Lexicographical Data: words and phrases on Wikidata

Léa Lacroix
Wikidata Labs, May 2019

Table of contents

Wikidata: the basics

- A knowledge base
- Part of the Wikimedia projects
- Structured data
- Linked to other databases
- Multilingual
- Collaborative
- Released under public domain (CC0)
- Based on facts and references
- Made for humans and machines

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

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Concept URI
Cite this page
Automatic addition

Douglas Adams (Q42)

English writer and humorist

Douglas Noël Adams | Douglas Noel Adams | Douglas N. Adams | DNA

edit

▾ In more languages

Language	Label	Description	Also known as
English	Douglas Adams	English writer and humorist	Douglas Noël Adams Douglas Noel Adams Douglas N. Adams DNA
German	Douglas Adams	britischer Schriftsteller	Douglas Noël Adams Douglas Noel Adams
French	Douglas Adams	écrivain anglais de science-fiction	Douglas Noel Adams Douglas Noël Adams
Italian	Douglas Adams	scrittore inglese	Douglas Noel Adams Douglas N. Adams

All entered languages

Statements

instance of	human ▸ 2 references	edit
	+ add value	
image	Douglas adams portrait cropped.jpg media legend: Portræt af Douglas Adams (Danish) Douglas Adams portrait. (English) Portrait de Douglas Adams. (French)	edit

Wikipedia (71 entries) edit

ar دوغلاس آدمز
arz دوجلاس ادامز
ast Douglas Adams
azb داقلاس آدامز
az Duqlas Adams
bar Douglas Adams
be_x_old Дуглас Адамз
be Дуглас Адамс
bg Дъглас Адамс
bn ডগলাস অ্যাডামস
bs Douglas Adams
ca Douglas Adams
cs Douglas Adams
cy Douglas Adams
da Douglas Adams
de Douglas Adams
el Ντάγκλας Άνταμς
en Douglas Adams
eo Douglas Adams
es Douglas Adams
et Douglas Adams
eu Douglas Adams
fa داگلاس آدامز
fi Douglas Adams
fr Douglas Adams
ga Douglas Adams
gl Douglas Adams
he דאגלס אדמס
hr Douglas Adams

Until 2018,

Wikidata was only describing

concepts

Now it also includes

words!

Wait, what's the difference?

Concept "mouse"

