european lexicographic infrastructure

D4.5 Sample dictionary drafts

Author(s): Miloš Jakubíček, Vít Suchomel, Marek Medveď, Marek Blahuš, Vojtěch Kovář

Date: January 31st, 2022

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

H2020-INFRAIA-2016-2017
Grant Agreement No. 731015
ELEXIS - European Lexicographic Infrastructure

D4.5 Sample dictionary drafts

Deliverable Number: D4.5
Dissemination Level: Public
Delivery Date: January 31st, 2022
Version: 1
Author(s): Miloš Jakubíček, Vít Suchomel,
Marek Medveď, Marek Blahuš, Vojtěch Kovář

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Deliverable/Document Information

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Document History

Version Date	Changes/Approval	Author(s)/Approved by
1, January 31st	Initial Draft	Jakubíček

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Languages

Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Irish, Hebrew, Croatian, Hungarian, Italian, Lithuanian, Latvian, Dutch, Polish, Portuguese, Romanian, Slovak, Slovenian, Swedish

Data format (common for all languages)

- NVH (plain text)
- NVH schema:

```
hw: +
    freq:
    example: +
    collocs:
        gramrel: *
            colloc: +
                freq:
                score:
                commonest:
    thesaurus:
        item: *
    form: *
        freq:
```

Entry composition (common for all languages)

- headword
  - denoted as `hw` in the NVH file
  - 10,000 most frequent lemposes (lemma-pos combinatios) according to the document frequency in the corpora that are part of D4.6
- headword frequency
  - the document frequency is available as `hw.freq`
- headword examples
  - up to 10 examples are available as `hw.example`
  - generated using the GDEX module in Sketch Engine [1]
- collocations
  - grouped by grammatical relations (`hw.collocs.gramrel`) according to word sketches in Sketch Engine [2]
  - up to 5 collocations available for each grammatical relation (`hw.collocs.gramrel.colloc`), each associated with a raw frequency, logDice score [3] and commnest phrase that it occurs in [4]
- thesaurus items
  - built using the distributional thesaurus in Sketch Engine [5]

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

    - up to 5 thesaurus items available as `hw.thesaurus.item`
- inflectional forms
    - all inflectional forms of the headword lemma with frequency higher than 1 % of the headword lemma available as `hw.form` and frequency as `hw.form.freq`.

Availability

1. ELEXIS Cloud - https://cloud.elex.is
2. Lexonomy - https://www.lexonomy.eu

Access conditions

Bibliography

[1] KILGARRIFF, Adam, et al. GDEX: Automatically finding good dictionary examples in a corpus. In: Proceedings of the XIII EURALEX international congress. Barcelona, Spain: Documenta Universitaria, 2008. p. 425-432.

[2] KILGARRIFF, Adam, et al. The Sketch Engine: ten years on. Lexicography, 2014, 1.1: 7-36.

[3] RYCHLÝ, Pavel. A Lexicographer-Friendly Association Score. In: RASLAN. 2008. p. 6-9.

[4] KILGARRIFF, Adam, et al. Longest–commonest Match. In: Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd, 2015. p. 11-13.

[5] RYCHLÝ, Pavel; KILGARRIFF, Adam. An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. 2007. p. 41-44.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.