Enhancing a spellcheck dictionary by Wikidata lexemes

LibreOffice Conference 2021

Stanislav Horáček

Czech LibreOffice team
The Document Foundation

24th September 2021

Introduction

Czech Hunspell dictionary: GNU GPL, ~2006

- FOSS applications, no update, missing/incorrect words

GNU GPL update – 2021

Czech morphological dictionary – CC0, 2019

Wikidata lexemes – CC0, >2018

current Czech Hunspell dictionaries

- updated GNU GPL – LibreOffice
- experimental CC0

GNU GPL update

done by Miroslav Pošta

thousands of words

- corrections of forms/patterns
- addition e.g. zoological terms
- removal of rarely used forms

pragmatic approach wrt. binary spellcheck

in LibreOffice 7.2 – “upstream” of Hunspell dictionaries

Czech morphological dictionary

Masaryk university & RedHat

2019, public domain licence (= CC0)

analysis of language corpus

nouns, adjectives, verbs; missing parts

~60 000 lexemes, ~650 000 forms

github.com/plin/slovnik

Lexicographical data at Wikidata

namespace with database of lexemes

lexeme – forms, characteristics, relations…

~13 000 lexemes (~700 000 forms) in Czech language

from 2018, CC0 licence

user interface to create and maintain lexemes

www.wikidata.org/wiki/Wikidata:Lexicographical_data

Lexicographical data at Wikidata

Wikimedia Foundation project

data usage

- spellcheck dictionary
- hyphenation dictionary
- dictionary of synonyms, antonyms, …
- adding diacritics
- foreign language dictionaries, translators
- grammar tools
- …

Lemma, senses, forms

(L45253) duchaplně
cs

edit

Language Czech
Lexical category adverb

Statements

+ add statement

Senses

L45253-S1 | Czech | duchaplným způsobem

edit

Statements about L45253-S1

+ add statement

+ add Sense

Forms

L45253-F1 | duchaplně
cs

edit

Grammatical features positive

Statements about L45253-F1

+ add statement

L45253-F2 | duchaplněji
cs

edit

Grammatical features comparative

Statements about L45253-F2

Characteristics, etymology

(L10536) **prasátko**

cs

edit

Language Czech
Lexical category noun

Statements

grammatical gender	neuter 0 references + add reference	edit
	+ add value	
derived from	prase 0 references + add reference	edit
	+ add value	

