Digital Dictionary Database and ELEXIS Dictionary Matrix

Simon Krek

Jožef Stefan Institute, Artificial Intelligence Laboratory

University of Ljubljana, Centre for Language Resources and Technologies

Slovenia

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

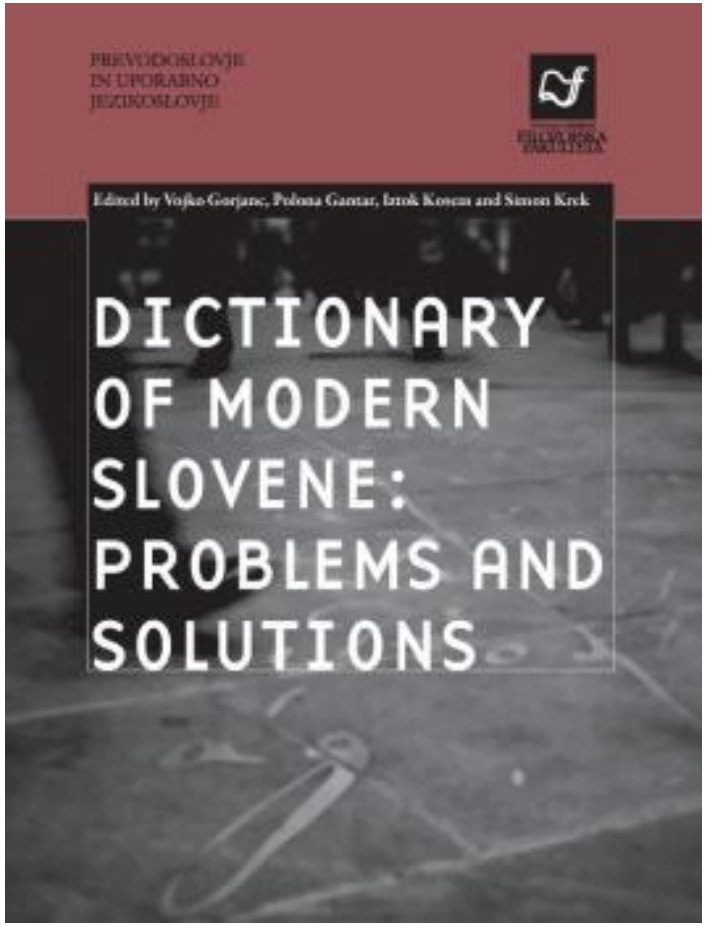
Topics (lexicography)

- Digital Dictionary Database (for Slovene)
  - Idea / Purpose
  - Automatic extraction of lexicographic data (from corpora)
  - Data Model
- ELEXIS (European Lexicographic Infrastructure)
  - (Dictionary) Sense Linking
  - Word Sense Disambiguation
  - Dictionary Matrix
- Universal Concepts (need for & possibility of)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Digital Dictionary Database (for Slovene) - history

- Slovene Lexical Database
  - Communication in Slovene (2008-2013)
  - Lexicographic data for human users and NLP
- Dictionary of Modern Slovene (proposal)
  - May 2013: http://www.sssj.si/
  - Sociolinguistics: https://www.simonkrek.si/blog/
- Dictionary of Modern Slovene: Problems and Solutions (monograph)
  - August 2017: publication link
- Digital Dictionary Database (June 2020)
  - Development of Slovene in Digital Environment
  - WP3: Semantic Resources and Technologies



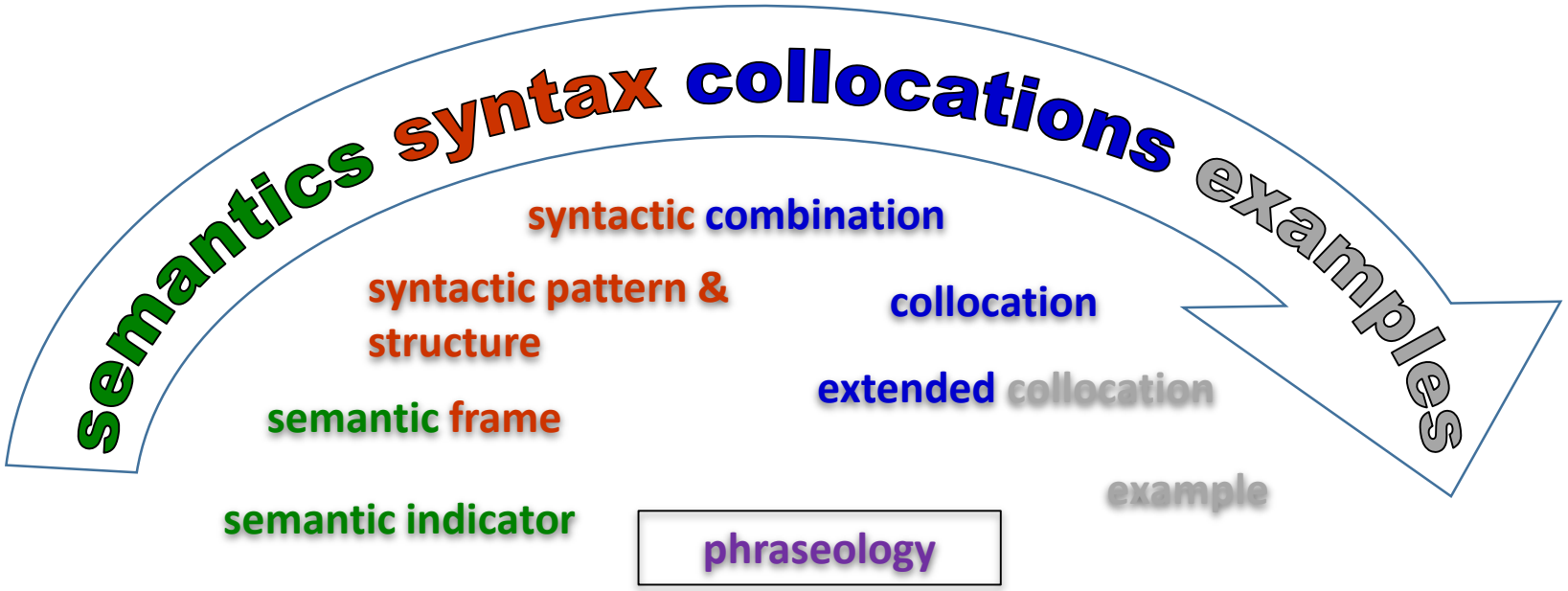
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Slovene Lexical Database (2013)

I. LEMMA	• **headword**	***svitati se (to dawn)***	
	• **part-of-speech**	verb	
II. SENSE	• **indicator**	***1. daniti se (day)***	***2. dojemati (understand)***
	• **semantic frame**	ko se svita **DAN.** začne vzhajati sonce	če se **ČLOVEKU** začne svitati o nekem **DOGAJANJU.** začne dojemati. kar prej ni vedel. ali pa je bilo to pred njim skrito
III. SYNTAX	• **label**	only in 3rd pers.	
	• **structure**	**gbz Inf-GBZ**	**rbz GBZ**
	• **pattern**	*kaj se svita (sth is dawning)*	*komu se svita o čem (sth is dawning to sb about sth)*
	• **synt. combin.**		
IV. COLLOC.	• **collocation**	[začeti. pričeti] se svitati	[počasi. malo. malce] se svita
V. EXAMPLES	• **example**	*Preden se začne zjutraj svitati. je najtemnejša noč.*	*Počasi se mi je začelo svitati. zakaj Jasni oči tako žarijo.*
		Na vzhodu se je že svital dan. ko sta se poslovila.	*Petru se pričenja svitati o nekdanji zvezi ned Chadom in Heather.*
	• **multi-word unit**		
VI. PHRASEOLOGY	• **phraseological units**		

This project has received funding from the Eu

SLD - from semantics to corpus examples

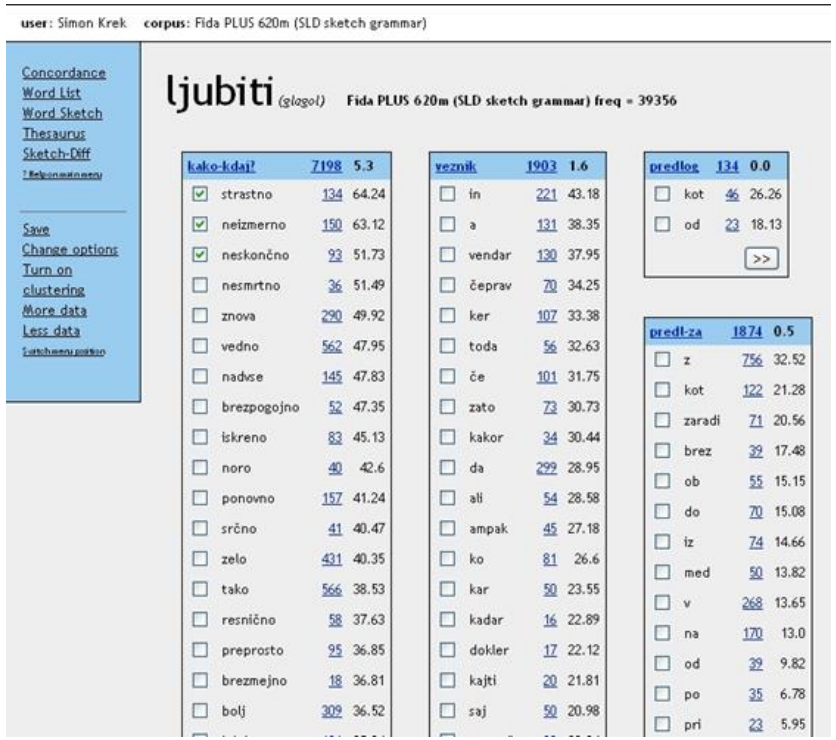


Book (2015) - **Polona Gantar**: Lexicographic Description of Slovene in Digital Environment
Data: Slovene lexical database 1.0 (CLARIN.SI repository)
Online: http://eng.slovenscina.eu/spletni-slovar

