Lexémy na Wikidatech a slovník pro kontrolu pravopisu

Wikikonference 2019

Stanislav Horáček

česká skupina kolem LibreOffice
The Document Foundation

Wikipedista:Strepon

23. listopadu 2019

Úvod

český slovník pro kontrolu pravopisu

- Hunspell
- licence GNU GPL

Toto je cesky slovnik pro kontrolu pravopisu zalozeny na ceskem slovniku pro ispell, verze z 29. 10. 2006, ktery vytvoril Petr Kolar spolu s desitkami dalsich prispevatelu.

Úvod

Český tvarotvorný slovník slovní zásoba

- Masarykova univerzita
- únor 2019, licence public domain (= CC0)
- analýza jazykového korpusu
- podstatná jména, přídavná jména, slovesa
- ~60 000 základních tvarů
- github.com/plin/slovnik

Úvod

Slovníková data na Wikidatech rozhraní

- oddělený prostor
- databáze slovní zásoby
- základní jednotka lexém
- ~3500 základních tvarů v češtině
- během roku 2018, licence CC0
- www.wikidata.org/wiki/Wikidata:Lexicographical_data/cs

Slovníková data na Wikidatech

projekt Wikimedia Foundation

Slovníková data na Wikidatech

projekt Wikimedia Foundation

možné využití

- slovník pro kontrolu pravopisu
- výkladový slovník
- slovník dělení slov
- slovník synonym, antonym, …
- cizojazyčné slovníky a překladače
- nástroje pro kontrolu gramatiky
- …