+ add statement

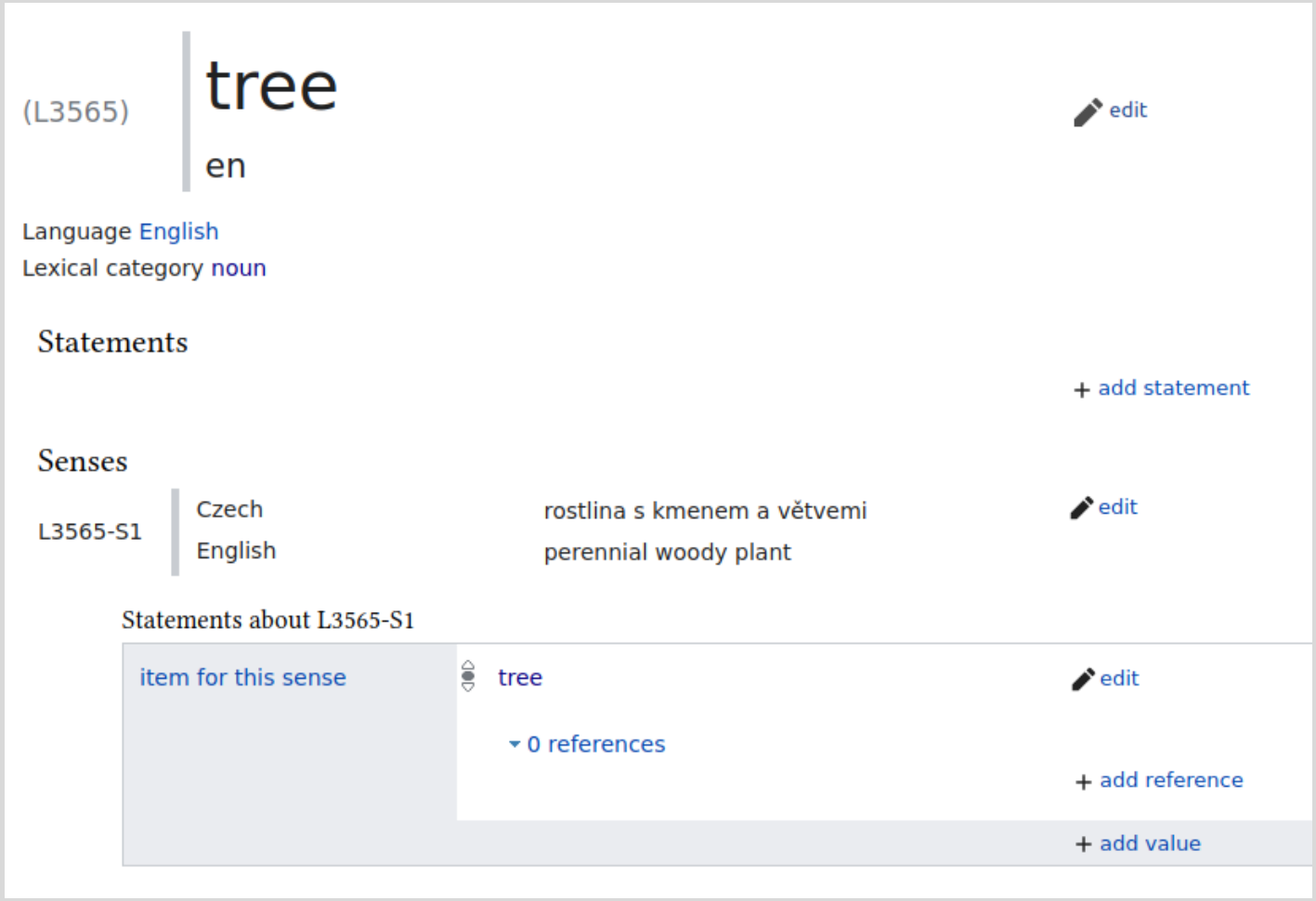
Senses

L10536-S1 | Czech | malé prase | edit

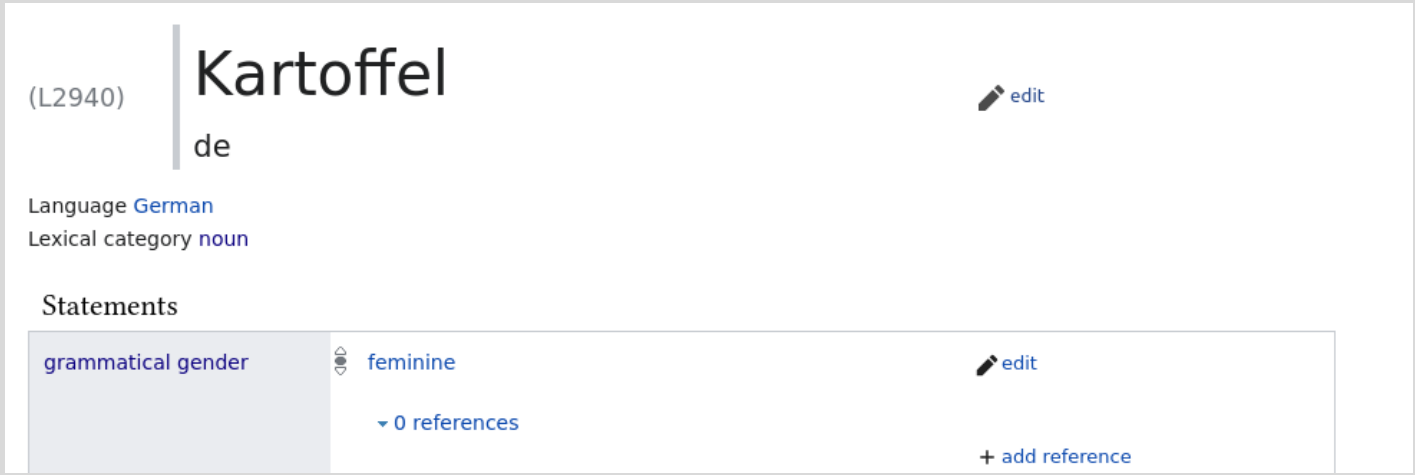