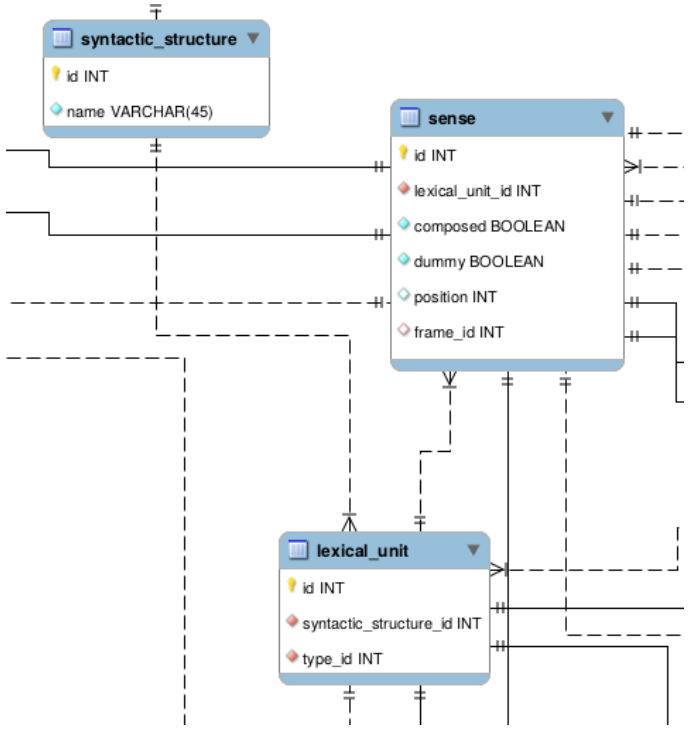
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Two aspects of subsequent work

Automation (of lexicographic work)



Data model (for lexicographic data)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

I. LEMMA

- headword — *svitati se (to dawn)*
- part-of-speech — verb

II. SENSE

- indicator — *1. daniti se (day)* — *2. dojemati (understand)*
- semantic — a **DAN.** ajati sonce — če se **ČLOVEKU** začne svitati o nekem **DOGAJANJU.** začne dojemati. kar prej ni vedel. ali pa je bilo to pred njim skrito

unary relations & constructions

gramrels

III. SYNTA

- lable — only in 3rd pers.
- structure — **gbz Inf-GBZ** — **rbz GBZ**
- pattern — *kaj se svita (sth is dawning)* — *komu se svita o čem (sth is dawning to sb about sth)*

word sketches

- synt. combin.
- collocation — [začeti. pričeti] se svitati — [počasi. malo. malce] se svita

GDEX

V. EXAMPLES

- example — *Preden se začne zjutraj svitati. je najtemnejša noč.* — *Počasi se mi je začelo svitati. zakaj Jasni oči tako žarijo.*

Na vzhodu se je že svital dan. ko sta se poslovila. — *Petru se pričenja svitati o nekdanji zvezi ned Chadom in Heather.*

- multi-word unit

VI. PHRASEOLOGY

- phraseological units

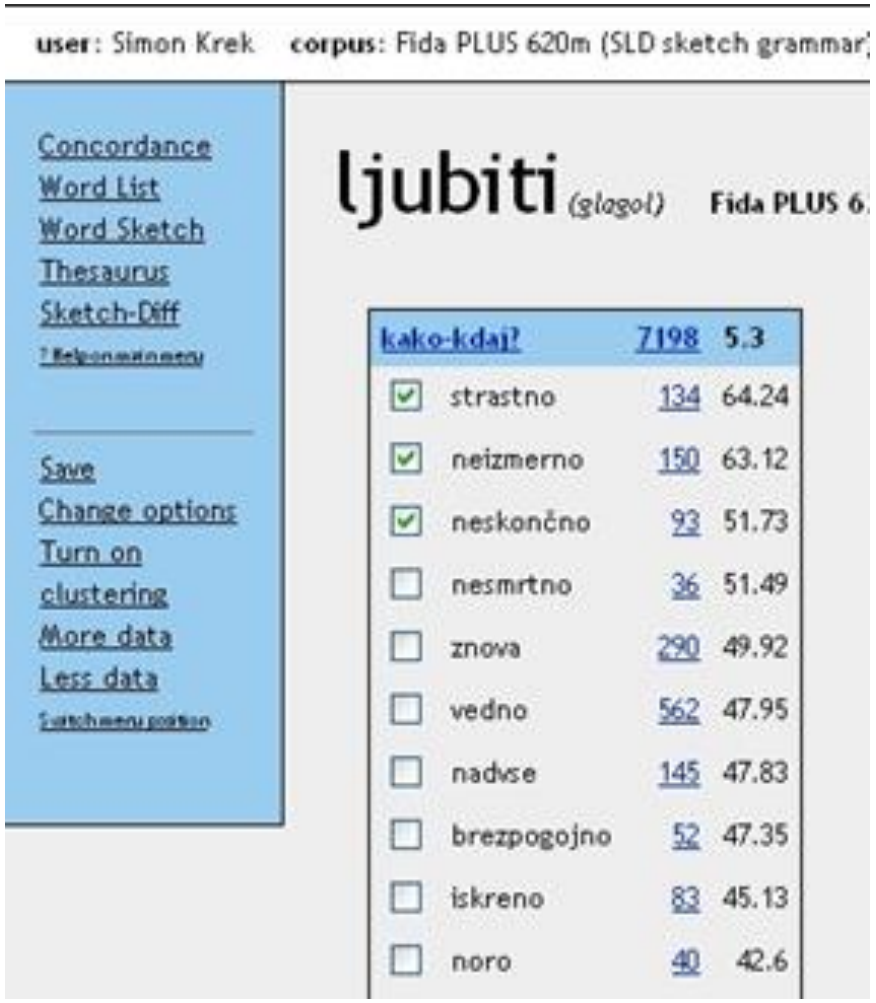
This project has

Sketch Engine: sketch grammar

- regular expressions over POS tags

```
=a_modifier/modifies
    2:[tag="P.*"] 1:[tag="S.*"]
```

- the name of the arguments (order)
- 1: 2: = words to be extracted as the first/second argument
- |, ., (), {} and * - standard metacharacters (RE)



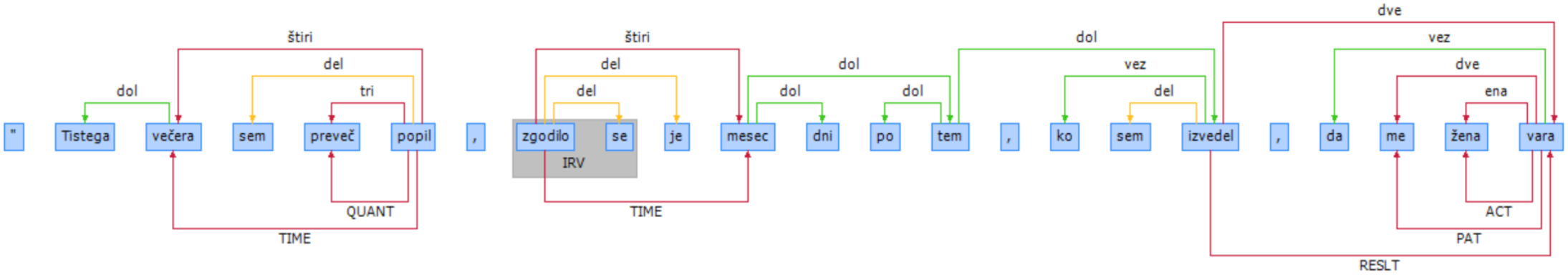
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

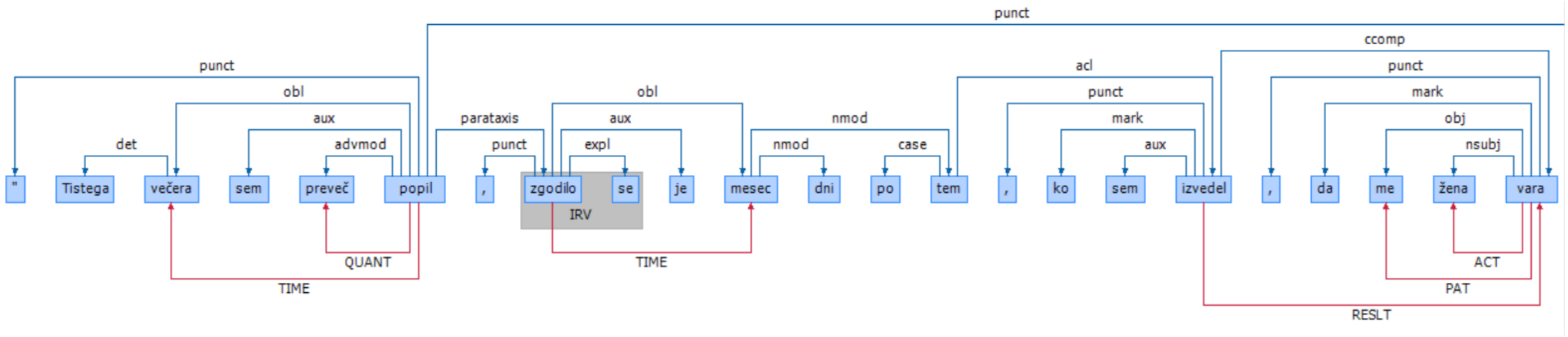
New formalism

- Two basic research projects (Slovene Research Agency)
  - New Grammar of Standard Slovene: Resources and Methods
    - WP 2: identification and extraction of collocations
  - Collocations as a Basis for Language Description: Semantic and Temporal Perspectives
- Method
  - Focused on dependency annotation layer
  - Combining restrictions & representation
    - using corpus annotations on morphological and syntactic levels

