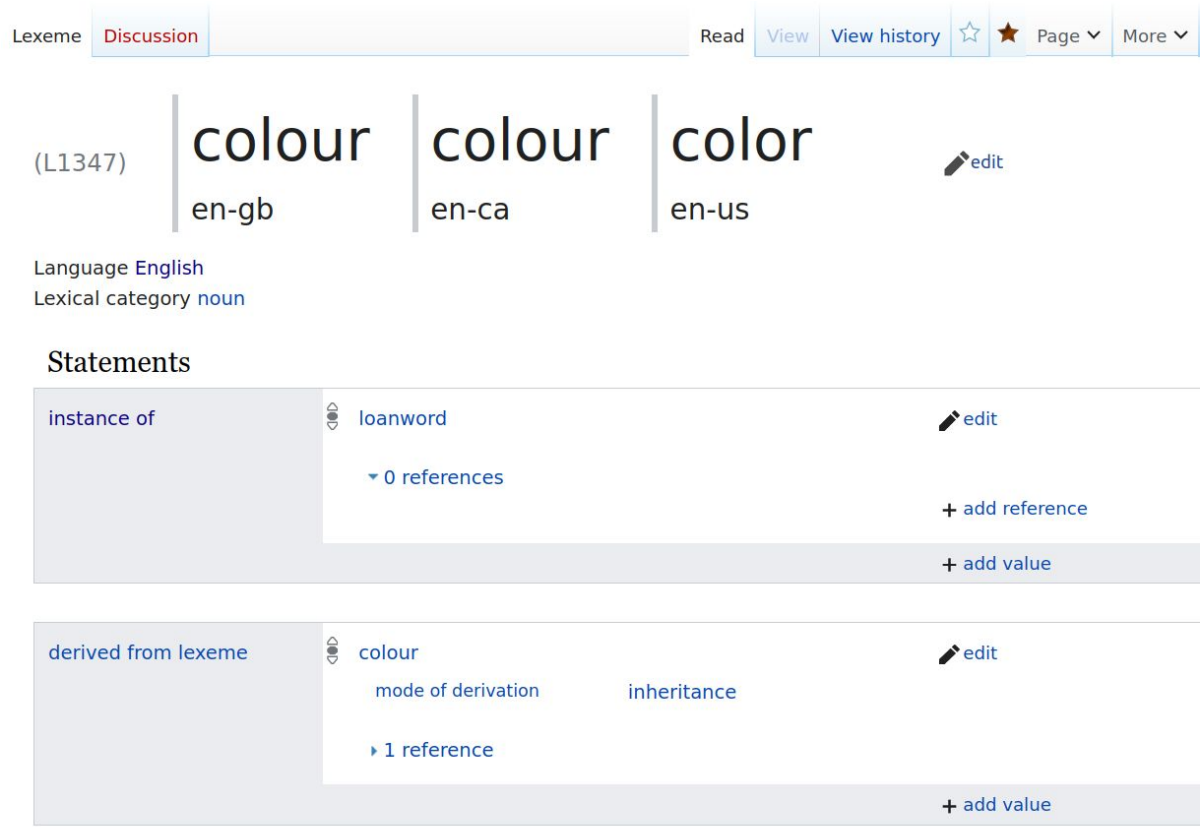
- Reference at least 1 form. Preferably on a usage example (P5831) statement
- or on another statement, eg.
  - described by source (P1343)
  - attested as (P7855)
  - attested in (P5323) are possible other properties

Frequently asked questions

Wikidata Lexicodays
28-30 June 2024

Can I enter dialects of a Lexeme?

Yes, simply edit to add a new lemma.



Can I enter different scripts of a Lexeme?

Yes!

You can add spellings of the same word in multiple scripts.

What if the dialect or script's language code isn't supported?

Use `mis-x-Q…` as language of the lemma

Does QuickStatements work for Lexemes?

- Not really :(
- Possible: add new statements on the Lexeme level for existing Lexemes
- Not possible: create new Lexemes, Forms or Senses

Is it possible to query Lexemes?

- Yes! The Query Service supports Lexemes
- All words in Indonesian https://w.wiki/AHRR
- Longest words in English https://w.wiki/45o
- More queries & ideas: Wikidata:Lexicographical_data/Ideas_of_queries
- Get help: Wikidata:Request_a_query