Statements about L10536-S1

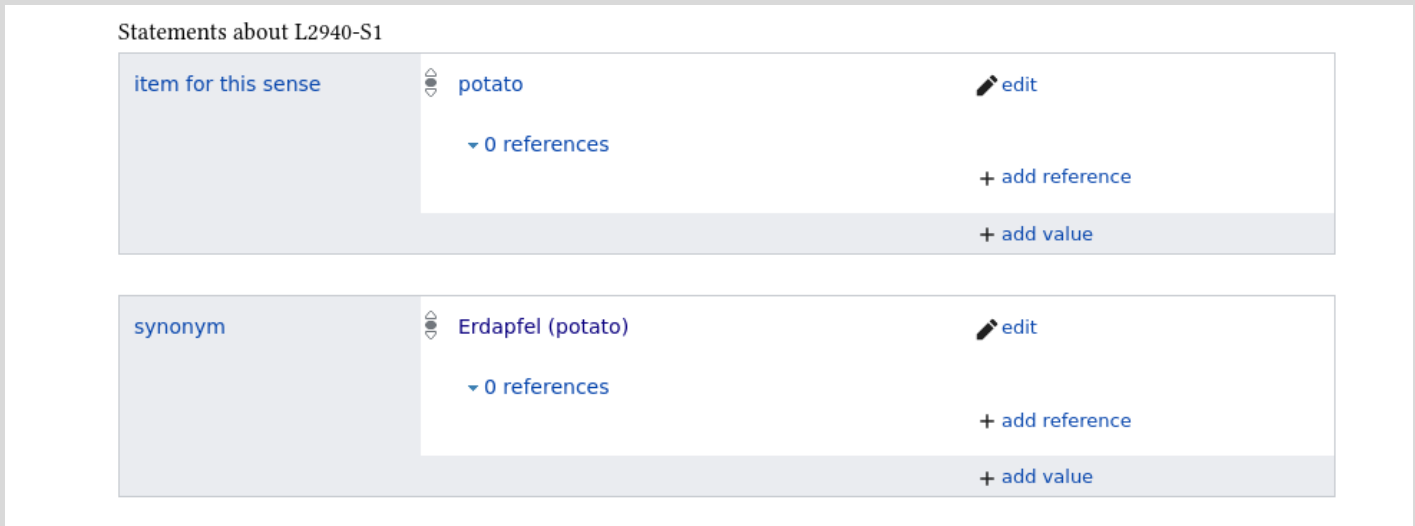
language style	diminutive 0 references + add reference	edit