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

JOS dependency <-> Universal Dependecies





This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Grammatical formalism for the description of collocations

Example: **V + N_{ACC}**

absorbirati snov = ‘to absorb substance‘

≠ ‘the substance absorbs‘

```
<system type="JOS">
<components>
    <component cid="1" type="core" name="gbz"/>
    <component cid="2" type="core" name="ref" status="optional"/>
    <component cid="3" type="other" status="forbidden"/>
    <component cid="4" type="core" name="sbz4"/>
</components>
<dependencies>
    <dependency from="#" to="1" label="#"/>
    <dependency from="1" to="2" label="del"/>
    <dependency from="4" to="3" label="dol"/>
    <dependency from="1" to="4" label="dve"/>
</dependencies>
```

```
<definition>
    <component cid="1">
     <restriction type="morphology">
      <feature POS="verb"/>
      <feature type="main"/>
     </restriction>
     <representation>
      <feature rendition="lemma"/>
     </representation>
    </component>
    <component cid="2">
     <restriction type="morphology">
      <feature POS="pronoun"/>
      <feature type="reflexive"/>
     </restriction>
     <representation>
      <feature rendition="word_form"/>
      <feature selection="all"/>
     </representation>
    </component>
...
```

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.

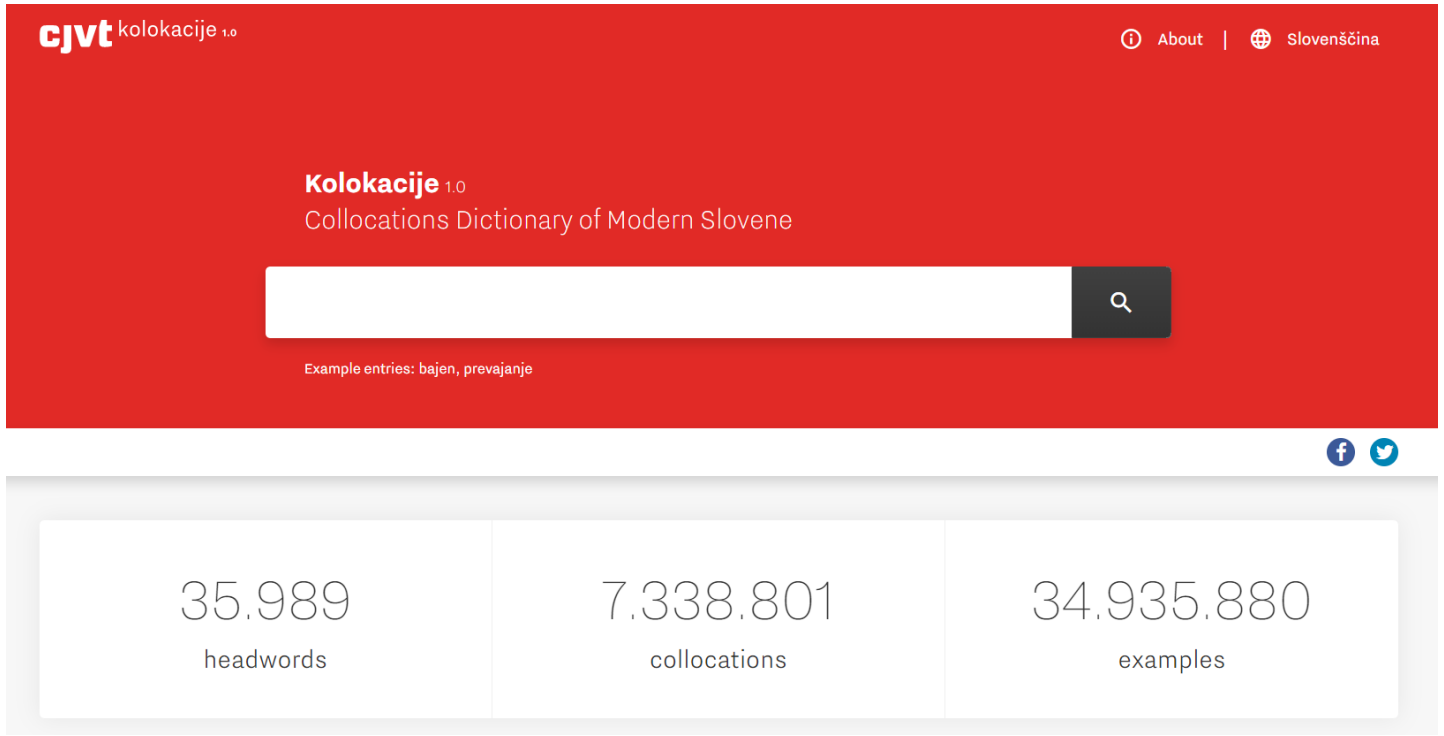
Improvements

- Better at identifying difficult relations (subject and object)
  - **substanca + absorbirati** ($N_{Nominative}$ + V) substance absorbs x
  - **absorbirati + substance** (V + $N_{Acusative}$) x absorbs substance
- Rendering typical forms of elements in collocations
  - finančna težava → finančne težave (financial + trouble → financial trubles)
  - stresti bonbon → stresti bonbone (drop + candy → drop candies)
  - dobra možnost → dobre možnosti (good chance → better chances)
- Enabling inclusion of all levels of annotation into the game
  - “Extended” collocations
  - From patterns (valency, frames + semantic types), to collocations (excluding phraseology or MWEs)

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.

elexis european lexicographic infrastructure

Collocations Dictionary of Modern Slovene

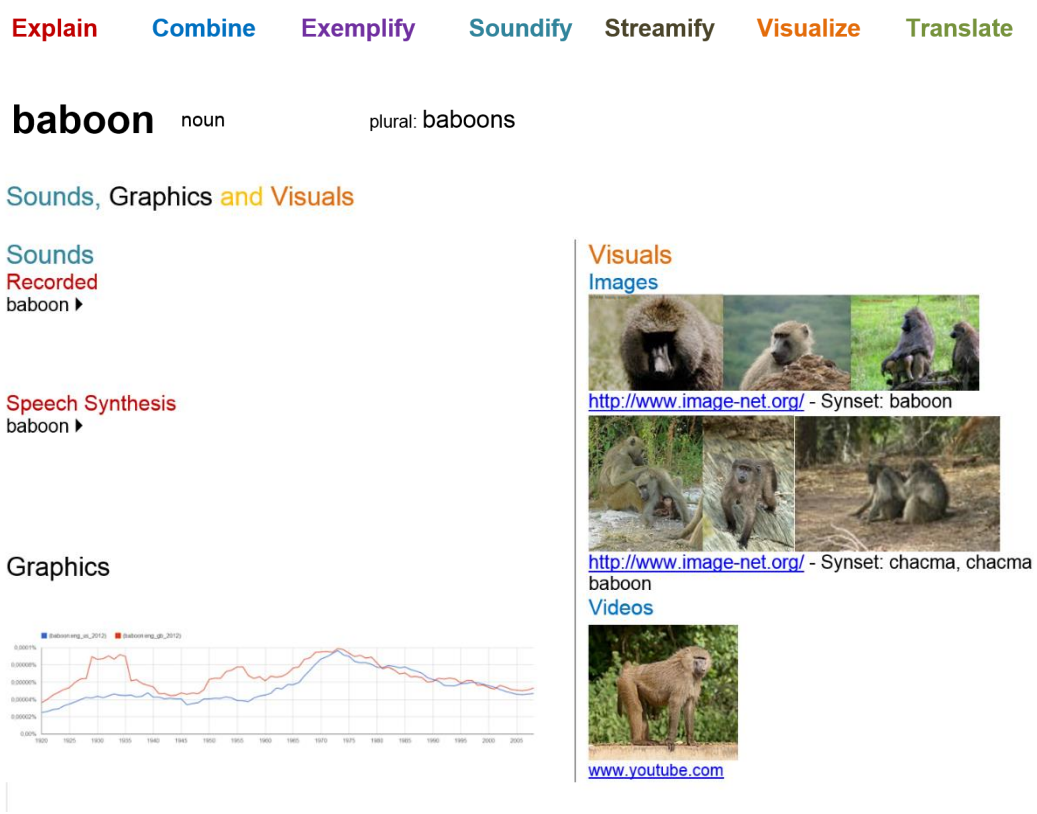


- Online: https://viri.cjvt.si/kolokacije/eng/
- Data: http://hdl.handle.net/11356/1250