Lemma, významy, tvary

(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

Charakteristika, etymologie

(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

Příznaky

(L58354) **měďák**

cs

edit

Language Czech
Lexical category noun

Statements

grammatical gender	masculine	edit
	0 references	
		+ add reference

Senses

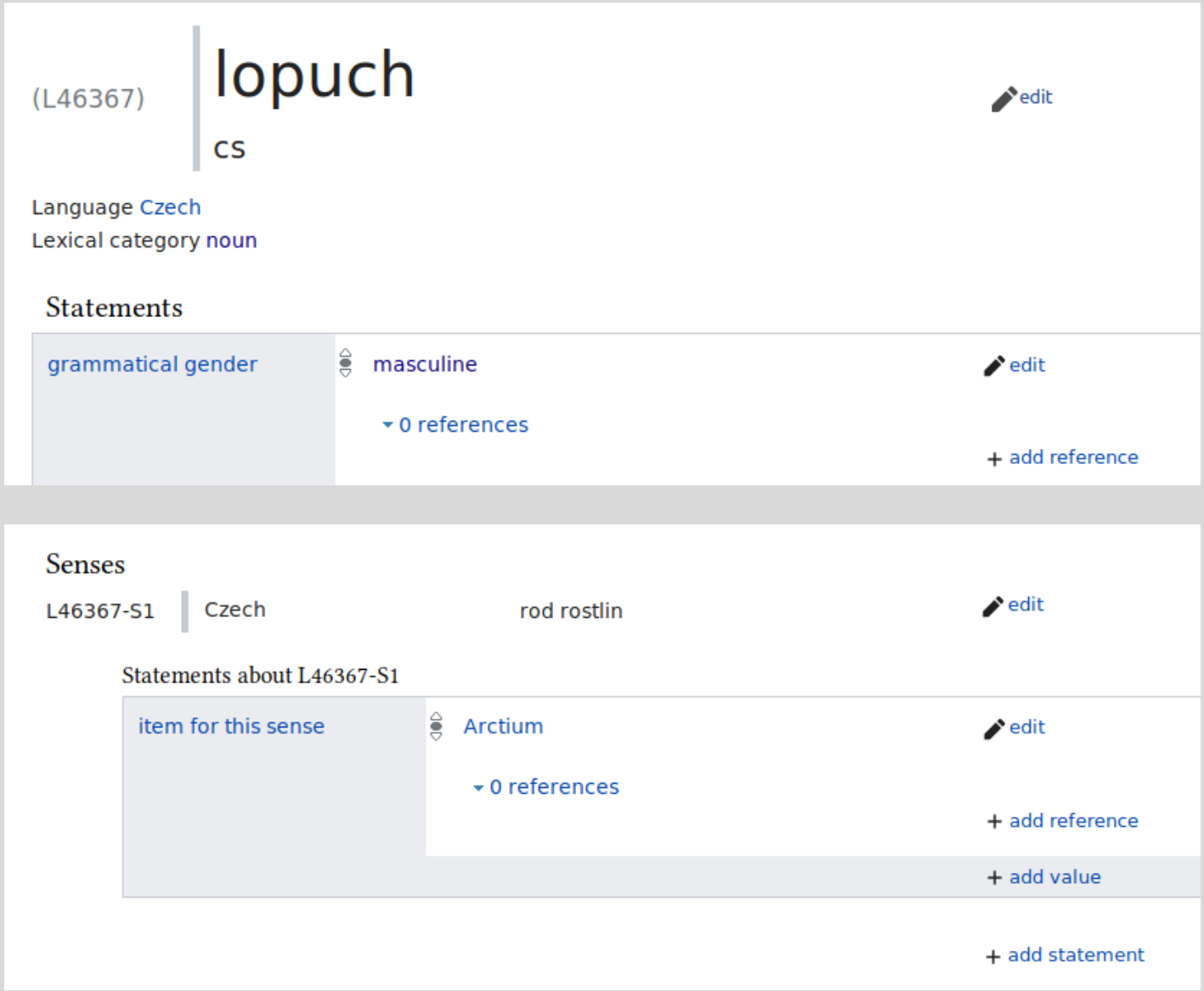
L58354-S1 | Czech | mince z mědi | edit

Statements about L58354-S1

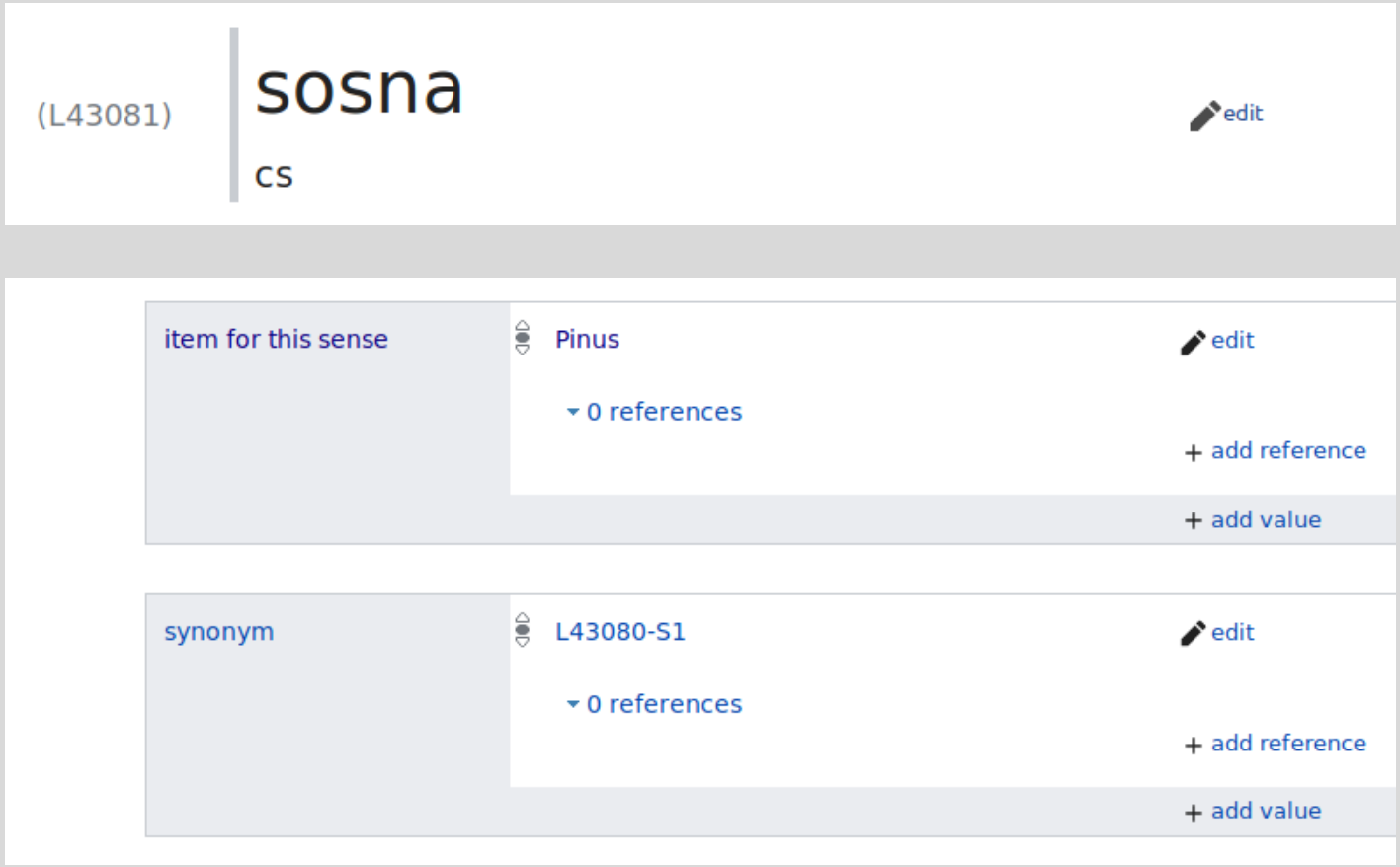
language style	Common Czech	edit
	0 references	
		+ add reference
		+ add value

