



# 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.

# 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

Glossary:

[Wikidata:Lexicographical\\_data/Glossary](https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Glossary)

Data model:

[MediaWiki:Extension:WikibaseLexeme/Data Model](https://www.wikidata.org/wiki/MediaWiki:Extension:WikibaseLexeme/Data_Model)

## L-id Lexeme

*a meaning carrying part of a language, such as a word or a phrase*

one

### Lemma

*standard form or dictionary form of the lexeme  
for verbs this is usually the infinitive form, for a noun the nominative singular, etc.*

one

### Lexical category

*also known as the part of speech or word class  
defines the lexeme to be either a noun, or a verb, or an adjective, etc.  
the set of possible values is open and taken from the Wikidata items*

one

### Language

*chosen in an open list from Wikidata items*

many

### Statements

*e.g. derived-from, region, period, usage example, homonym, etc.*

### Forms

*specific, conjugated or inflexed forms of the lexeme*

one

### Representation

*the actual string value realizing a given form*

many

many

### Grammatical features

*e.g. normative, comparative, past tense, etc.*

many

### Statements

*e.g. region, period, pronunciation, etc.*

### Senses

one

### Gloss

*short description, translatable in all languages of Wikidata*

many

many

### Statements

*e.g. translations, synonyms, connotation, register, refers-to-concept, etc.*



# Creating Your First Lexeme



# Search for Lexemes

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

## Search results

To search for Wikidata items by their title on a given site, use [Special:ItemByTitle](#).

Q L:Mouse

Advanced search: [Sort by relevance](#)

Search in: [\(Main\) X](#) [Property X](#)

[mouse \(L1119\)](#)

English, noun...

14 statements, 4 forms - 13:23, 3 June 2024

[mouse \(L332274\)](#)

English, verb...

4 statements, 5 forms - 17:49, 2 May 2024

[moucha \(L19721\)](#)

Czech, noun...

1 statement, 14 forms - 20:06, 22 December 2019



Special page

Search Wikidata

# Create a new Lexeme

You can check whether a Lexeme already exists by using [the search](#). You can also learn more about Lexemes in the help box below.

Lemma \*

Base form of a word, e.g. 'cat'

Lexeme's language \*

The Lexeme's language, e.g. 'English'

Lexical category \*

The Lexeme's lexical category, e.g. '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

## About Lexemes

Lexemes contain [lexicographical data](#) which is data about words or phrases, such as language, etymology, inflections, etc. Here is an example:

cat (L7)	cat	Language: <a href="#">English</a>
	en	Lexical category: <a href="#">noun</a>

Lexemes don't contain general data (date of birth, opening date, author, country, coordinates, website, etc.) about the entity or concept to which they refer. If you want to submit general data, you need to [create an Item](#) instead.

- Main page
- Community portal
- Project chat
- Create a new Item
- Recent changes
- Random Item
- Query Service
- Nearby
- Help
- Donate

Lexicographical data

- Create a new Lexeme
- Recent changes
- Random Lexeme

Tools

- What links here
- Special pages
- Printable version
- Page information
- Get shortened URL
- Download QR code



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

# Create a new Lexeme

You can check whether a Lexeme already exists by using [the search](#). You can also learn more about Lexemes in the help box below.

Lemma \*

mouse

Lexeme's language \*

English

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 

## About Lexemes

Lexemes contain [lexicographical data](#) which is data about words or phrases, such as language, etymology, inflections, etc. Here is an example:

cat (L7) | cat Language: English  
en Lexical category: noun

Lexemes don't contain general data (date of birth, opening date, author, country, coordinates, website, etc.) about the entity or concept to which they refer. If you want to submit general data, you need to [create an Item](#) instead.

Hit the "Create Lexeme" button



Lexeme

[Discussion](#)

Read

[View history](#)



(L4223)

# mouse

en

 [edit](#)

Language [English](#)

Lexical category [noun](#)

## Statements

[+ add statement](#)

## Senses

[+ add Sense](#)

## Forms

[+ add Form](#)

# Take note:

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 the

Lemma \*

Lexeme's language \*

Spelling variant of the Lemma \* [\(Help\)](#)

Lexical category \*

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

Not all  
languages are  
supported for  
Lexemes yet!