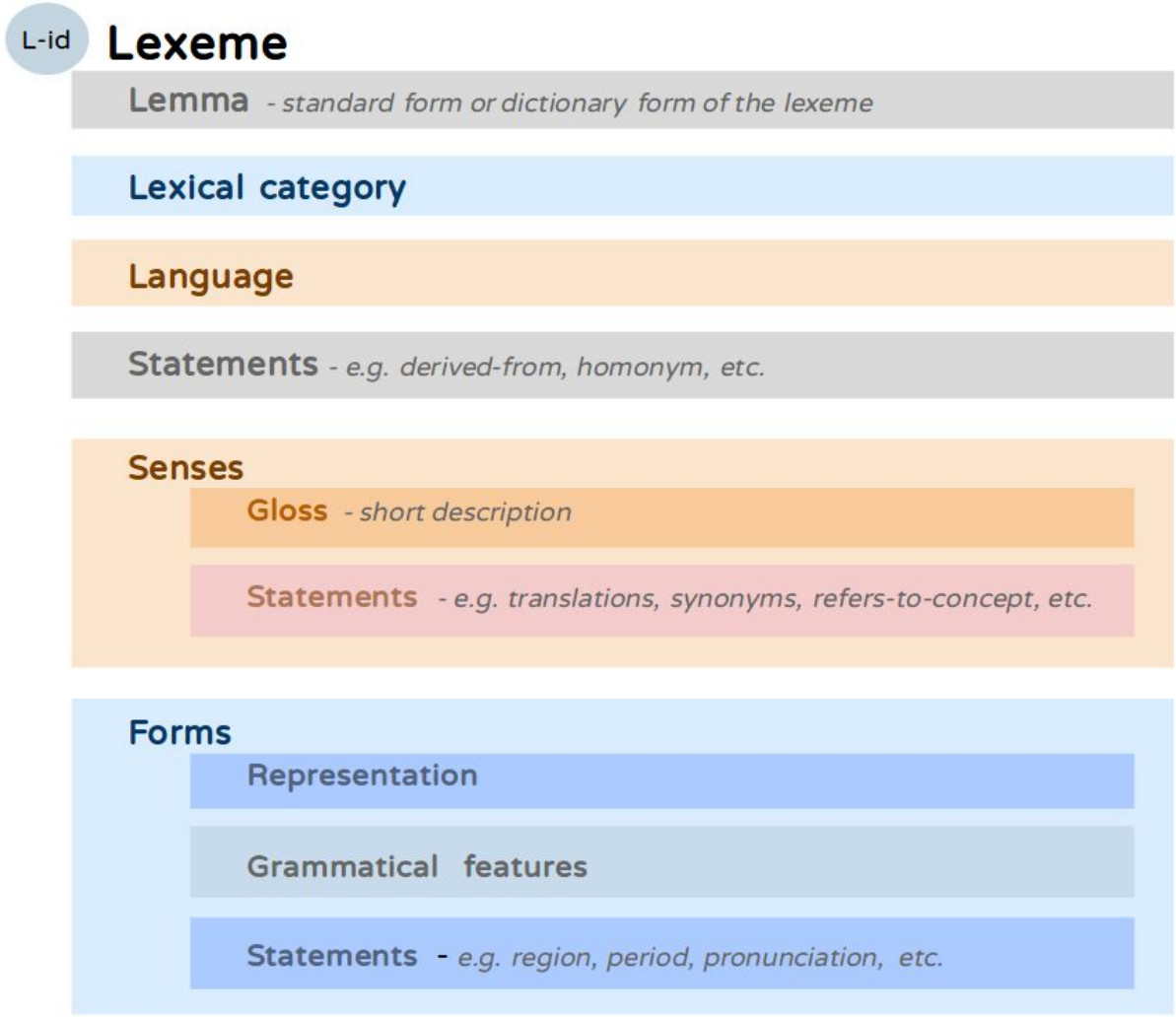
- Specie of mammal
- Taxon name
- Average size
- Picture
- Encyclopedia of Life ID

File:House mouse.jpg, public domain

Lexeme "mouse"

- Language: English
- Lexical category: noun
- Plural form: mice (irregular)
- Etymology: Proto-Germanic *mūs
- Senses: animal, computer device
- Translations: souris (fr), rato (pt)
- Audio pronunciation ▶

/maʊs/



More info:
[mw:Extension:Wikibase Lexeme/Data Model]

(L407)

casa

pt

edit

Language Portuguese
Lexical category noun

grammatical gender	feminine 0 references + add reference	edit
	+ add value	
derived from	casa 0 references + add reference	edit
	+ add value	

Senses

L407-S1	English	building, habitation	edit
	Portuguese	prédio	

Statements about L407-S1

translation	house (structure built or serving as an abode of human beings) 0 references	edit + add reference
		+ add value