+ add statement

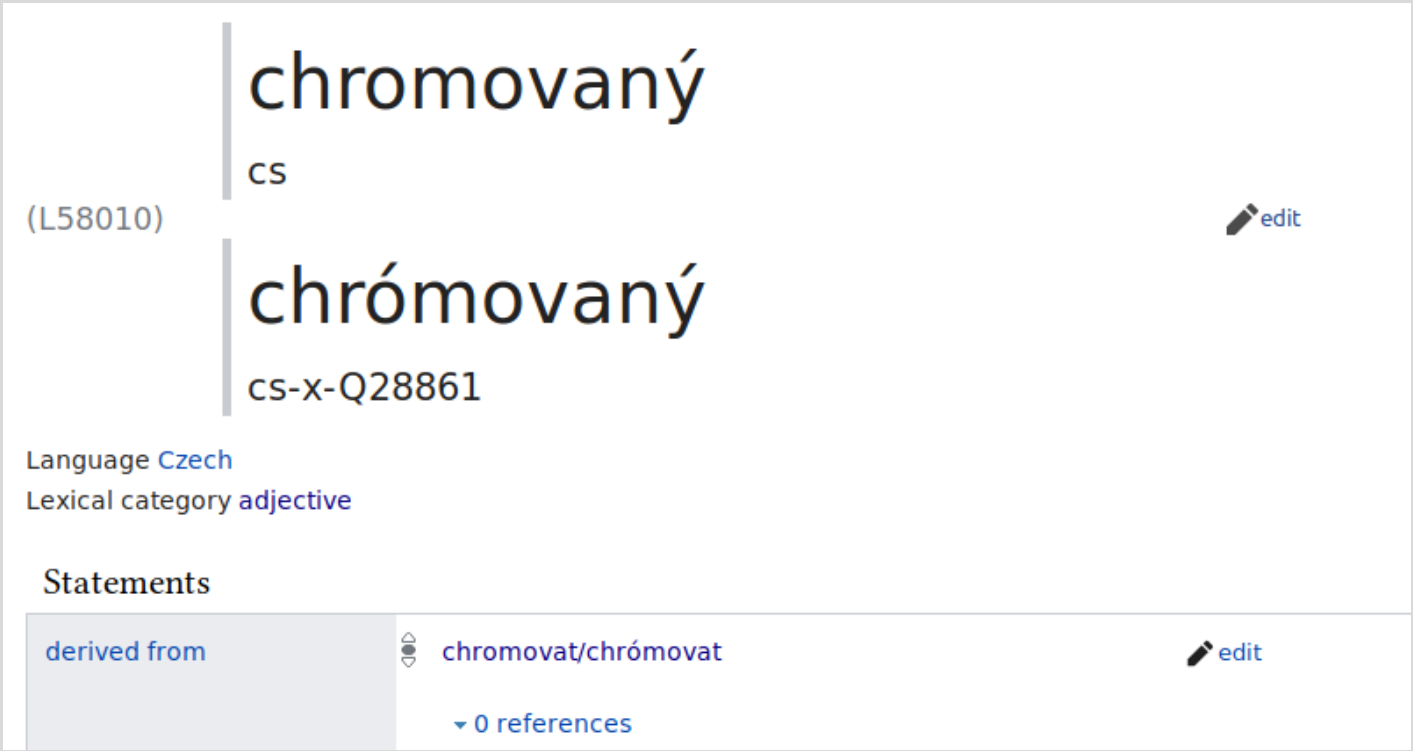
Propojení s položkami Wikidat



Vztahy mezi lexémy



Varianty


chromovaný
cs
(L58010)
edit
chrómovaný
cs-x-Q28861
Language Czech
Lexical category adjective
Statements
derived from
chromovat/chrómovat
edit
0 references
L58010-F3
chromovanému
cs
chrómovanému
cs-x-Q28861
edit
Grammatical features singular, dative case, animate masculine
Statements about L58010-F3
+ add statement

