european lexicographic infrastructure

D4.6

Semantically annotated corpora

Author(s): Miloš Jakubíček, Vít Suchomel, Federico Martelli, Roberto Navigli

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

D4.6 Semantically annotated corpora

Deliverable Number: D4.6
Dissemination Level: Public
Delivery Date: January 31st, 2022
Version: 1
Author(s): Miloš Jakubíček, Vít Suchomel,
Federico Martelli, Roberto Navigli

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.

Data format (common for all languages)

- Compressed plain text files: GZIP compression.
- Plain text file with the form of a vertical text consisting of three tab-separated columns.
- Each line in the vertical file represents a token consisting of attributes word form, part-of-speech tag and lemma in TAB-separated columns,
- additional morphological categories may be put into additional columns (in case they are provided by the morphological analyser/tagger for the particular language).
- Original documents are split into paragraphs and sentences.
- Duplicate and near-duplicate paragraphs were removed (i.e. paragraphs consisting of at least 5 words and containing more than 90 % of 5-tuples of words present elsewhere in this data were removed from this data).
- Encoding: UTF-8.
- Attribute gender_lemma is the base form of the word (lemma) respecting the gender of the particular word form.

Data structures (common for all languages)

- Encoded as XML elements with attributes.
- Documents, marked by <doc> and </doc>, representing a single web page or a single text document,
- paragraphs, marked by and , delimited by , or double
 in HTML documents or double end of line in other documents in the source data,
- sentences, marked by <s> and </s>, identified using punctuation,
- word join markers (“glue”), marked <g/>, indicate that there was no space between the surrounding words in the original document.

Corpora

ELEXIS Bulgarian Web 2020

Corpus size in tokens: 1224491231
Time span of crawling: 2021-09 to 2021-10
Tokenization, lemmatization and morphological tagger pipeline: TreeTagger (LC pipeline v. 2)
Token attributes (columns): word, tag, lempos
POS suffixes: ",adjective,-a,adverb,-d,conjunction,-c,family name,-h,interjection,-i,noun,-n,numeral,-m,particle,-t,preposition,-r,pronoun,-p,verb,-v"
Tagset documentation: https://www.sketchengine.co.uk/bulgarian-treebank-part-of-speech-tagset/

ELEXIS Czech Web 2019

Corpus size in tokens: 1151331996

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.

Time span of crawling: 2019-12 to 2020-02
Tokenization, lemmatization and morphological tagger pipeline: Majka (LC pipeline v. 2)
Token attributes (columns): word, tag, lempos, gender_lemma
POS suffixes: ",noun,-n,adjective,-j,pronoun,-d,numeral,-m,verb,-v,adverb,-a,preposition,-p,conjunction,-c"
Tagset documentation: https://www.sketchengine.co.uk/tagset-reference-for-czech

ELEXIS Danish Web 2020

Corpus size in tokens: 1172484020
Time span of crawling: 2020-06 to 2020-08
Tokenization, lemmatization and morphological tagger pipeline: TreeTagger (LC pipeline v. 2)
Token attributes (columns): word, tag, lempos, gender_lemma
POS suffixes: ",verb,-v,adjective,-a,numeral,-l,noun,-n,pronoun,-p,adverb,-d,interjection,-i,preposition,-t,conjunction,-c,unique,-u,residual,-x"
Tagset documentation: https://www.sketchengine.co.uk/danish-epos-part-of-speech-tagset/

ELEXIS German Web 2020

Corpus size in tokens: 1228166621
Time span of crawling: 2020-06 to 2020-07
Tokenization, lemmatization and morphological tagger pipeline: RFTagger (LC pipeline v. 3)
Token attributes (columns): word, tag, lempos
POS suffixes: ",adjective,-j,preposition,-i,adverb,-a,conjunction,-c,noun,-n,numeral,-m,pronoun,-p,verb,-v"
Tagset documentation: https://www.sketchengine.co.uk/german-rftagger-part-of-speech-tagset/

ELEXIS Greek Web 2020

Corpus size in tokens: 1193338965
Time span of crawling: 2019-12 to 2020-01
Tokenization, lemmatization and morphological tagger pipeline: TreeTagger (LC pipeline v. 2)
Token attributes (columns): word, tag, lempos
POS suffixes: ",noun,-n,adjective,-j,numeral,-m,article,-a,verb,-v,pronoun,-p,adverb,-r,adposition,-i,conjunction,-c,interjection,-q,particle,-e,residual,-y"
Tagset documentation: https://www.sketchengine.co.uk/greek-intera-part-of-speech-tagset/

ELEXIS English Web 2020

Corpus size in tokens: 1178335836
Time span of crawling: 2020-12
Tokenization, lemmatization and morphological tagger pipeline: TreeTagger (LC pipeline v. 3.1)
Token attributes (columns): word, tag, lempos

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.