Corresponding Wikidata items



Relation between lexemes





Pronunciation, …

(L442) **být**

cs

edit

L442-F14 | být
cs

edit

Grammatical features infinitive

Statements about L442-F14

pronunciation audio

0:00

Cs-být.ogg
1.1 s; 14 KB

language of work or name: Czech

0 references

+ add reference

edit

+ add value

IPA transcription

bi:t

language of work or name: Czech

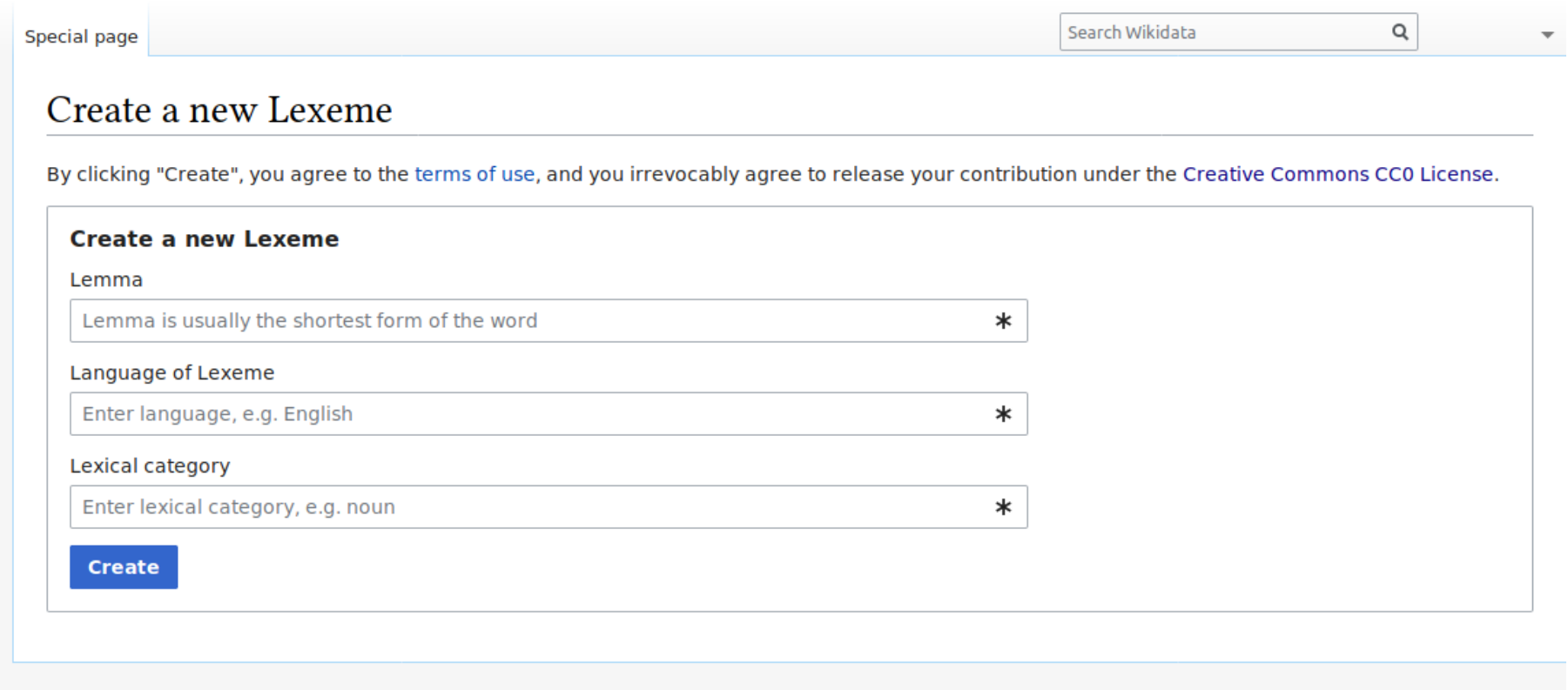
0 references

+ add reference

edit

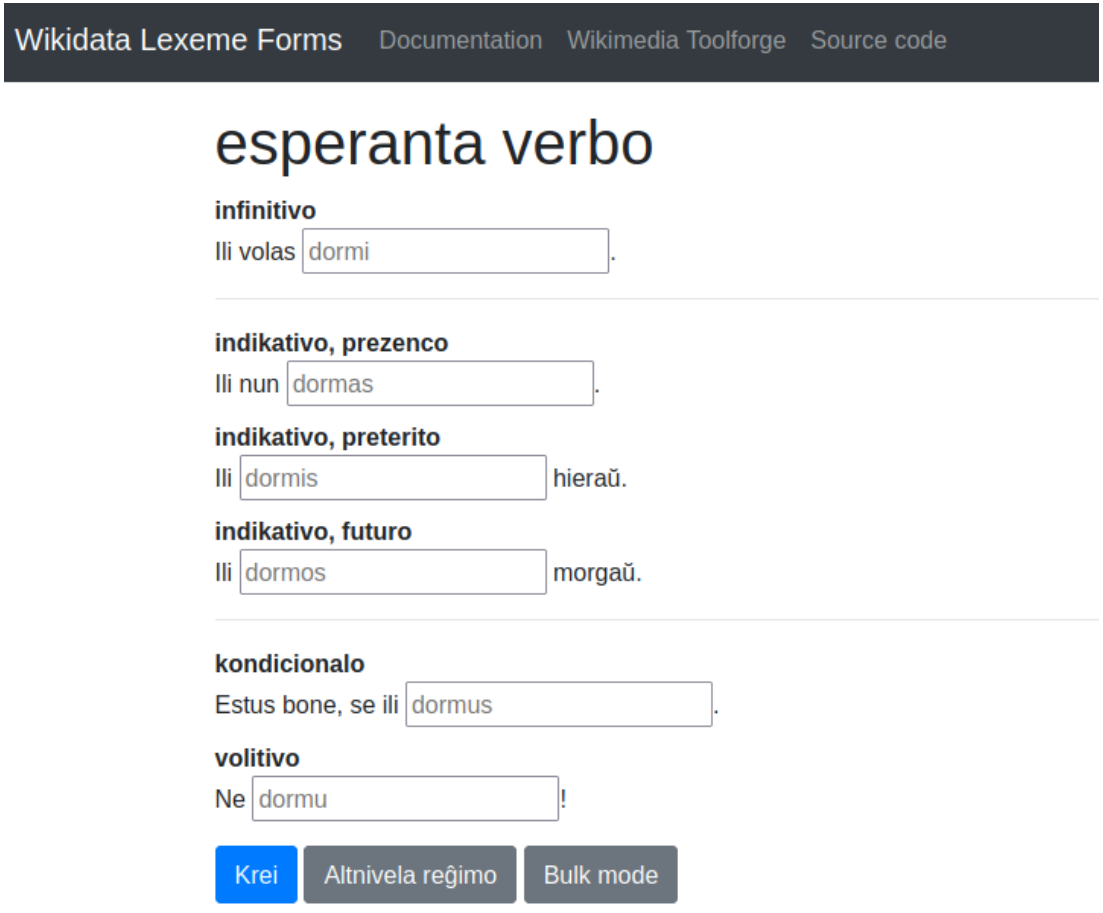
Creating new lexeme

www.wikidata.org/wiki/Special:NewLexeme



Templates for different parts of speech

tools.wmflabs.org/lexeme-forms/