Výslovnost

(L442) **být**

cs

edit

L442-F14 | být | cs

edit

Grammatical features infinitive

Statements about L442-F14

pronunciation audio

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

Příklady užití

(L205210) **Pavel** edit

cs

S Pavlem si dobře rozumíme. (Czech) edit

demonstrates form Pavlem

demonstrates sense L205210-S1

0 references

+ add reference

My o Pavlu, a Pavel za dveřmi. (Czech) edit

demonstrates form Pavel

Pavlu

demonstrates sense L205210-S1

0 references

+ add reference

Ve třídě máme dva Pavly. (Czech) edit

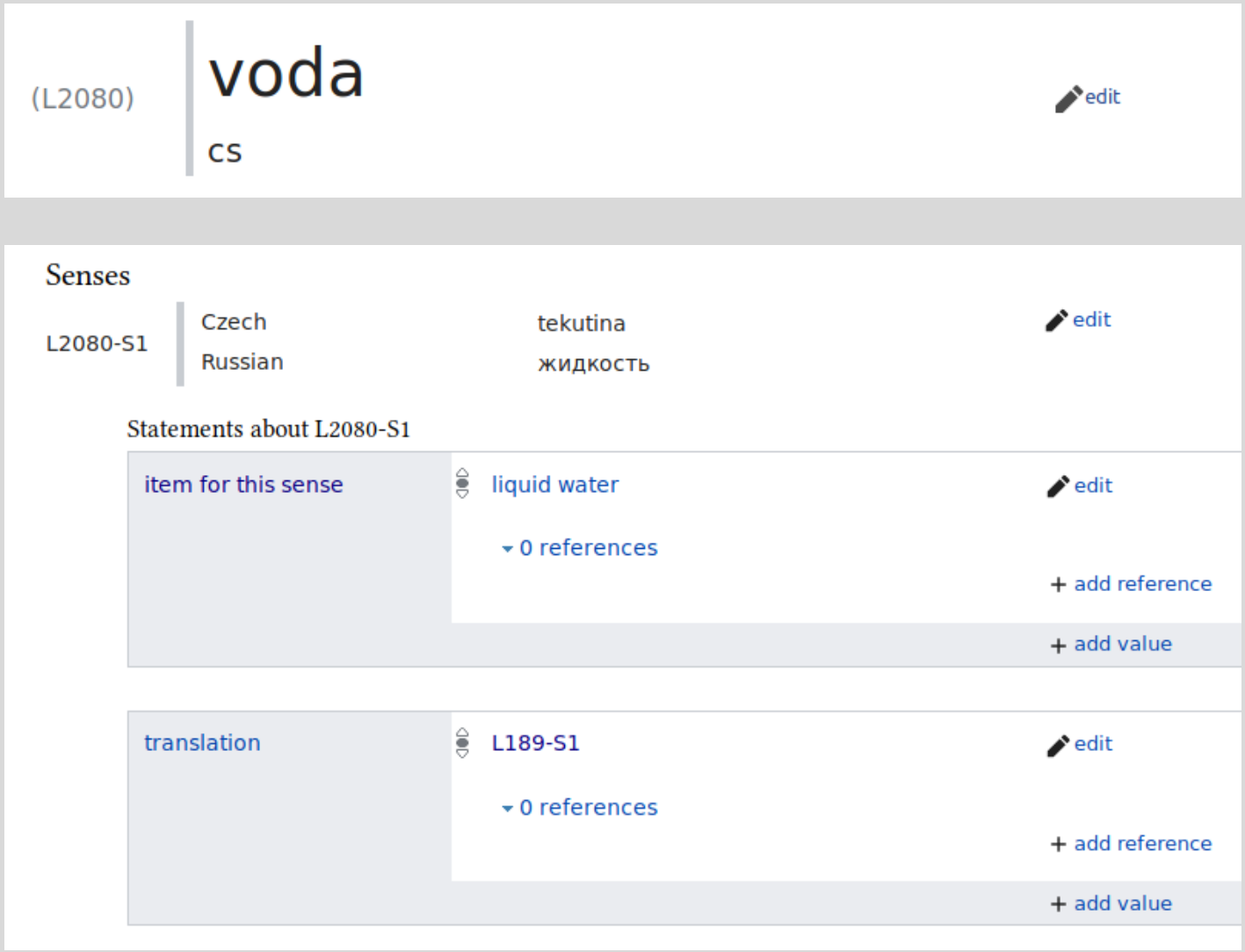
demonstrates form Pavly

demonstrates sense L205210-S1

0 references

+ add reference

Překlady?


(L2080)
voda
cs
edit
Senses
L2080-S1
Czech
tekutina
Russian
жидкость
edit
Statements about L2080-S1
item for this sense
liquid water
edit
0 references
+ add reference
+ add value
translation
L189-S1
edit
0 references
+ add reference
+ add value

Citování?

(L10856) ječmen

cs

edit

L10856-F3 ječmena

cs

edit

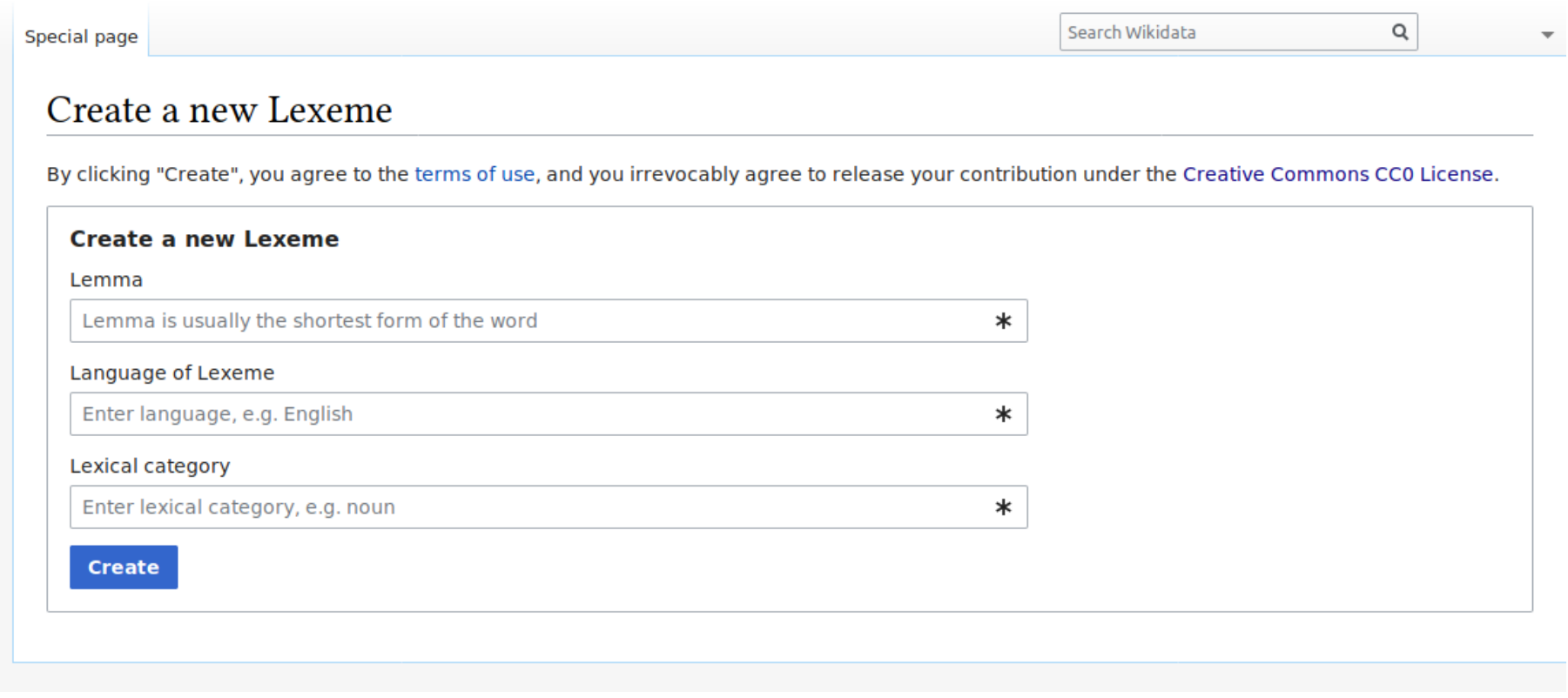
Grammatical features singular, genitive case