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

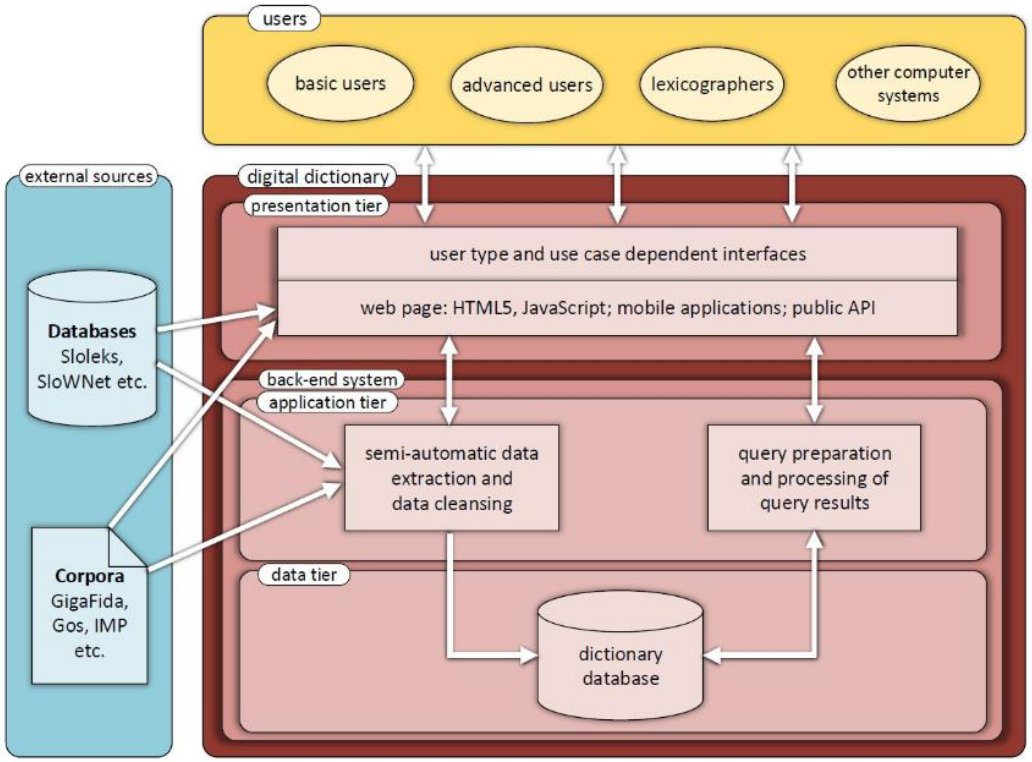
Automation on all levels of description

- Headword/Lemma list + POS
- Syntactic structures & Collocations
- Valency & Semantic Roles
- Sense
  - Word Sense Disambiguation
  - Word Sense Induction
  - Definition Extraction
  - Labels (domain, register etc)
- Phraseology
- Sounds, images, video etc.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Data model (for all types of lexicographic data)



- Dictionary of Modern Slovene: Problems and Solutions (monograph)
- Bojan Klemenc, Marko Robnik-Šikonja, Luka Fürst, Ciril Bohak and Simon Krek: Technological Design of a State-of-the-art Digital Dictionary

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

elexis european lexicographic infrastructure

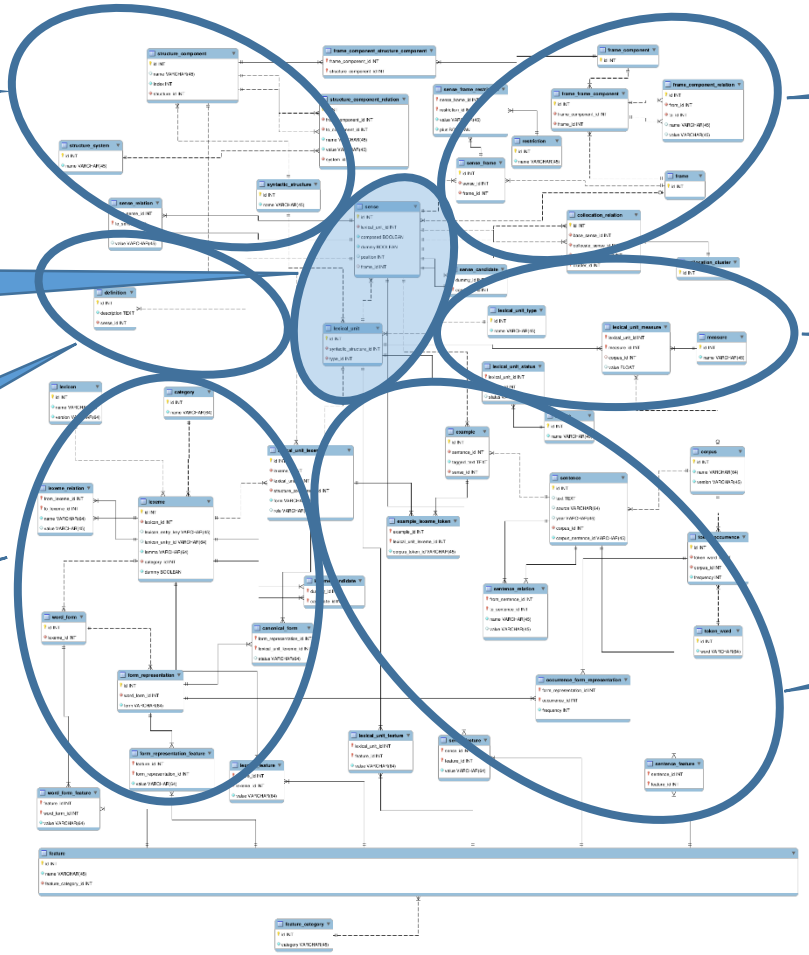
Digital Dictionary Database (for Slovene)

SYNTACTIC STRUCTURES

LEMMA-SENSE = CONCEPT

DEFINITIONS TRANSLATIONS

MORPHOLOGY PRONUNCIATION



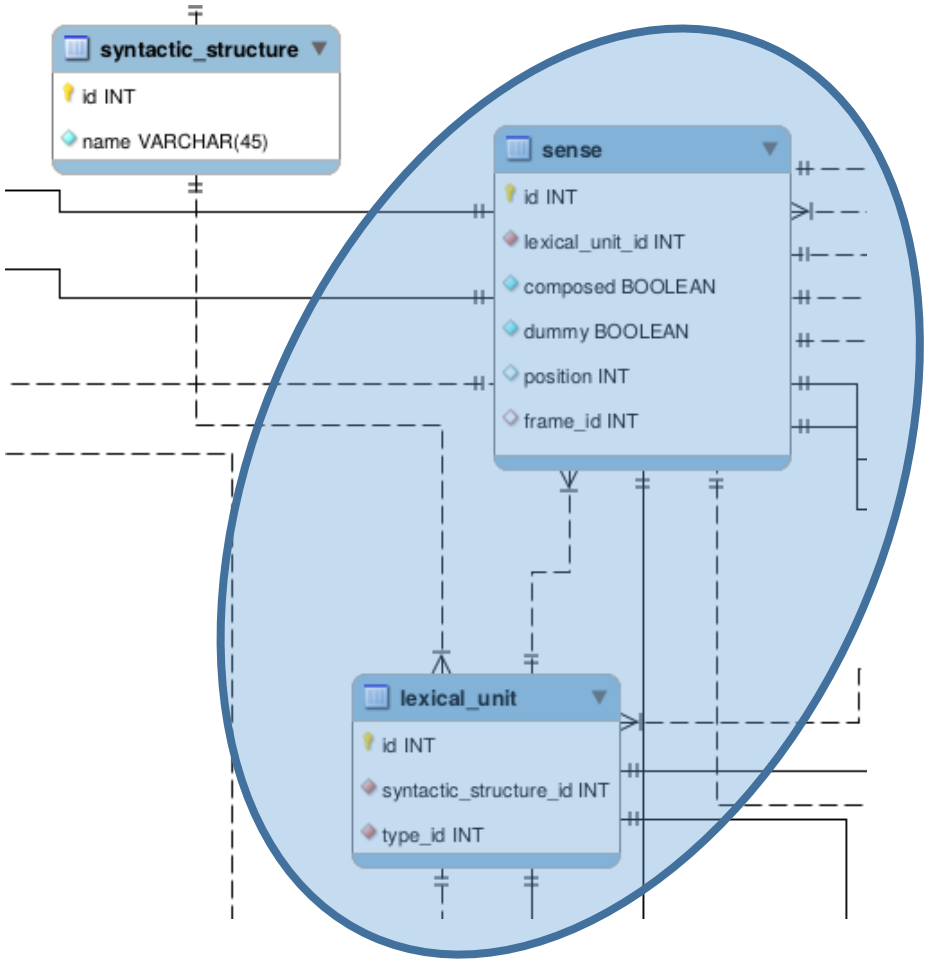
SEMANTIC FRAMES / TYPES

COLLOCATIONS

ATTESTATIONS CORPUS DATA

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Lexical unit + Sense = Concept (stable id)



- Lexicographic Data as a Service (LDaaS)
- Available through API service
- Open Access (CC BY-SA)
- In CLARIN.SI repository

- Development of Slovene Language in Digital Environment (2020-2022)