POS suffixes: ",adjective,-j,adverb,-a,conjunction,-c,noun,-n,preposition,-i,pronoun,-d,verb,-v"
Tagset documentation: https://www.sketchengine.co.uk/english-treetagger-pipeline-3/

ELEXIS Spanish Web 2020

Corpus size in tokens: 1166796560
Time span of crawling: 2020-06
Tokenization, lemmatization and morphological tagger pipeline: FreeLing (LC pipeline v. 5)
Token attributes (columns): word, tag, lempos, gender_lemma, tags, morphemes
POS suffixes: ",adjective,-j,adverb,-r,conjunction,-c,noun,-n,numeral,-m,preposition,-i,pronoun,-p,verb,-v"
Tagset documentation: https://www.sketchengine.co.uk/spanish-freeling-part-of-speech-tagset/

ELEXIS Estonian Web 2021

Corpus size in tokens: 1237456410
Time span of crawling: 2021-05 to 2021-09
Tokenization, lemmatization and morphological tagger pipeline: NLTK+Filosoft (LC pipeline v. 2)
Token attributes (columns): word, longtag, lempos, features, root_tokens, root, suffix, clitic
POS suffixes: ",adjective positive,-a,adjective comparative,-c,adverb,-d,indeclinable adjective,-g,proper noun,-h,interjection,-i,conjuction,-j,adposition,-k,cardinal number,-n,ordinal number,-o,pronoun,-p,common noun,-s,superlative adjective,-u,verb,-v,verb particle,-x"
Tagset documentation: https://www.sketchengine.co.uk/estonian-filosoft-part-of-speech-tagset/

ELEXIS Finnish Web 2020

Corpus size in tokens: 1239257389
Time span of crawling: 2019-10 to 2020-01
Tokenization, lemmatization and morphological tagger pipeline: TreeTagger (LC pipeline v. 3)
Token attributes (columns): word, tag, lempos
POS suffixes: ",adjective,-j,adverb,-r,conjunction,-c,interjection,-i,noun,-n,numeral,-m,pronoun,-p,verb,-v"
Tagset documentation: https://www.sketchengine.eu/finnish-treetagger-part-of-speech-tagset/

ELEXIS French Web 2020

Corpus size in tokens: 1256258596
Time span of crawling: 2020-12 to 2021-02
Tokenization, lemmatization and morphological tagger pipeline: FreeLing (LC pipeline v. 2)
Token attributes (columns): word, tag, lempos, gender_lemma, tags, morphemes

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.

POS suffixes: ",adjective,-j,adverb,-r,conjunction,-c,noun,-n,numeral,-m,preposition,-i,pronoun,-p,verb,-v"
Tagset documentation: https://www.sketchengine.eu/french-freeling-part-of-speech-tagset/

ELEXIS Irish Web 2021

Corpus size in tokens: 66441473
Time span of crawling: 2020-01 and 2021-03
Tokenization, lemmatization and morphological tagger pipeline: RFTagger (LC pipeline v. 1)
Token attributes (columns): word, tag, lempos
POS suffixes: ",adjective,-j,adverb,-a,conjunction,-c,noun,-n,numeral,-m,preposition,-i,pronoun,-d,verb,-v"
Tagset documentation: https://universaldependencies.org/ga/index.html

ELEXIS Hebrew Web 2021

Corpus size in tokens: 1194724650
Time span of crawling: 2020-02 to 2020-04 and 2020-12 and 2021-02 to 2021-03
Tokenization, lemmatization and morphological tagger pipeline: YAP (LC pipeline v. 1)
Token attributes (columns): word, tag, lempos, morphemes, id, head, deprel
POS suffixes: ",adjective,-j,conjunction,-c,interjection,-i,noun,-n,numeral,-m,preposition,-p,pronoun,-d,verb,-v"
Tagset documentation: https://www.sketchengine.eu/hebrew-yap-part-of-speech-tagset/

ELEXIS Croatian Web 2020

Corpus size in tokens: 1194668784
Time span of crawling: 2020-07 to 2020-12
Tokenization, lemmatization and morphological tagger pipeline: RFTagger (LC pipeline v. 1)
Token attributes (columns): word, tag, lempos, gender_lemma
POS suffixes: ",adjective,-a,adverb,-r,conjunction,-c,noun,-n,numeral,-m,preposition,-s,pronoun,-p,verb,-v"
Tagset documentation: https://www.sketchengine.co.uk/multext-east-croatian-part-of-speech-tagset/

ELEXIS Hungarian Web 2020

Corpus size in tokens: 1233677633
Time span of crawling: 2020-12 to 2021-02
Tokenization, lemmatization and morphological tagger pipeline: HunPOS (LC pipeline v. 2)
Token attributes (columns): word, tag, lempos
POS suffixes: ",noun,-n,verb,-v,adjective,-j,adverb,-r"
Tagset documentation: https://www.sketchengine.co.uk/hungarian-emmorph-based-part-of-speech-tagset/

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.