Statements about L10856-F3

described by source	Internet Language Reference Book	edit
	URL http://prirucka.ujc.cas.cz/?slovo=je%C4%8Dmen	
	0 references	
	+ add reference	
	+ add value	

Vytvoření nového lexému

www.wikidata.org/wiki/Special:NewLexeme



Šablony pro různé slovní druhy

tools.wmflabs.org/lexeme-forms/

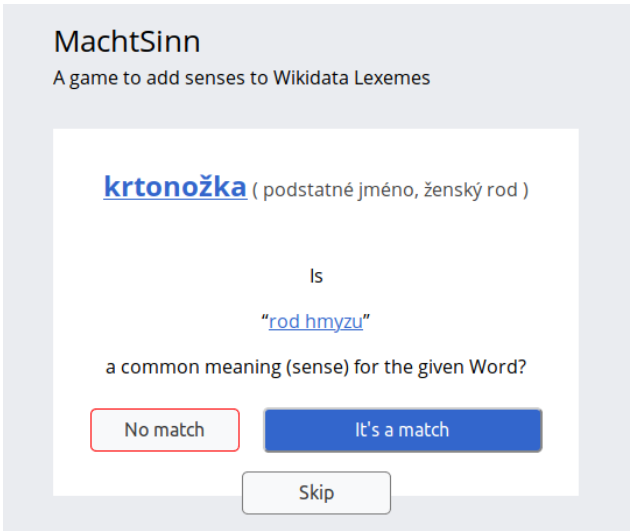
Šablony pro různé slovní druhy

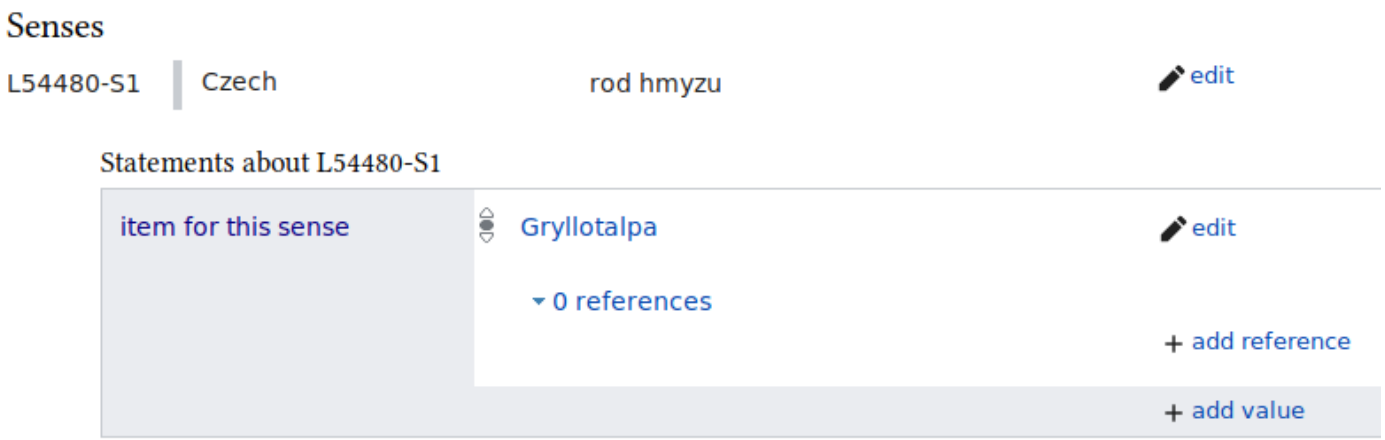
hromadné zadání