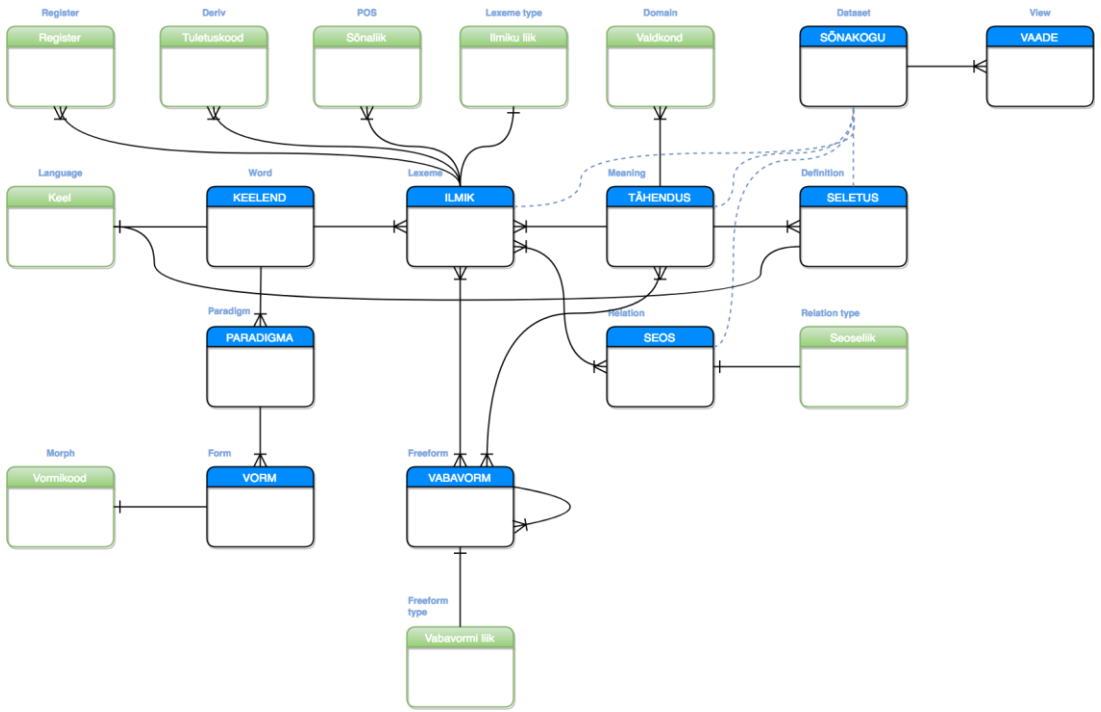
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Why? Artificial Intelligence

Denmark's National Strategy for Artificial Intelligence (2019)

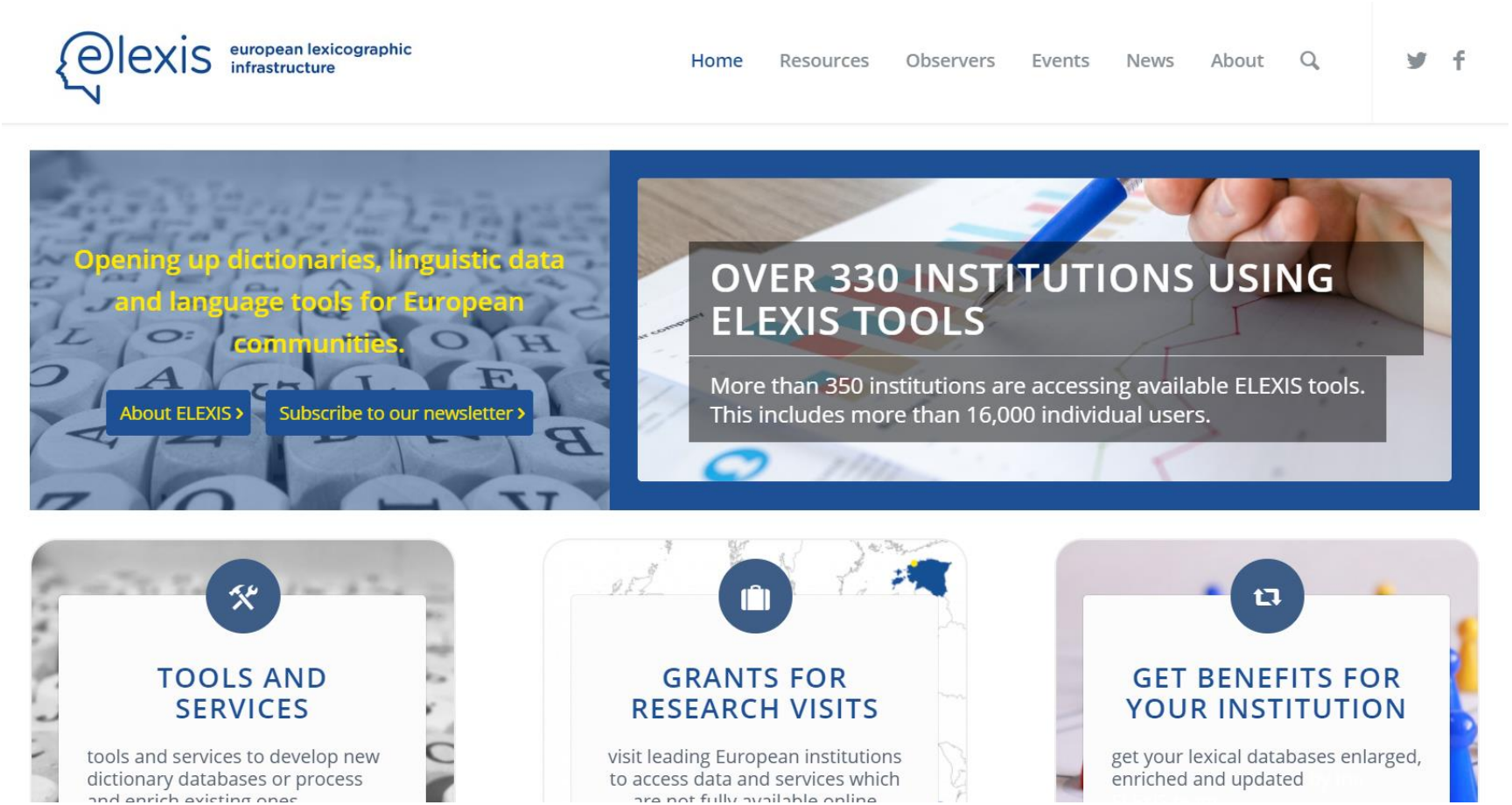
- A common Danish language resource will be established to support and accelerate the development of language-technology solutions in Danish. The language resource will be freely available, enabling suppliers to build on existing knowledge to create new solutions within voice recognition and language understanding to benefit citizens, authorities and businesses.

Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX (2019)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

ELEXIS (European Lexicographic Infrastructure)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

elexis european lexicographic infrastructure

ELEXIS FACT SHEET

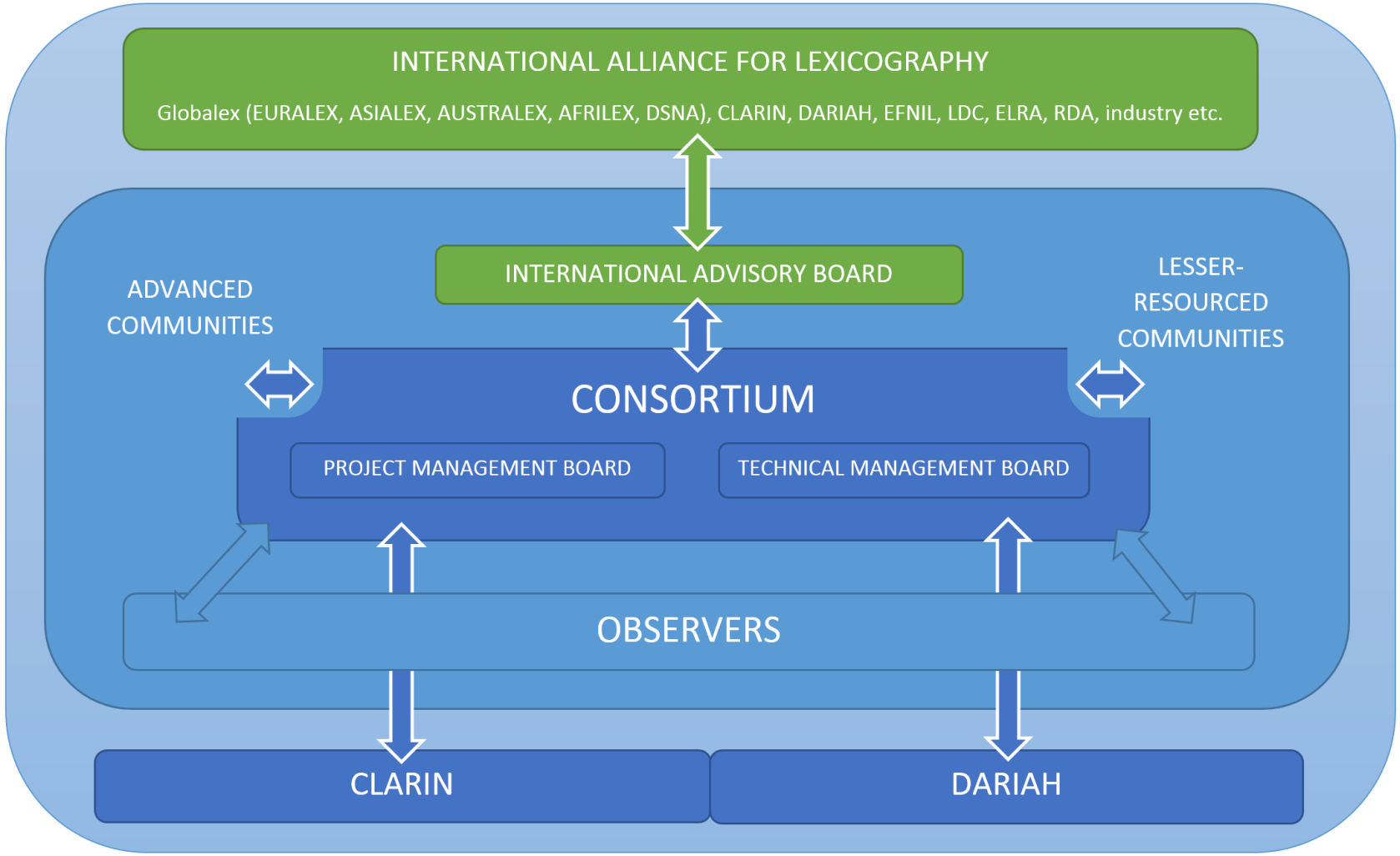
- Call & Topic: INFRAIA-02-2017 (Infrastructures)
  - Integrating Activities for **Starting Communities**
- Start date: 1 February 2018
- Duration: 48 months (31 January 2022) – PROBABLY EXTENDED
- Total cost: 5M €
- Coordinator: Jožef Stefan Institute, Ljubljana, Slovenia
- Number of **partners**: **17** from **15** countries
- Number of **observers**: **45** institutions from **34** countries
- Web site: www.elex.is