Templates for different parts of speech

batch loading

API access

SPARQL for Wikidata

- query.wikidata.org

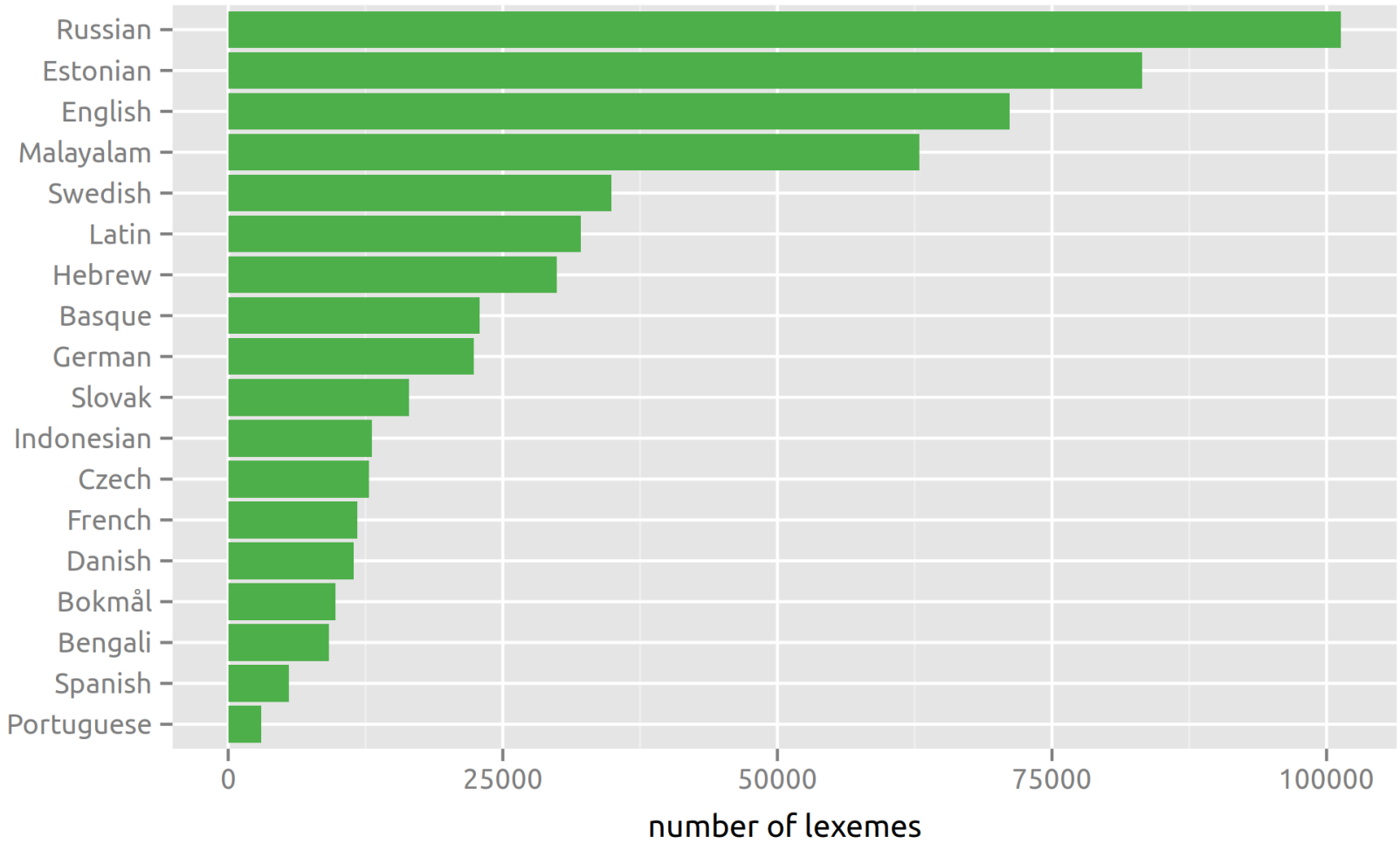
Pywikibot

- www.mediawiki.org/wiki/Manual:Pywikibot

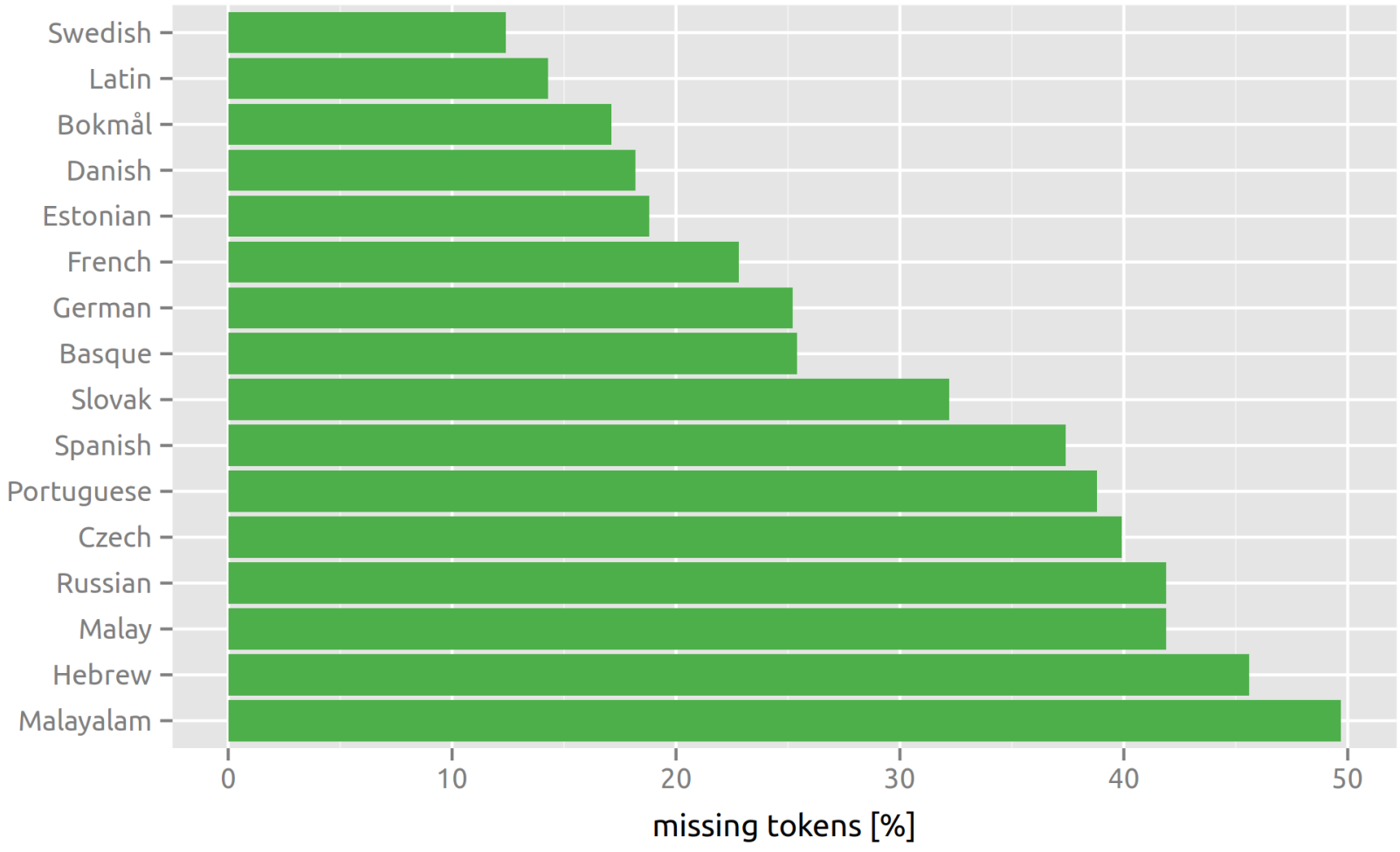
LexData

- nudin.github.io/LexData

Lexemes count by language