Nástroje

MachtSinn – propojení významů s položkami Wikidat

tools.wmflabs.org/machtsinn/





Přístup přes API

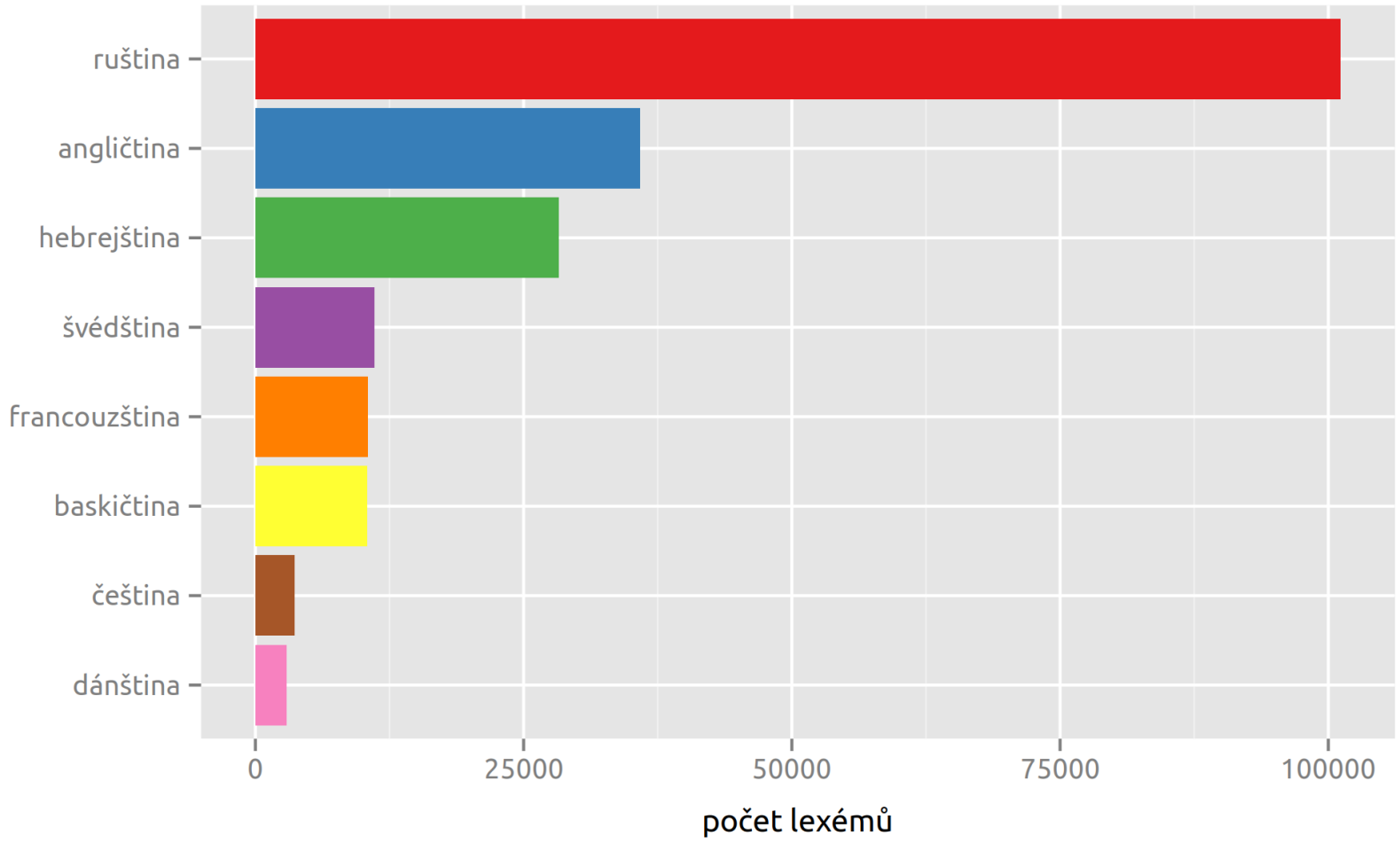
dotazovací jazyk SPARQL pro Wikidata

- query.wikidata.org

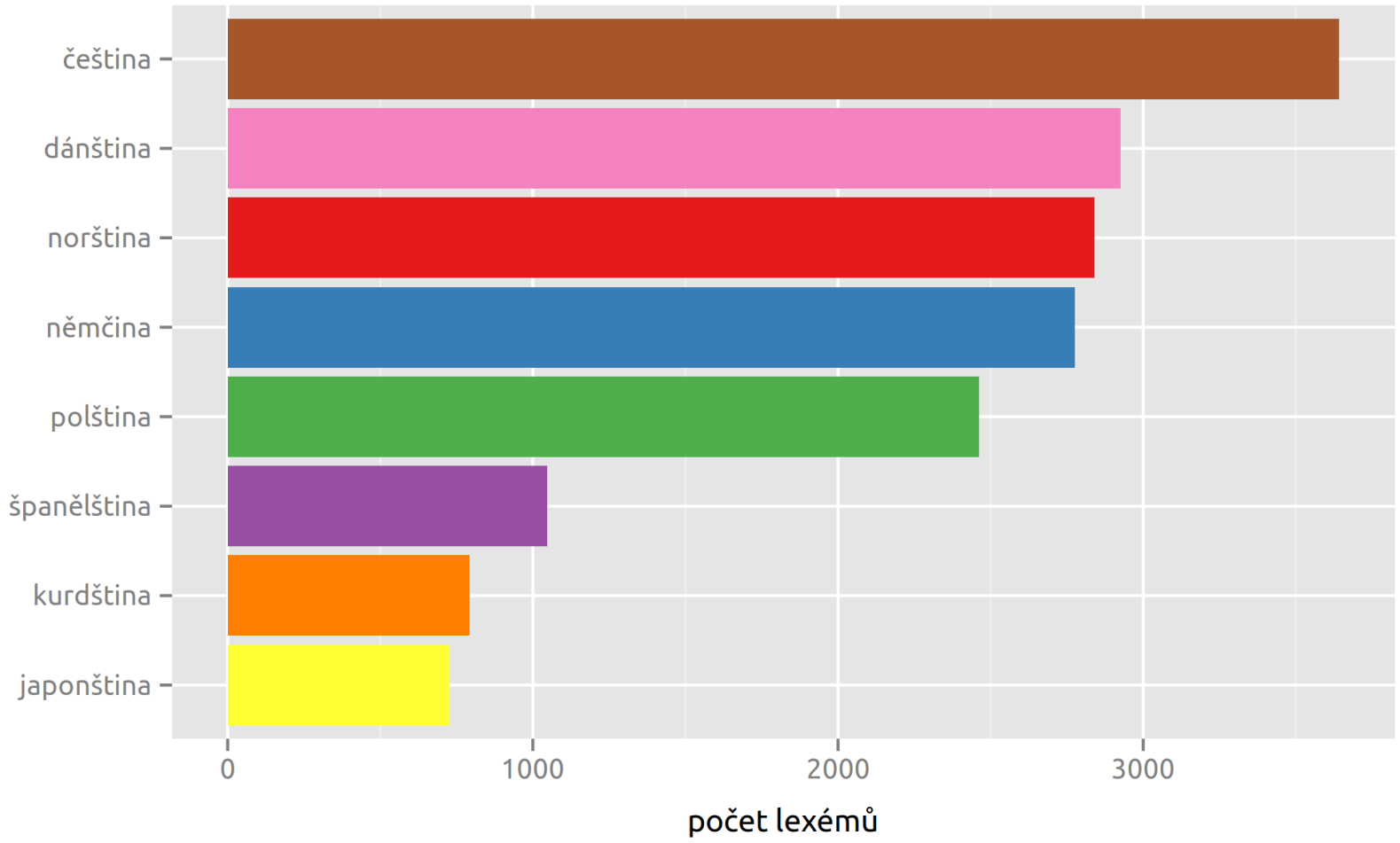
Pywikibot

- www.mediawiki.org/wiki/Manual:Pywikibot
- neúplná podpora pro lexémy na Wikidatech
- phabricator.wikimedia.org/T189321
- nefunkční generátory – WikidataSPARQLPageGenerator