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

elexis european lexicographic infrastructure

ELEXIS

European Lexicographic Infrastructure



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

ELEXIS project summary (1)

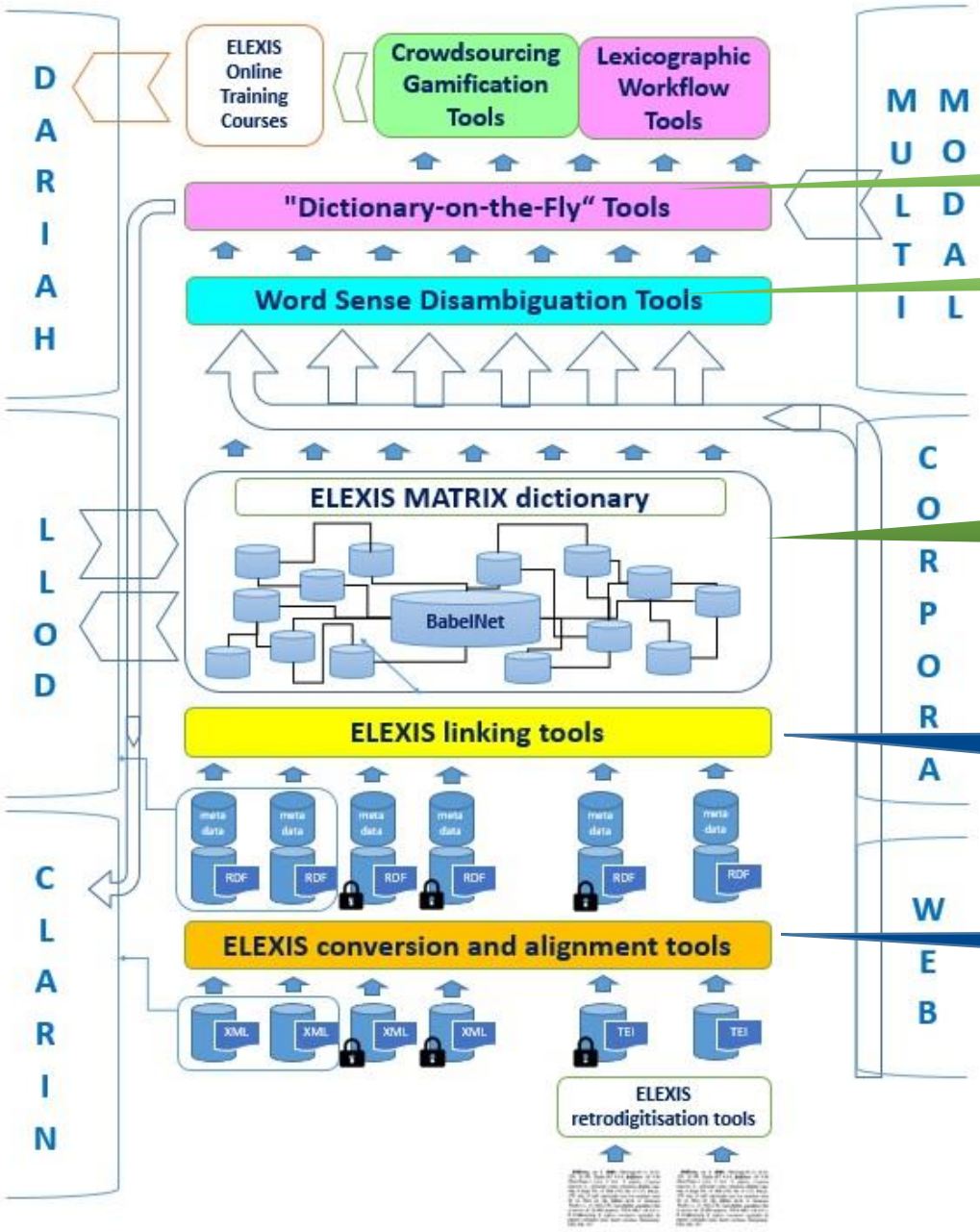
- The project proposes to integrate, extend and harmonise **national** and **regional efforts** in the field of lexicography, both modern and historical, with the goal of creating a sustainable infrastructure which will
- (1) enable **efficient access to high quality lexical data** in the digital age, and
- (2) bridge the gap between more advanced and lesser-resourced scholarly communities working on lexicographic resources.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

ELEXIS project summary (2)

- Current lexicographic resources, both modern and historical, have different levels of **structuring** and are not equally suitable for application in other fields, e.g. **Natural Language Processing**.
- The project will develop strategies, **tools** and **standards** for extracting, structuring and linking lexicographic resources to unlock their full potential for **Linked Open Data** and the **Semantic Web**, as well as in the context of digital humanities.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.