ELEXIS Italian Web 2020

Corpus size in tokens: 1195071361
Time span of crawling: 2020-12
Tokenization, lemmatization and morphological tagger pipeline: TreeTagger (LC pipeline v. 2)
Token attributes (columns): word, tag, lempos
POS suffixes: ",adjective,-j,adverb,-r,conjunction,-c,noun,-n,numeral,-m,preposition,-i,pronoun,-p,verb,-v"
Tagset documentation: https://www.sketchengine.co.uk/italian-treetagger-part-of-speech-tagset/

ELEXIS Lithuanian Web 2021

Corpus size in tokens: 1065842741
Time span of crawling: 2021-08 to 2021-10
Tokenization, lemmatization and morphological tagger pipeline: Lithuanian tagger (LC pipeline v. 1)
Token attributes (columns): word, tag, lempos
POS suffixes: ",adjective,-j,adverb,-a,conjunction,-c,noun,-n,numeral,-m,pronoun,-d,verb,-v"
Tagset documentation: http://nl.ijs.si/ME/Vault/V4/msd/html/index.html

ELEXIS Latvian Web 2021

Corpus size in tokens: 1293988029
Time span of crawling: 2021-08 to 2021-10
Tokenization, lemmatization and morphological tagger pipeline: Latvian tagger (LC pipeline v. 1)
Token attributes (columns): word, tag, lempos
POS suffixes: ",adjective,-j,adverb,-a,conjunction,-c,noun,-n,numeral,-m,pronoun,-d,verb,-v"
Tagset documentation: https://www.sketchengine.co.uk/latvian-part-of-speech-tagset/

ELEXIS Dutch Web 2020

Corpus size in tokens: 1189551736
Time span of crawling: 2020-06 to 2020-07
Tokenization, lemmatization and morphological tagger pipeline: FreeLing (LC pipeline v. 2)
Token attributes (columns): word, tag, lempos
POS suffixes: ",adjective,-j,adverb,-r,conjunction,-c,noun,-n,numeral,-m,preposition,-i,pronoun,-p,verb,-v"
Tagset documentation: https://www.sketchengine.eu/dutch-treetagger-tagset/

ELEXIS Polish Web 2019

Corpus size in tokens: 1201790229
Time span of crawling: 2019-12
Tokenization, lemmatization and morphological tagger pipeline: RFTagger (LC pipeline v. 1)

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.

Token attributes (columns): word, tag, lempos, gender_lemma
POS suffixes: ",adjective,-j,adverb,-a,conjunction,-c,noun,-n,preposition,-i,pronoun,-d,verb,-v"
Tagset documentation: https://www.sketchengine.co.uk/polish-nkjp-part-of-speech-tagset/

ELEXIS Portuguese Web 2020

Corpus size in tokens: 1200470569
Time span of crawling: 2020-06 to 2020-07
Tokenization, lemmatization and morphological tagger pipeline: FreeLing (LC pipeline v. 4)
Token attributes (columns): word, tag, lempos, ao, tags, morphemes
POS suffixes: ",adjective,-j,adverb,-r,conjunction,-c,noun,-n,numeral,-m,preposition,-i,pronoun,-p,verb,-v"
Tagset documentation: https://www.sketchengine.co.uk/portuguese-freeling-part-of-speech-tagset/

ELEXIS Romanian Web 2021

Corpus size in tokens: 1189744243
Time span of crawling: 2021-03
Tokenization, lemmatization and morphological tagger pipeline: RFTagger (LC pipeline v. 1)
Token attributes (columns): word, tag, lempos
POS suffixes: ",adjective,-a,adverb,-r,conjunction,-c,noun,-n,numeral,-m,pronoun,-p,verb,-v"
Tagset documentation: https://www.sketchengine.co.uk/romanian-tagset/

ELEXIS Slovak Web 2021

Corpus size in tokens: 1198933594
Time span of crawling: 2021-11
Tokenization, lemmatization and morphological tagger pipeline: RFTagger (LC pipeline v. 1)
Token attributes (columns): word, tag, lempos, gender_lemma
POS suffixes: ",noun,-n,adjective,-j,pronoun,-p,numeral,-m,verb,-v,adverb,-a,preposition,-s,conjunction,-c"
Tagset documentation: https://www.sketchengine.co.uk/tagset-reference-for-czech

ELEXIS Slovenian Web 2020

Corpus size in tokens: 1218262261
Time span of crawling: 2020-07 to 2020-12
Tokenization, lemmatization and morphological tagger pipeline: TreeTagger (LC pipeline v. 2)
Token attributes (columns): word, tag, lempos, gender_lemma
POS suffixes: ",noun,-s,verb,-g,adjective,-p,adverb,-r,pronoun,-z,adposition,-d,conjunction,-v,particle,-l,interjection,-m,numeral,-k,abbreviation,-o,residual,-n"
Tagset documentation: https://www.sketchengine.co.uk/slovene-tagset-multext-east-v4/

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.