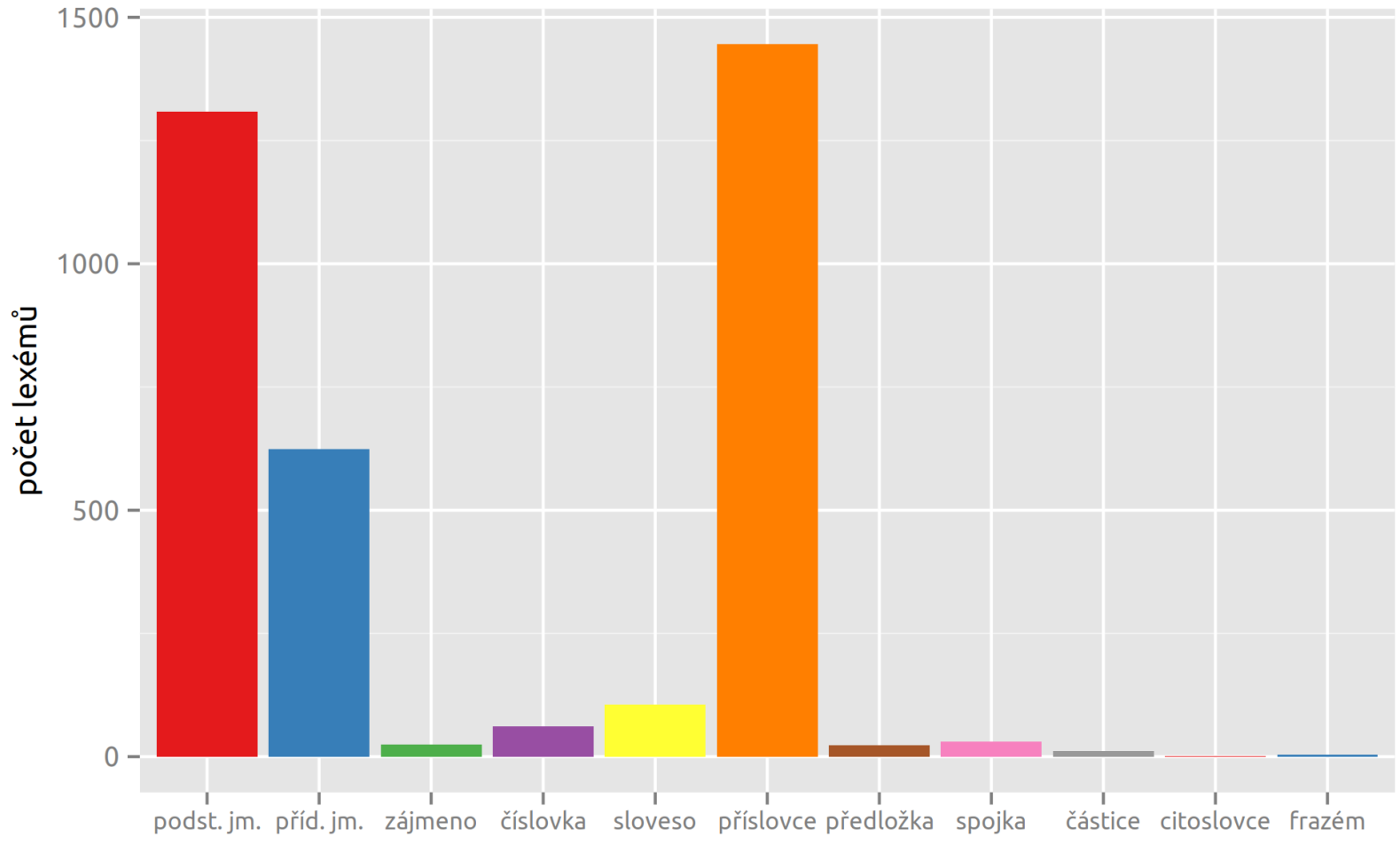
Lexémy podle jazyků



Lexémy podle jazyků



Slovní druhy pro češtinu



České CC0 slovníky

Český tvarotvorný slovník

\+ slovníková data z Wikidat

= české CC0 slovníky

experimentální!

České CC0 slovníky

rozšíření pro LibreOffice

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

EXTENSIONS & TEMPLATES

Log in Register Search Site Search

only in current section

Home Extensions Templates Events

Home › Extensions › Czech CC0 dictionaries / České CC0 slovníky

Czech CC0 dictionaries / České CC0 slovníky

Czech spell check dictionary licensed under the Creative Commons CC0 License / Slovník kontroly pravopisu pro češtinu zveřejněný pod licencí Creative Commons CC0

13 1

Project Description

The dictionary uses data from the Czech morphological dictionary created by the Masaryk University in Brno and from Czech Wikidata lexemes.

Repository of this dictionary can be found at GitLab.

The dictionary is considered as experimental (many words missing), with a lot of space for improvements.

For more details and instructions how to contribute, see the webpage ceskeslovniky.cz.

Slovník využívá data Českého tvaroslovného slovníku vytvořeného na Masarykově univerzitě v Brně a česká slovníková data z Wikidat.

Repozitář s tímto českým slovníkem je k dispozici na GitLabu.

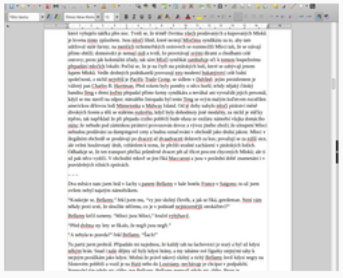
Slovník je třeba považovat za experimentální (mnoho slov chybí) a je zde značný prostor pro jeho vylepšování.

Podrobnosti a návod jak se zapojit naleznete na stránkách ceskeslovniky.cz.

Project Resources

- External Project Page

Screenshot



Install Instructions

To install an extension, follow these steps:

- Download an extension and save it anywhere on your computer.
- In LibreOffice, select Tools --> Extension Manager from the menu bar.
- In the Extension Manager dialog click Add.

České CC0 slovníky

srovnání úspěšnosti

– procentuální podíl slov označených jako chybná

	červen 2019	listopad 2019	GNU GPL
Dobrodružství Sherlocka Holmese	9,02	6,78	2,63
Evangelium podle Jana	7,46	4,61	0,67
LibreOffice Writer: Praktický průvodce	6,00	4,56	3,33
R.U.R.	16,82	11,87	8,37
Ústava České republiky	7,89	5,57	0,90
Wikikonference 2019	14,68	12,61	8,51

Jak vylepšit

- doplňování slov na Wikidata
- import z Tvarotvorného slovníku
- import z Wikislovníku
- import z Wikidat
- nové šablony
- nové nástroje
- kampaň pro určité texty

Shrnutí

slovníková data na Wikidatech

- databáze slovní zásoby pod licencí CC0
- wikidata.org/wiki/Wikidata:Lexicographical_data/cs
- wikidata.org/wiki/Wikidata_talk:Lexicographical_data
- wikidata.org/wiki/Wikidata:Mezi_bajty

nový český slovník kontroly pravopisu

- experimentální, pod licencí CC0
- Tvarotvorný slovník a lexémy z Wikidat
- budoucnost?
- ceskeslovniky.cz