Resources and Tools

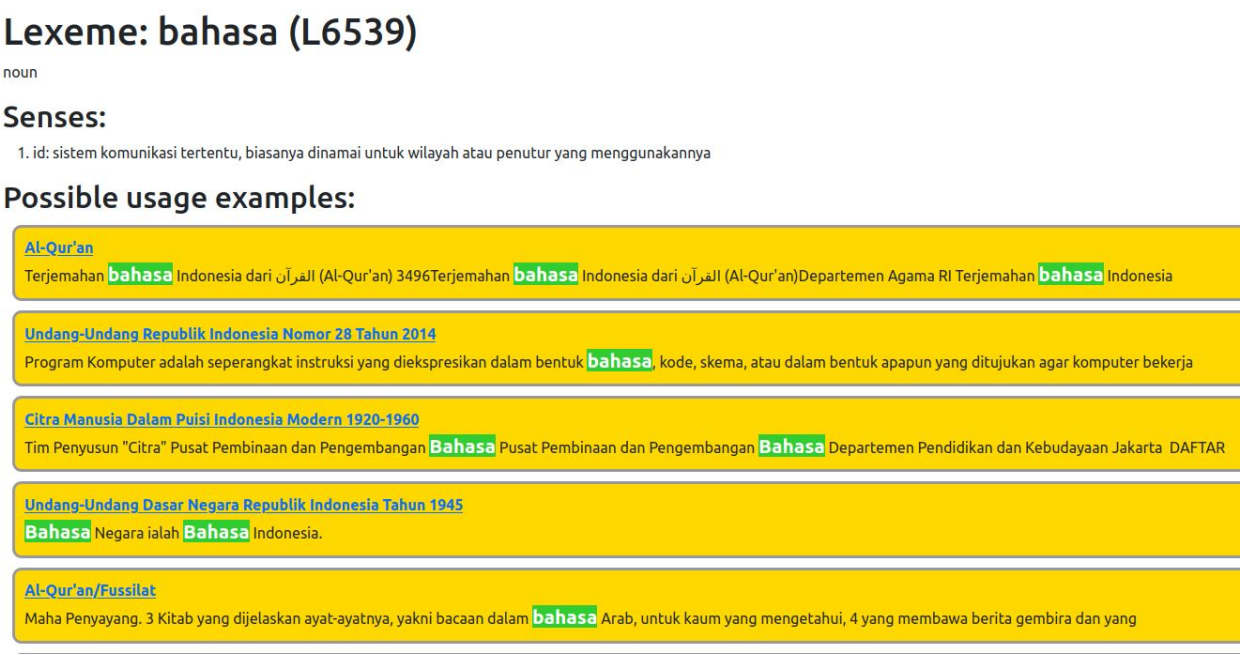
Wikidata Lexicodays
28-30 June 2024

Tools to edit Lexemes: Wikidata Lexeme Forms

- Creates Lexemes and generate a lot of Forms https://tools.wmflabs.org/lexeme-forms/
- Possibility to create new templates (eg. conjugations, declinations)

Tools to edit Lexemes: Luthor

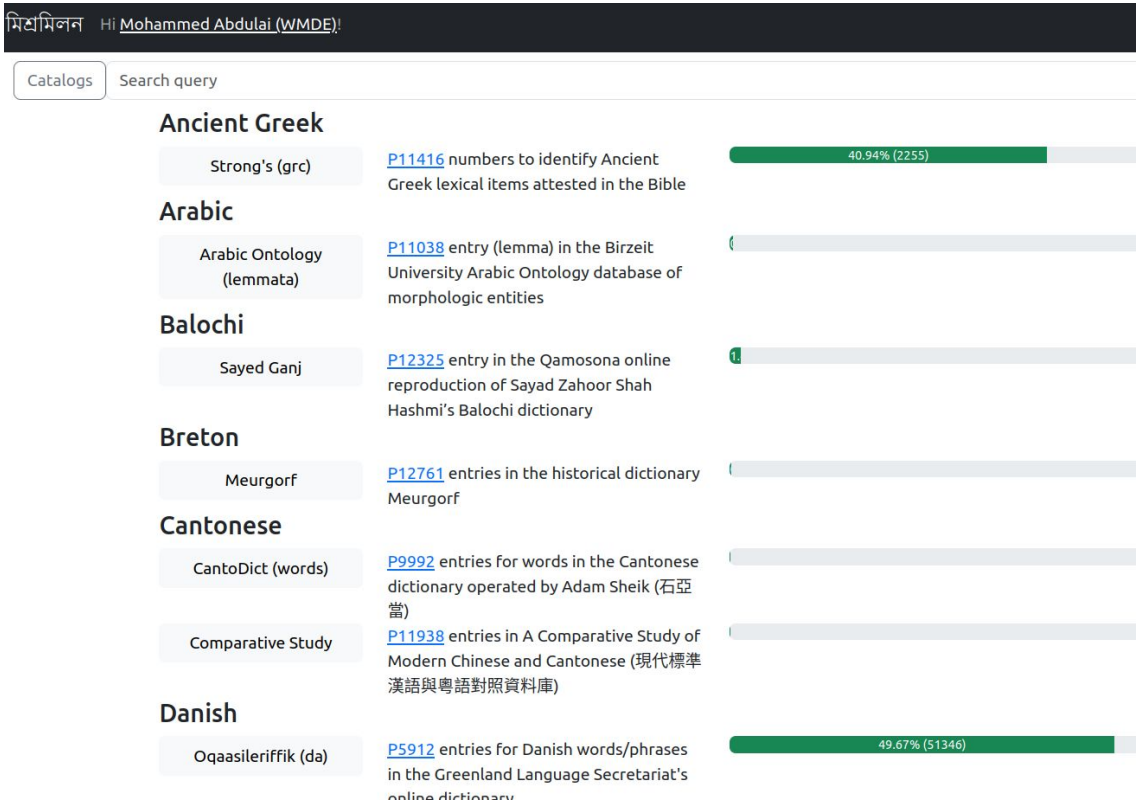
Finds examples from Wikisource as potential `usage examples` to Lexemes