Forms

L407-F1	casas pt	edit

Grammatical features plural

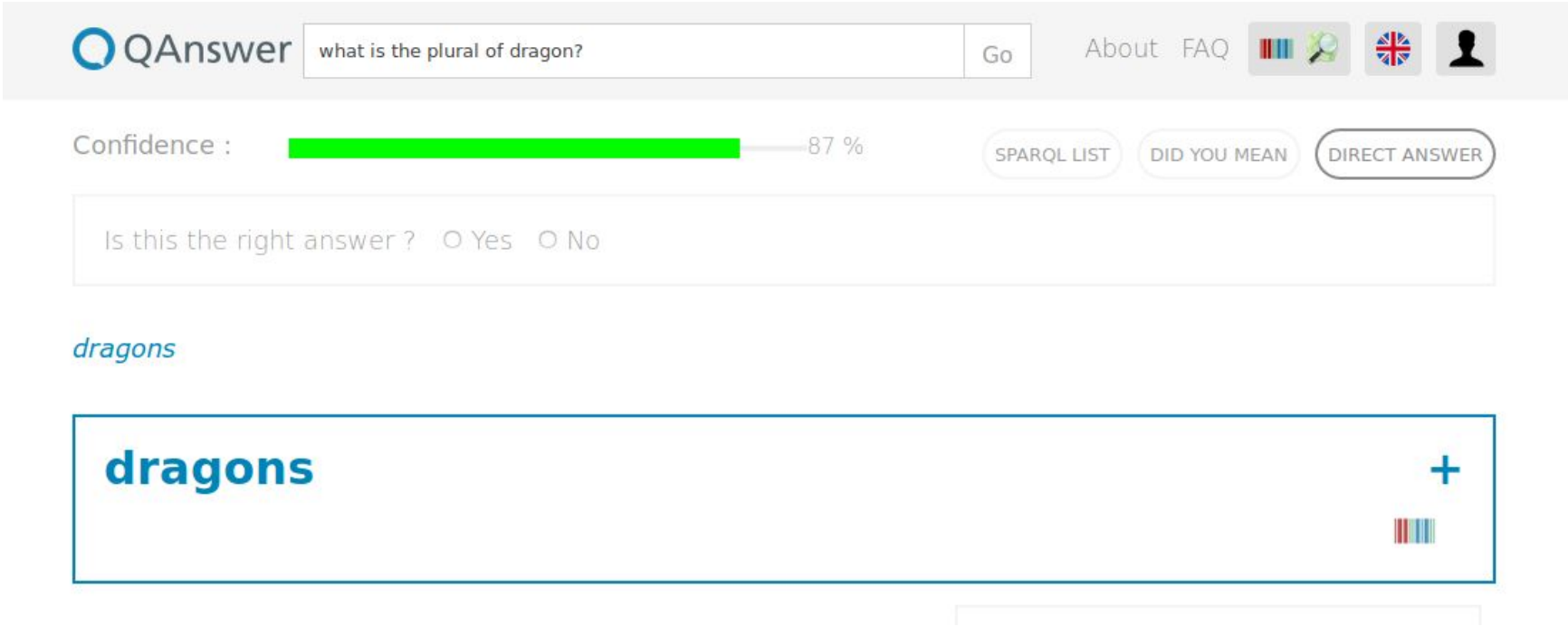
Statements about L407-F1

IPA transcription	/'ka.zas/		edit
	language of work or name	Brazilian Portuguese	
	0 references		+ add reference

Why is it interesting?

- Structured data = machine readable
- Plain text = only for humans, need to build complex algorithms
- Can be reused by tools, research, dictionaries, translation services
- CC0 = open knowledge, can be reused by all
- Huge variety of languages, including undeserved ones
- International community = more people to help

Examples of reuse: question answering algorithms



QAnswer https://qanswer-frontend.univ-st-etienne.fr/

Examples of reuse: language-learning games

Der Die Das http://auregann.fr/derdiedas/

Other uses in the future

- Support for Wiktionary, Wikisource & other Wikimedia projects
- Research
- Text analysis, text mining
- Translation tools
- New generation of dictionaries
- Language-learning tools

A few numbers

- **45400** Lexemes in **340** languages and dialects
- **42** Lexemes in Portuguese
- **12000** Senses (definitions)
- More statistics: https://tools.wmflabs.org/ordia/

Frequently asked questions

What's the difference with Wiktionary?