## Create a new Lexeme

You can check whether a Lexeme already exists by using [the search](#). You can also learn more about Lexemes in the

Lemma \*

Lexeme's language \*

Spelling variant of the Lemma \* [\(Help\)](#)

Lexical category \*

*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](#)



# 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 the

Lemma \*

Lexeme's language \*

Spelling variant of the Lemma \* [\(Help\)](#)

Lexical category \*

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



Lexeme

[Discussion](#)

Read

[View history](#)



(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)

(L1119)

mouse

en

 edit

Language English

Lexical category noun

## Statements



instance of

 count noun edit

▼ 0 references

+ add reference

+ add value

described by source


 Merriam-Webster online dictionary edit

▼ 0 references

+ add reference

+ add value

usage example

 When the mouse laughs at the cat, there's a hole nearby.  
(English) edit

▼ 0 references



## Senses

L1119-S1

English  
Spanish

any small rodent of the genus Mus  
mamífero roedor de pequeño tamaño, del género Mus

edit

Persian  
French

بستانداری کوچک از راستهٔ جوندگان  
petit rongeur du genre Mus

Hebrew

מכרסם קטן מהסוג Mus

Kannada

ಇಲಿಯು ದಂಚಕಗಳ ಗಣಕ್ಕೆ ಸೇರಿದೆ

Portuguese

qualquer pequeno roedor do gênero Mus

### Statements about L1119-S1

image



edit

Apodemus sylvaticus bosmuis.jpg

2,160 × 1,524; 2.53 MB

▼ 0 references

+ add reference

+ add value

item for this sense



mouse

edit

▼ 0 references

+ add reference

## Senses

L1119-S1

English	any small rodent of the genus Mus	<a href="#">edit</a>
Spanish	mamífero roedor de pequeño tamaño, del género Mus	
Persian	بستانداری کوچک از راستهٔ جوندگان	
French	petit rongeur du genre Mus	
Hebrew	מכרסם קטן מהסוג Mus	
Kannada	ಇಲಿಯು ದಂಚಕಗಳ ಗಣಕ್ಕೆ ಸೇರಿದೆ	
Portuguese	qualquer pequeno roedor do gênero Mus	

## Statements about L1119-S1

image



[edit](#)

Apodemus sylvaticus bosmuis.jpg

2,160 × 1,524; 2.53 MB

▼ 0 references

[+ add reference](#)

[+ add value](#)

item for this sense




mouse

[edit](#)

▼ 0 references

[+ add reference](#)

 L1119-F2 | mice  
en

Grammatical features plural

Statements about L1119-F2

 edit

+ add statement

L1119-F3 | mouse's  
en

Grammatical features genitive case

Statements about L1119-F3

 edit

+ add statement

L1119-F4 | mouses  
en

Grammatical features plural

 edit





## 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

# 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

# Can I enter dialects of a Lexeme?

Lexeme [Discussion](#) [Read](#) [View](#) [View history](#) [☆](#) [★](#) [Page](#) [More](#)

(L1347) **colour** | **colour** | **color** [edit](#)  
en-gb | en-ca | en-us

Language [English](#)  
Lexical category [noun](#)

## Statements

instance of [loanword](#) [edit](#)  
▼ 0 references  
[+ add reference](#)  
[+ add value](#)

derived from lexeme [colour](#) [edit](#)  
mode of derivation inheritance  
▶ 1 reference  
[+ add value](#)

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.

Lexeme [Discussion](#) [Read](#) [View](#) [View history](#) [★](#) [★](#)

(L8303) **ruwa** **رُوَا** [edit](#)  
ha ha-arab

Language [Hausa](#)  
Lexical category [noun](#)

### Statements

<a href="#">grammatical gender</a>	masculine <a href="#">edit</a> ▼ 0 references <a href="#">+ add reference</a> <a href="#">+ add value</a>
<a href="#">usage example</a>	Ruwa (ana amfanidashi domin inganta rayuwan dan adam (Hausa) <a href="#">edit</a> ▼ 0 references <a href="#">+ add reference</a>

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

(L940072)

<b>مِس</b>	<b>mis</b>
bal	bal-x-Q123750890

 edit

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

**МІС**

bal-x-Q123734863

Language [Balochi](#)  
Lexical category [noun](#)

### Statements

described by source	 <a href="#">Baluchi Glossary</a> 
	page(s) 111
	<a href="#">▶ 1 reference</a>
<a href="#">+ add value</a>	

# 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](Wikidata:Lexicographical_data/Ideas_of_queries)
- Get help: [Wikidata:Request\\_a\\_query](Wikidata:Request_a_query)





# Resources and Tools

# 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)

## English verb

### present

They  every day.

### third-person singular

He  every day.

### simple past

He  every day last week.

### present participle

They are  right now.

### past participle

We have  for hours.