elexis european lexicographic infrastructure

SKE One-click Dictionary

BN VerbAtlas / SyntagNet

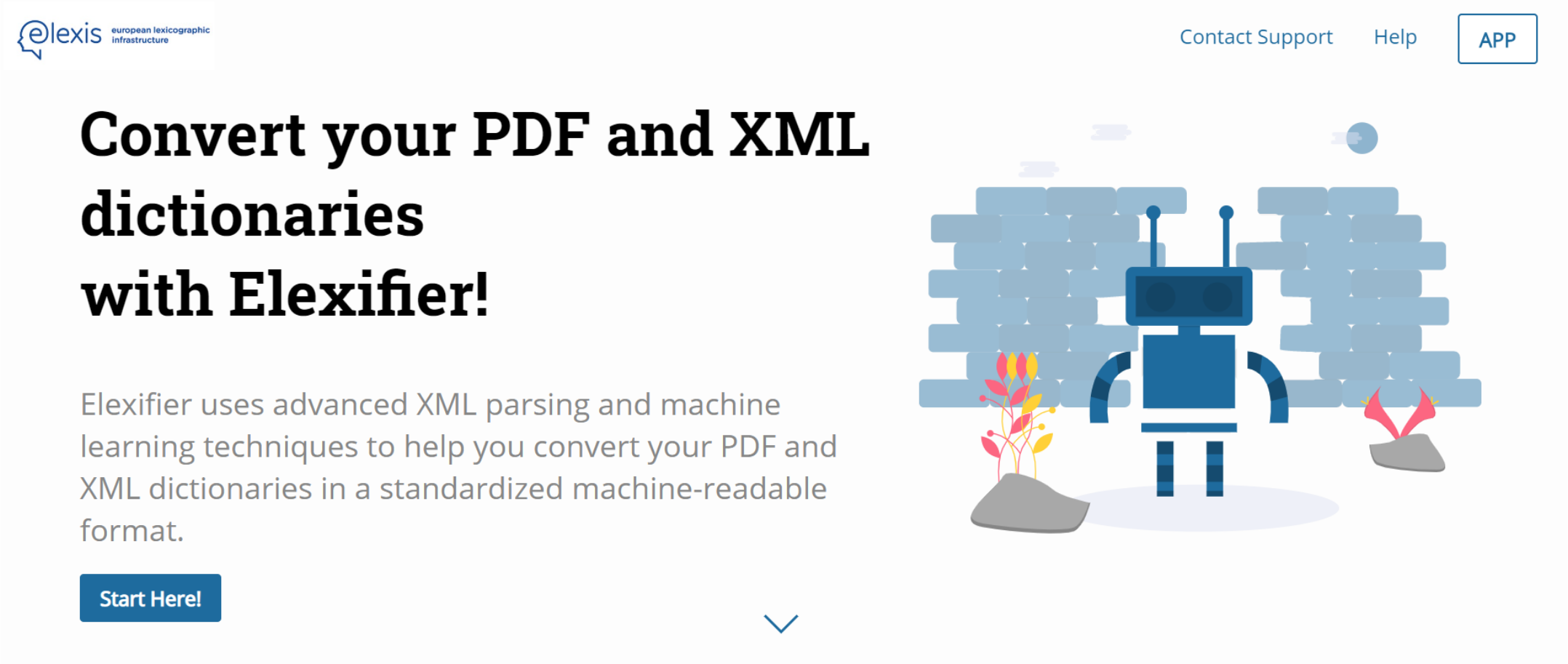
DICTIONARY MATRIX / MATRIX DICTIONARY

MANUAL: LEXONOMY
AUTOMATIC: NAISC

ELEXIFIER.ELEX.IS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

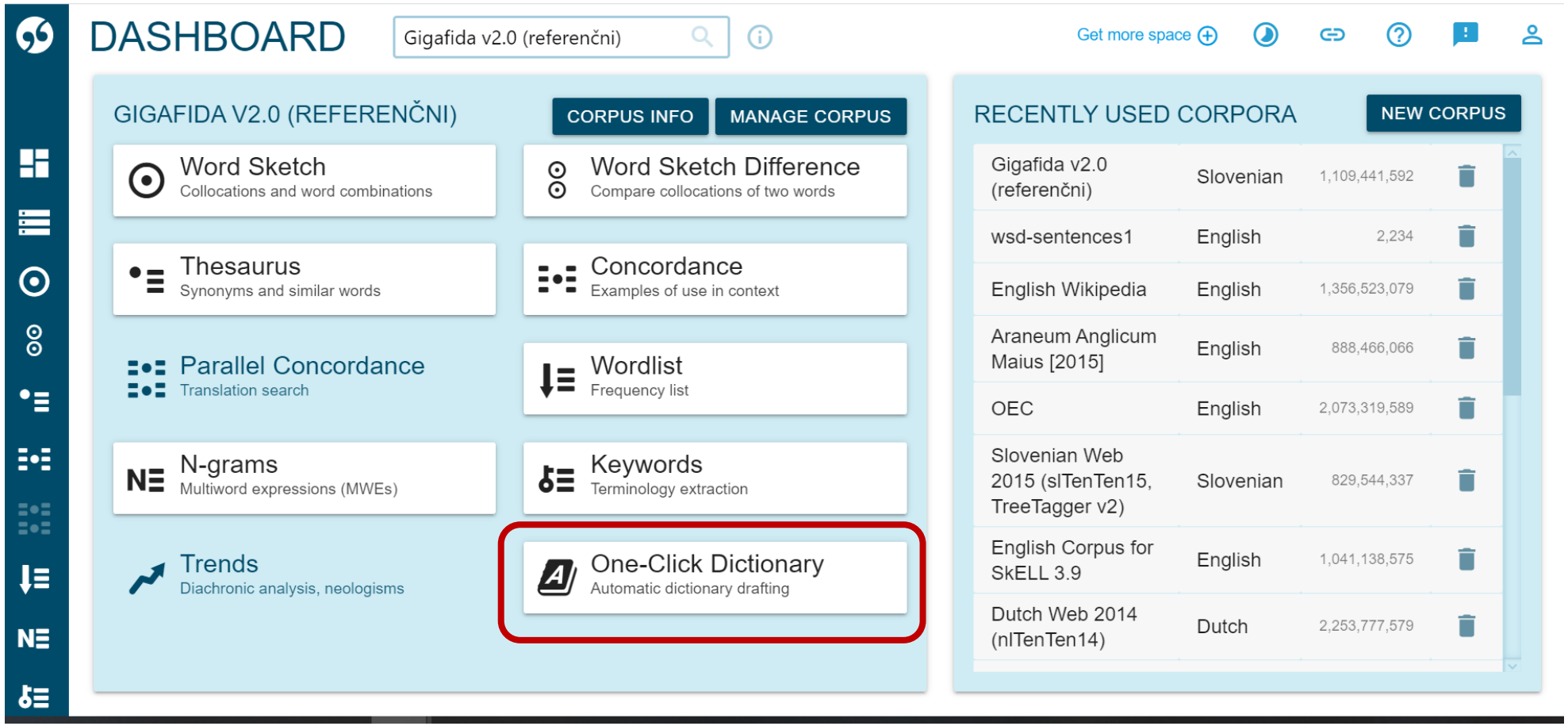
Conversion tool (https://github.com/elexis-eu/elexifier)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

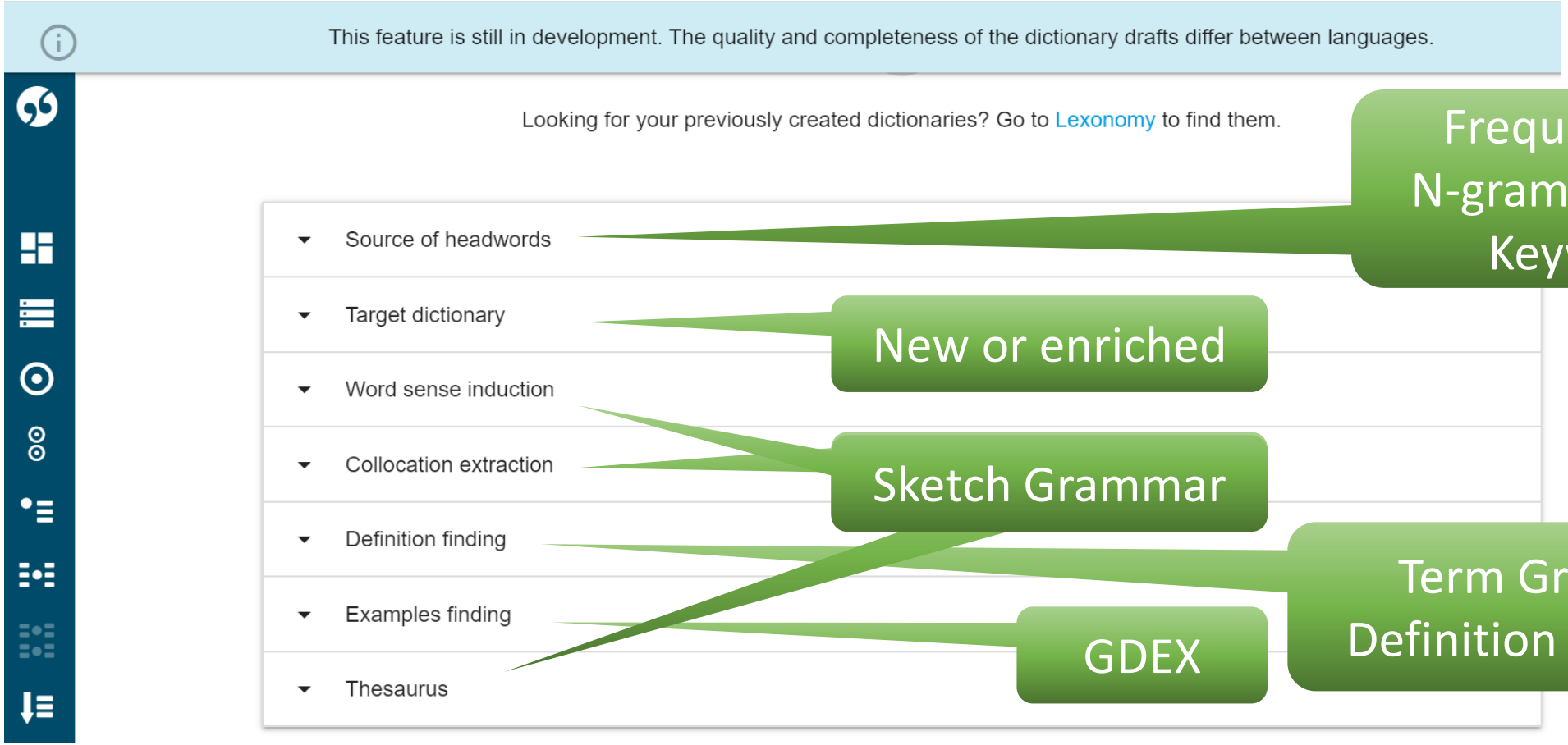
elexis european lexicographic infrastructure

“Dictionary-on-the-fly”



This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.