- Wiktionary = plain text + templates, Wikidata = structured data
- Wikidata can be easily parsed and reused
- Wikidata works with Lexemes, Wiktionary combines Lexemes
- Wiktionary may have extra info (examples, quotes…)
- Wikidata aims to support Wiktionaries (if they want to)

Frequently asked questions

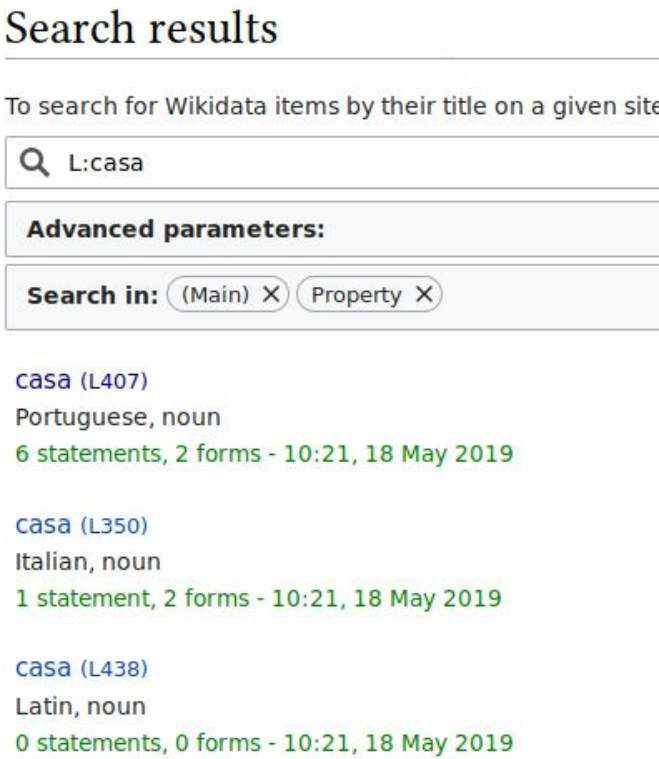
How to enter dialects or scripts?

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

Frequently asked questions

How to search for Lexemes?

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



Frequently asked questions

Is it possible to query Lexemes?

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

Frequently asked questions

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

Any questions so far?

Editing Lexemes

- Check if the Lexeme exists (L:singen)
- Find an example https://www.wikidata.org/wiki/Lexeme:L2058
- Create a new Lexeme https://www.wikidata.org/wiki/Special:NewLexeme
- Add statements (auxiliary verb)
- Add a Sense
- Add translation (needs another Sense)
- Add a Form (present, first-person singular)

A few tips for editing

- 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 doing wrong :)
- In doubt: ask the other editors for help (LexData talk page, Telegram)

Everything is new!

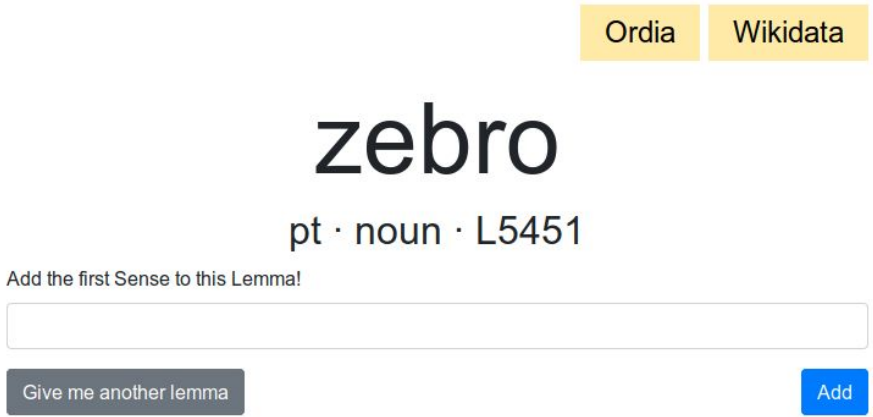
- Started from scratch in May 2018
- Everything is to create
- Still plenty of questions, ways to models
- Maybe new properties to be created
- Not everything works perfectly (eg mobile interface)
- Some documentation needs improvement