ELEXIS Swedish Web 2020

Corpus size in tokens: 1162253496
Time span of crawling: 2020-08
Tokenization, lemmatization and morphological tagger pipeline: HunPOS (LC pipeline v. 4)
Token attributes (columns): word, tag, lempos, gender_lemma
POS suffixes: ",noun,-n,verb,-v,adjective,-a,pronoun,-p,determiner,-d,adverb,-r,preposition,-s,conjunction,-c,numeral,-m,interjection,-i,particle,-q"
Tagset documentation: https://www.sketchengine.co.uk/swedish-part-of-speech-tagset/

Semantic annotation

- Approx. 2 million token samples of each corpus were semantically annotated.
- The annotation for each corpus document is stored in a separate file in JSON format.
- Each file contains a list of sentences in the corresponding document.
- There is a list of annotations for each sentence in the document, each item in this list is an annotation of a token.
- The annotation of a token includes disambiguation with the BabelNet synset, WordNet offset, and WordNet id according to NLTK.
- Only content word tokens are annotated.

Availability

1. ELEXIS Cloud - https://cloud.elex.is
2. Sketch Engine - https://app.sketchengine.eu

Access conditions

All corpora are available under the conditions specified by the Lexical Computing Web Corpus Agreement as given in Annex 1.

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.

Annex 1: Lexical Computng Web Corpus Agreement

Lexical Computing CZ s.r.o.,
Botanická 68a, 602 00 Brno, Czech Republic
VAT: CZ29295491
(further as “LC” or “Lexical Computing”)

Lexical Computing Web Corpus Agreement

to use the Lexical Computing data set.
(further as “**Agreement**”)

I, _____, am a person engaging in research and development of natural-language-processing, information-retrieval or document-understanding systems
Official mail address _____

_____

Telephone _____

E-mail _____

I hereby agree to use the collection designated as _____ data set collected by LC (the "**Collection**"). By signing this Agreement I hereby agree to abide by the following understandings, terms and conditions. These understandings, terms and conditions apply equally to all or to part of the Collection, including any updates or new versions of the Collection supplied under this Agreement.

Copyright

1. The Collection has been obtained by crawling the Internet. Due to the size of the Collection it was not practicable to obtain permission from copyright owners to provide the Collection for the uses permitted under this Agreement (“Permitted Uses”).
2. I understand that all the documents in the Collection are documents which were at some time made publicly available on the Internet and which were collected using a process which respects the commonly accepted methods (such as robots.txt) for indicating that the documents should not be so collected.
3. Owners of copyright in individual documents may choose to request deletion of these documents from the Collection.
4. The limitation on permitted use contained in the following section is intended to reduce the risk of any action being brought by copyright owners, but if this happens I agree to bear all associated liability.

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.

Permitted Uses

1. The Collection may only be used for research and development of natural-language processing, information-retrieval or document-understanding systems.
2. Summaries, analyses and interpretations of the linguistic properties of the Collection may be derived and published, provided it is not possible to reconstruct the Collection from these summaries.
3. Small excerpts of the Collection may be displayed to others or published in a scientific or technical context, solely for the purpose of describing the research and development carried out and related issues.
4. All efforts must be made not to infringe the rights of any third party including, but not limited to, the authors and publishers of any excerpts used in accordance with clause 3 above in this “Permitted Uses” section.
5. I must make sure that I only display the Collection to or share the Collection with persons who also signed this Agreement with LC.

Agreement to Delete Data on Request

I undertake to delete within thirty days of receiving notice all copies of any nominated document
that is part of the Collection whenever requested to do so by either:

1. LC; or
2. the owner of copyright for the particular document.

No Warranty

The Collection is provided "as is", without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose and noninfringement. In no event shall LC be liable for any claim, damages or other liability, whether in an action of contract, tort or otherwise, arising in any way of the use of the Collection.

Termination

Either LC or me may terminate this Agreement at any time by notifying the other party in writing. On termination of the Agreement I shall destroy all copies of the Collection.

Applicable Law

This Agreement is governed by the laws of the Czech Republic.

I hereby execute this Agreement in favour and for the benefit of LC.

By the Individual:

Signature _____

Date _____

Name (please print) _____

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.