https://luthor.toolforge.org/

Tools to edit Lexemes: Mishramilan

- Lists entries of some external databases
- Allows users to match those entries against Wikidata lexemes.

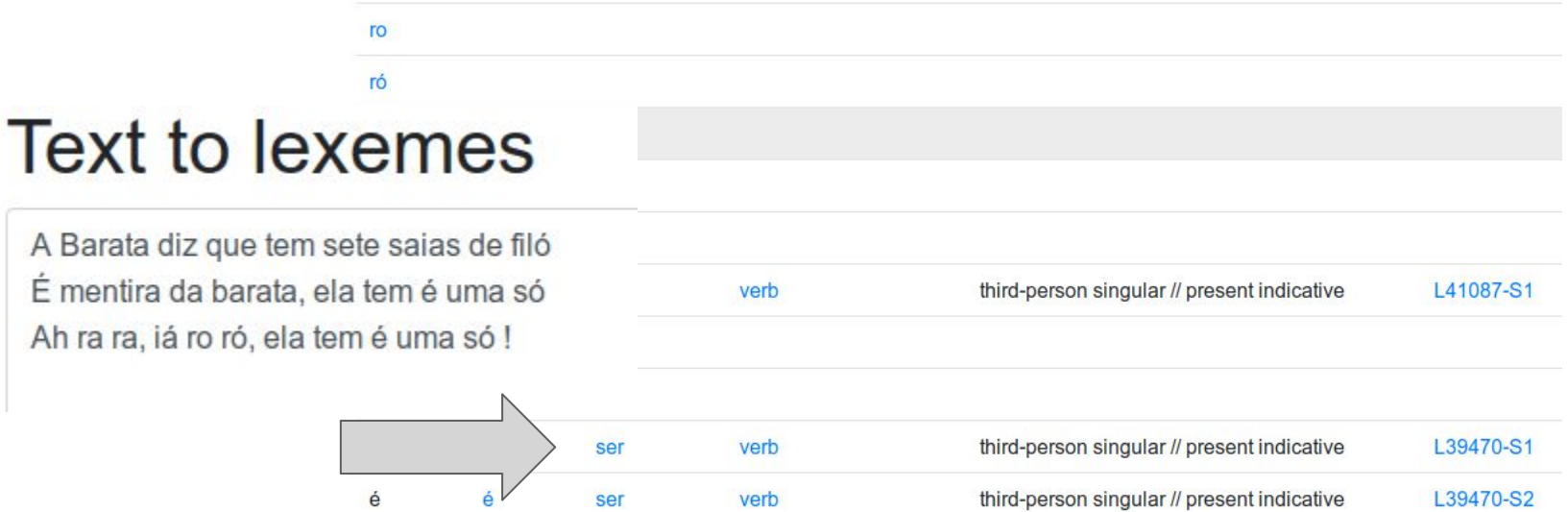


https://mishramilan.toolforge.org/#/

Wikidata Lexicodays
28-30 June 2024

Tools to edit Lexemes: text-to-lexemes

- Ordia text-to-Lexeme transforms a text into Lexemes and helps with creation https://tools.wmflabs.org/ordia/text-to-lexemes



Tools to edit Lexemes: Lingua Libre

- Record words in your language
  https://lingualibre.fr/wiki/LinguaLibre:Main_Page
- Generate lists of words from Lexemes
  https://lingualibre.fr/wiki/Help:Create_your_own_lists
- Automatically adds the pronunciation as a statement

Tools to reuse Lexemes

- The **Query Service**
  - Lexemes don’t appear in the documentation… yet ^^
- The **API**

How to help

There are many ways to improve lexicographical data on Wikidata!

d:Wikidata:Lexicographical_data/How_to_help

Thanks for your attention!

Get in touch with us:

Lydia Pintscher
lydia.pintscher@wikimedia.de
@nightrose

Mohammed Abdulai
mohammed.abdulai@wikimedia.de
@masssly

Wikidata Lexicodays
28-30 June 2024