Tools to edit Lexemes

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

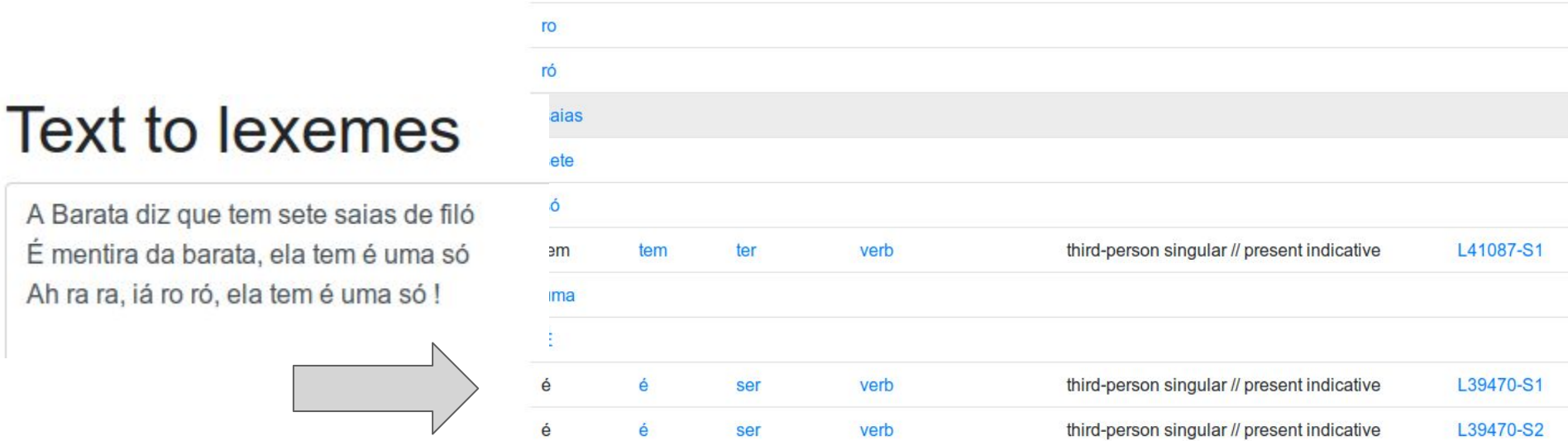
Tools to edit Lexemes

- **Wikidata Senses**, to add Senses to existing Lexemes https://tools.wmflabs.org/lexeme-senses/
- Add Senses for Portuguese words in Portuguese: https://tools.wmflabs.org/lexeme-senses/add/pt



Tools to edit 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

Helpful pages

- Create and edit Lexemes Special:NewLexeme
- Suggest or discuss new properties Wikidata:Property_proposal/Lexemes
- Discuss with the community about how to model words &
- Report any bug or wish for the future Wikidata_talk:Lexicographical_data
- Try the existing tools Wikidata:Tools/Lexicographical_data
- Suggest ideas of tools Wikidata:Lexicographical_data/Ideas_of_tools
- Ideas & examples of queries https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Ideas_of_queries

More questions?

Activities

- Add Senses in Portuguese
  https://tools.wmflabs.org/lexeme-senses/add/pt
- Create more Lexemes in Portuguese
  - Using text-to-Lexemes
    https://tools.wmflabs.org/ordia/text-to-lexemes

Activities

- Create Portuguese templates in Wikidata Lexeme Forms
  - Create the page for Portuguese https://www.wikidata.org/wiki/Wikidata:Wikidata_Lexeme_Forms#Language_support
  - Help translating the interface
  - Add new templates (example: https://www.wikidata.org/wiki/Wikidata:Wikidata_Lexeme_Forms/French)
  - Ping Lucas (user:Lucas Werkmeister) to include the templates in the tool (or for any question)

Activities

- Once templates are ready: create Forms in Portuguese
- Developers: create a tool or a game using lexicographical data (see also: ideas of tools)

Thank you!

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