Wikipedia coverage by language



Czech CC0 dictionaries

Czech morphological dictionary

+ Wikidata lexemes

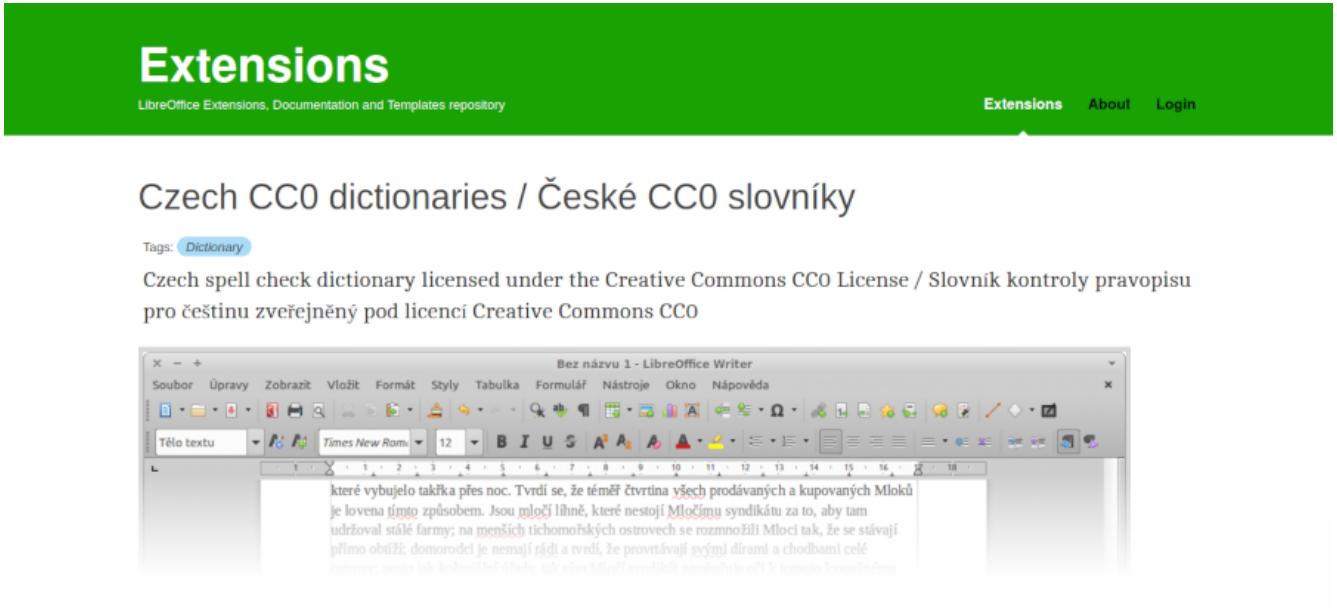
= Czech CC0 dictionaries

experimental

Czech CC0 dictionaries

LibreOffice extension

extensions.libreoffice.org/extensions/czech-cc0-dictionaries-ceske-cc0-slovniky

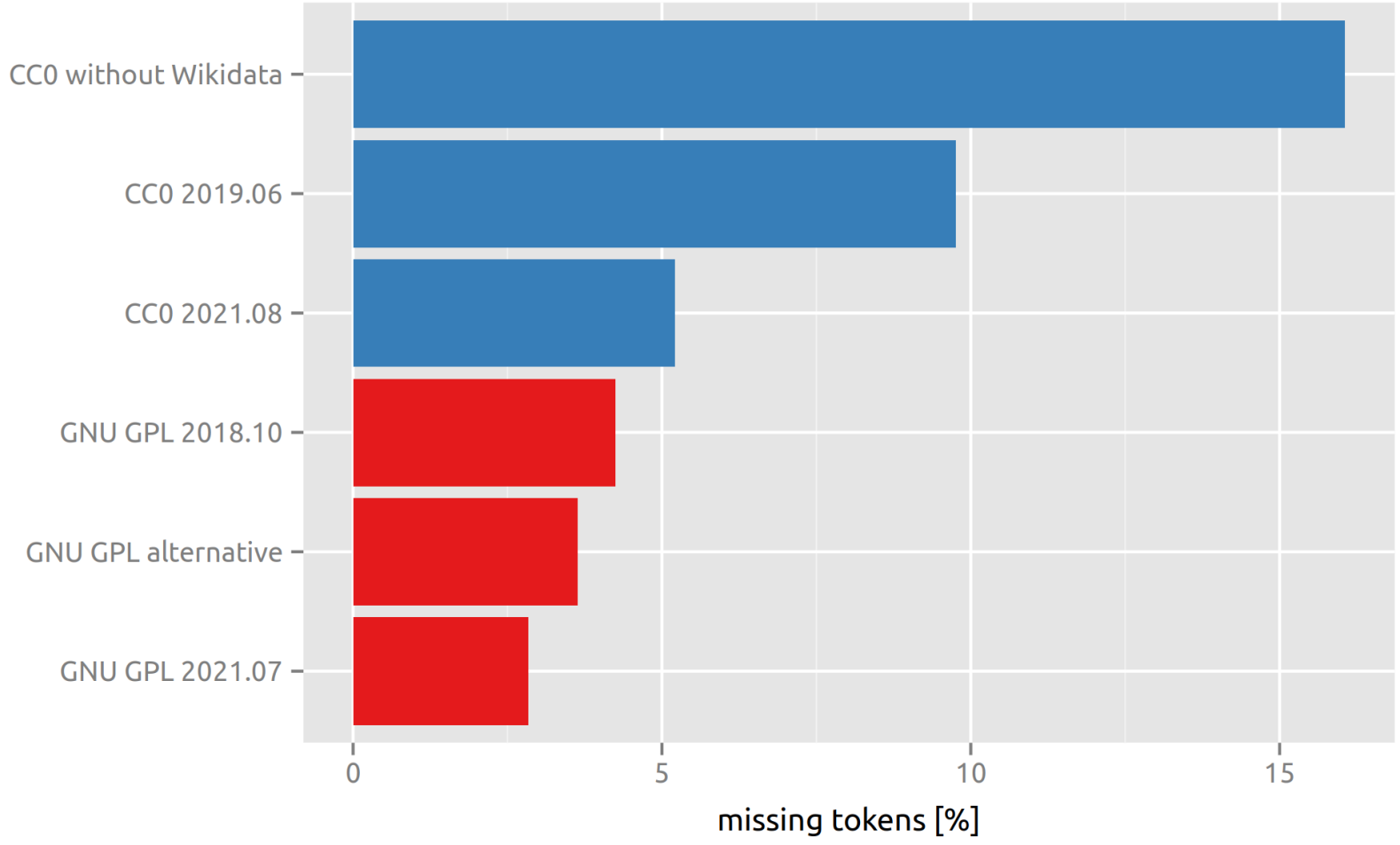


Mozilla addon

ceskeslovniky.cz

Dictionaries comparison

missing tokens in ~250 000 words



Summary

lexemes at Wikidata

- word database licensed under CC0
- for any language
- user interface, API
- one of sources for the experimental Czech dictionary

check the lexemes

use them

contribute to them