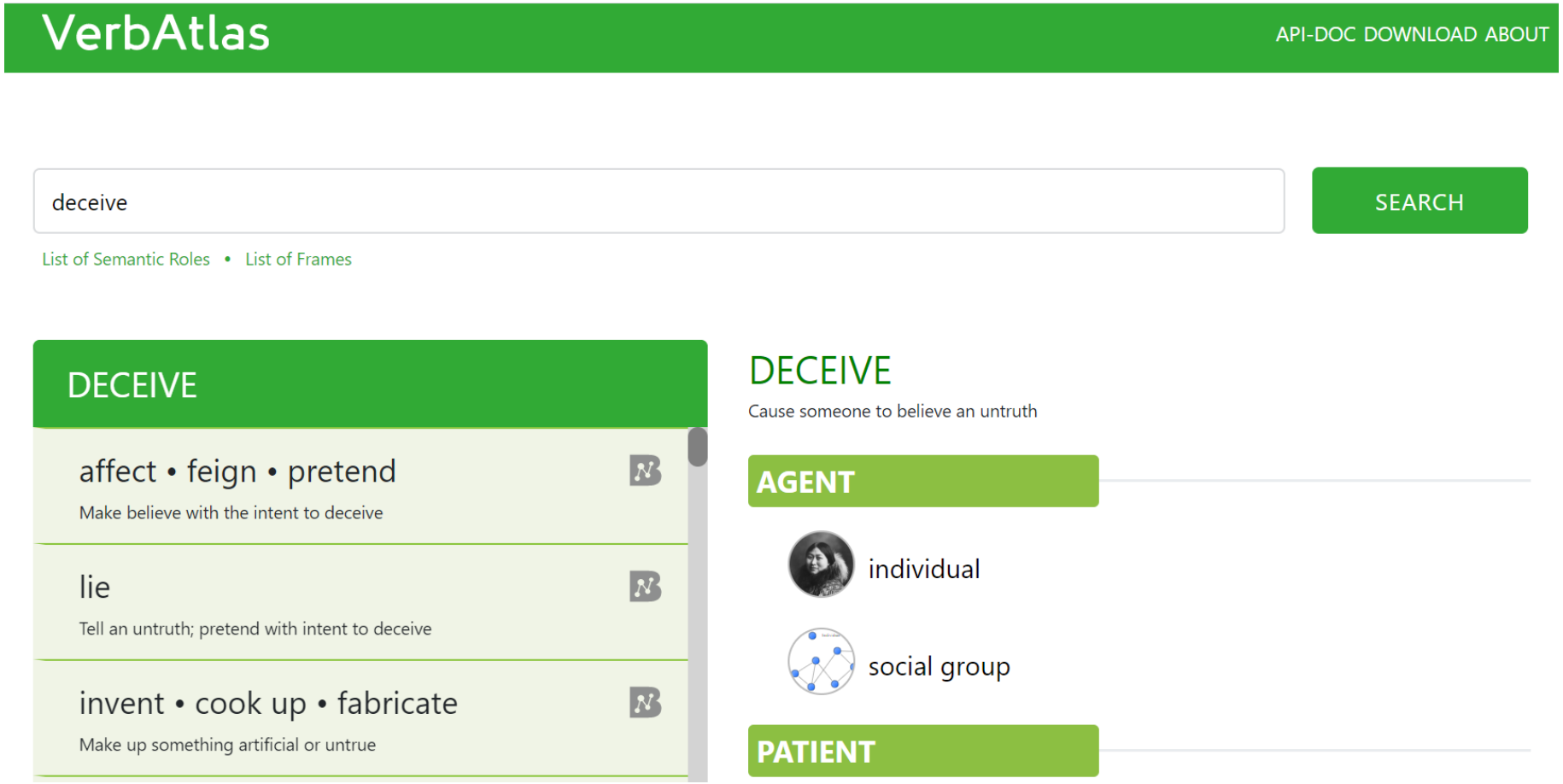
One click dictionary



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

elexis european lexicographic infrastructure

VerbAtlas – Semantic Roles & Frames



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

elexis european lexicographic infrastructure

SyntagNet – Word Sense Disambiguation

We must learn to live together as brothers or perish together as fools.

English

DISAMBIGUATE

View:

Concept Named Entities

We must | learn | to | live | together | as | brothers | or | perish

VERB

learn

Gain knowledge or skills

WordNet (license)

VERB

live

Inhabit or live in; be an inhabitant of

WordNet (license)

ADV

together

In each other's company

WordNet (license)

NOUN

brother

A male with the same parents as someone else

WordNet (license)

VERB

perish

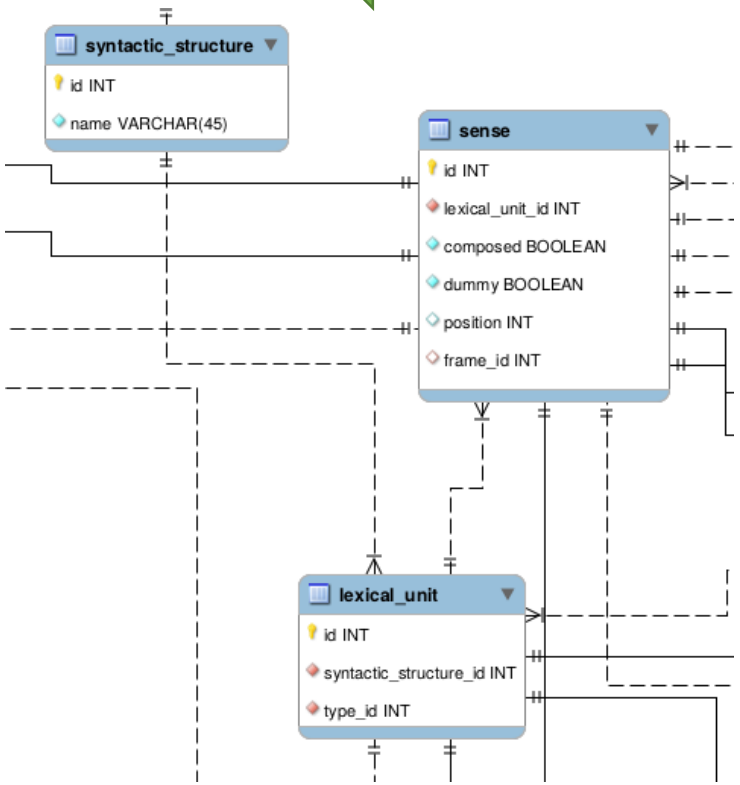
Pass from physical life and lose all bodily attributes and functions necessary to sustain life

WordNet (license)

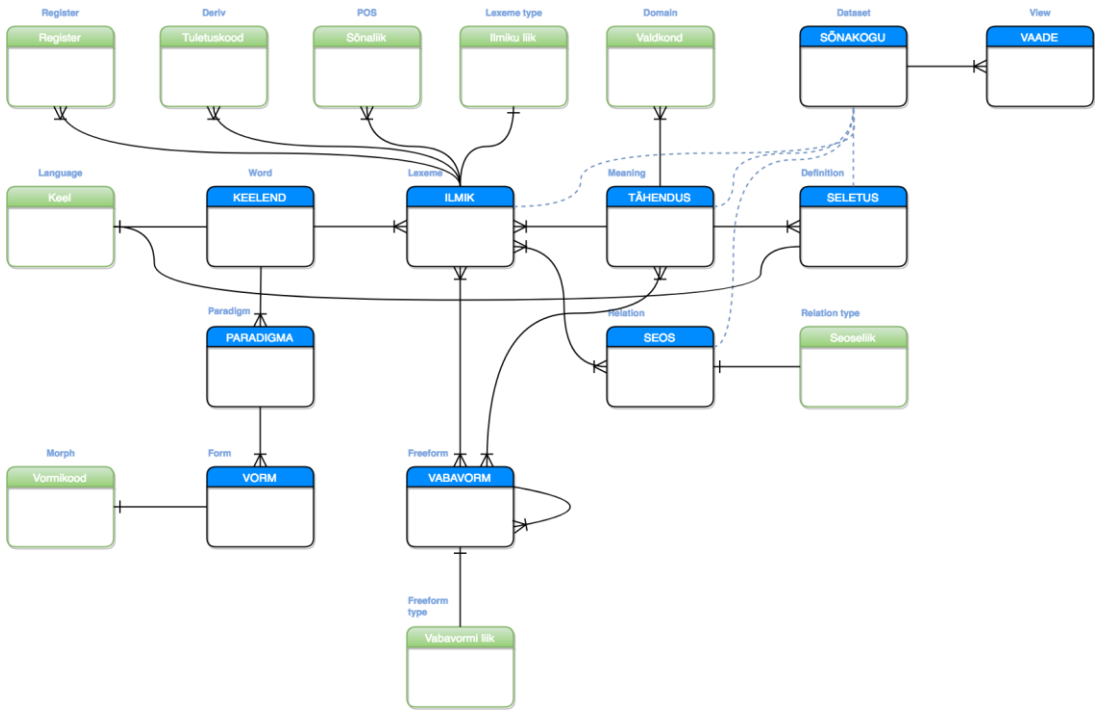
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Dictionary Matrix – the cross-lingual challenge

Slovenian



Estonian



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Globalex 2020 Workshop on Linked Lexicography

- **ELEXIS Monolingual Word Sense Alignment Task**
- Monolingual word sense alignment is a challenging task of finding matching senses between two dictionary entries and will play a crucial role in the development of new lexical resources.
- Training data – monolingual dictionaries: Basque, Bulgarian, Danish, Dutch, English, Estonian, German, Hungarian, Irish, Italian, Portuguese, Serbian, Slovenian, Spanish, Russian
  - Competition: https://competitions.codalab.org/competitions/22163
  - Proceedings: https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/GLOBALEX2020book.pdf

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Task: predicting the relationship between two senses

- Five categories
  - **Exact**: The sense are the same, for example the definitions are simply paraphrases
  - **Broader**: The sense in the first dictionary completely covers the meaning of the sense in the second dictionary and is applicable to further meanings
  - **Narrower**: The sense in the first dictionary is entirely covered by the sense of the second dictionary, which is applicable to further meanings
  - **Related**: There are cases when the senses may be equal but the definitions in both dictionaries differ in key aspects
  - **None**: There is no match for this sense

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Dictionary Matrix

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Dictionary / semantic network

- BabelNet (CC BY-**NC**-SA)
  - is both a **multilingual encyclopedic dictionary**, with lexicographic and encyclopedic coverage of terms, and a **semantic network** which connects concepts and named entities in a very large network of semantic relations, made up of about 15 million entries, called Babel synsets.
- ConceptNet (CC BY-SA)
  - is a freely-available **semantic network**, designed to help computers understand the meanings of words that people use. ConceptNet aims to give computers access to common-sense knowledge, the kind of information that ordinary people know but usually leave unstated.
  - BabelNet is very similar in structure to ConceptNet, but very different in openness.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Universal Concepts

WIKIDATA

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

Tools

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

Print/export

Download as PDF

Lexeme | Discussion

(L1) ama mis-x-Q36790 mis-x-Q401

Language Sumerian
Lexical category noun

Statements

image	 Ama-gi.svg 872 × 222; 17 KB ▾ 0 references
described by source	Pennsylvania Sumerian Dictionary ▸ 1 reference

L1-S1

Czech	matka
Cantonese	媽
Serbian	мајка
Korean	어머니
Chinese	妈
Hindi	माँ
Telugu	తల్లి
Catalan	mare
Georgian	დედა
Kazakh	ана
Azerbaijani	ana
Cherokee	ᎠᏂᏥ
Swedish	mor
Slovak	matka
Lithuanian	motė
Latvian	māte

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Thank you

Simon Krek

Jožef Stefan Institute, Artificial Intelligence Laboratory

University of Ljubljana, Centre for Language Resources and Technologies

Slovenia

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.