Create

Advanced

# Tools to edit Lexemes: Luthor

Finds examples  
from Wikisource as  
potential usage  
examples to  
Lexemes

## Lexeme: bahasa (L6539)

noun

### Senses:

- id: sistem komunikasi tertentu, biasanya dinamai untuk wilayah atau penutur yang menggunakannya

### Possible usage examples:

#### [Al-Qur'an](#)

Terjemahan **bahasa** Indonesia dari القرآن (Al-Qur'an) 3496Terjemahan **bahasa** Indonesia dari القرآن (Al-Qur'an)Departemen Agama RI Terjemahan **bahasa** Indonesia

#### [Undang-Undang Republik Indonesia Nomor 28 Tahun 2014](#)

Program Komputer adalah seperangkat instruksi yang diekspresikan dalam bentuk **bahasa**, kode, skema, atau dalam bentuk apapun yang ditujukan agar komputer bekerja

#### [Citra Manusia Dalam Puisi Indonesia Modern 1920-1960](#)

Tim Penyusun "Citra" Pusat Pembinaan dan Pengembangan **Bahasa** Pusat Pembinaan dan Pengembangan **Bahasa** Departemen Pendidikan dan Kebudayaan Jakarta DAFTAR

#### [Undang-Undang Dasar Negara Republik Indonesia Tahun 1945](#)

**Bahasa** Negara ialah **Bahasa** Indonesia.

#### [Al-Qur'an/Fussilat](#)

Maha Penyayang. 3 Kitab yang dijelaskan ayat-ayatnya, yakni bacaan dalam **bahasa** Arab, untuk kaum yang mengetahui, 4 yang membawa berita gembira dan yang

<https://luthor.toolforge.org/>

# Tools to edit Lexemes: Mishramilan

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

মিশ্রমিলন Hi Mohammed Abdulaj (WMDE)

Catalogs Search query

<b>Ancient Greek</b>		
Strong's (grc)	<a href="#">P11416</a> numbers to identify Ancient Greek lexical items attested in the Bible	40.94% (2255)
<b>Arabic</b>		
Arabic Ontology (lemmata)	<a href="#">P11038</a> entry (lemma) in the Birzeit University Arabic Ontology database of morphologic entities	
<b>Balochi</b>		
Sayed Ganj	<a href="#">P12325</a> entry in the Qamosona online reproduction of Sayad Zahoor Shah Hashmi's Balochi dictionary	
<b>Breton</b>		
Meurgorf	<a href="#">P12761</a> entries in the historical dictionary Meurgorf	
<b>Cantonese</b>		
CantoDict (words)	<a href="#">P9992</a> entries for words in the Cantonese dictionary operated by Adam Sheik (石亞當)	
Comparative Study	<a href="#">P11938</a> entries in A Comparative Study of Modern Chinese and Cantonese (現代標準漢語與粵語對照資料庫)	
<b>Danish</b>		
Oqaasileriffik (da)	<a href="#">P5912</a> entries for Danish words/phrases in the Greenland Language Secretariat's online dictionary	49.67% (51346)

# 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>

ro

ró

## Text to lexemes

A Barata diz que tem sete saias de filó  
É mentira da barata, ela tem é uma só  
Ah ra ra, iá ro ró, ela tem é uma só !

verb

third-person singular // present indicative

[L41087-S1](#)



é

é

ser

verb

third-person singular // present indicative

[L39470-S1](#)

ser

verb

third-person singular // present indicative

[L39470-S2](#)

# Tools to edit Lexemes: Lingua Libre

- Record words in your language  
[https://lingualibre.fr/wiki/LinguaLibre:Main\\_Page](https://lingualibre.fr/wiki/LinguaLibre:Main_Page)
- Generate lists of words from Lexemes  
[https://lingualibre.fr/wiki/Help:Create\\_your\\_own\\_lists](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**

```
wbladdform  
wbladdsense  
wbleditformelements  
wbleditsenseelements  
wblinktitles  
wblmergelexemes  
wblremoveform  
wblremovesense
```

# How to help

There are many ways to improve lexicographical data on Wikidata!

- 1 First steps, regardless of language
  - 1.1 Check what lexemes already exist in your language
  - 1.2 Check any language-specific documentation for your language
- 2 Tasks for all languages
  - 2.1 Add pronunciation audio to existing lexemes
  - 2.2 Find more specific issues to resolve with Wikidata queries
  - 2.3 Make it easier to add specific types of lexemes
  - 2.4 Add usage examples to existing lexemes
  - 2.5 See if text can be generated using the lexemes in your language
- 3 Tasks for higher-resourced languages
  - 3.1 Add external identifiers to existing lexemes
  - 3.2 Add senses to lexemes that currently don't have them
- 4 Tasks for lower-resourced languages
  - 4.1 Add lexemes for concepts commonly encountered in linguistics
  - 4.2 Add lexemes for concepts from the weekly Lexemes Challenge

[d:Wikidata:Lexicographical\\_data/How\\_to\\_help](https://www.wikidata.org/wiki/d:Wikidata:Lexicographical_data/How_to_help)





# Thanks for your attention!

Get in touch with us:

Lydia Pintscher  
[lydia.pintscher@wikimedia.de](mailto:lydia.pintscher@wikimedia.de)  
@nightrose

Mohammed Abdulai  
[mohammed.abdulai@wikimedia.de](mailto:mohammed.abdulai@wikimedia.de)  
@masssly

## Credits

This presentation is licensed under CC BY-SA 4.0.