



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



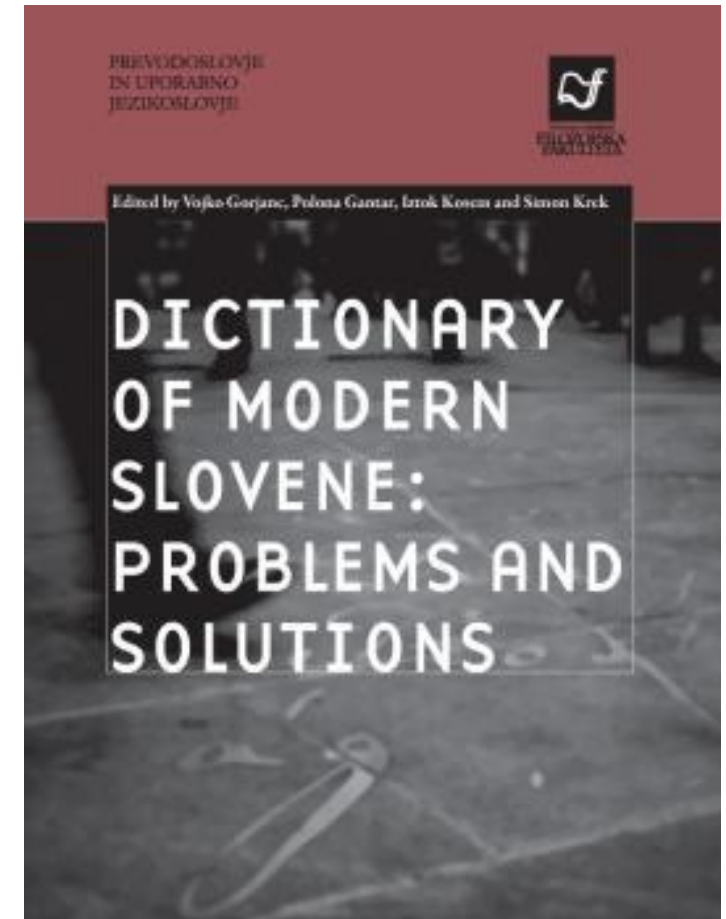
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

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)

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



Slovene Lexical Database (2013)

I. LEMMA

- headword
- part-of-speech

svitati se (to dawn)

verb

II. SENSE

- indicator
- semantic frame

1. daniti se (day)

ko se svita **DAN**.
začne vzhajati sonce

2. dojemati (understand)

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
- structure
- pattern
- synt. combin.

only in 3rd pers.

gbz Inf-GBZ

kaj se svita
(sth is dawning)

rbz GBZ

komu se svita o čem
(sth is dawning to sb about sth)

IV. COLLOC.

- collocation [začeti. pričeti] se svitati

[počasi. malo. malce] se svita

V. EXAMPLES

- example Preden se začne zjutraj
svitati. je najtemnejša noč.

Na vzhodu se je že svital
dan. ko sta se poslovila.

Počasi se mi je začelo svitati.
zakaj Jasni oči tako žarijo.

Petru se pričenja svitati o nekdanji
zvezi ned Chadom in Heather.

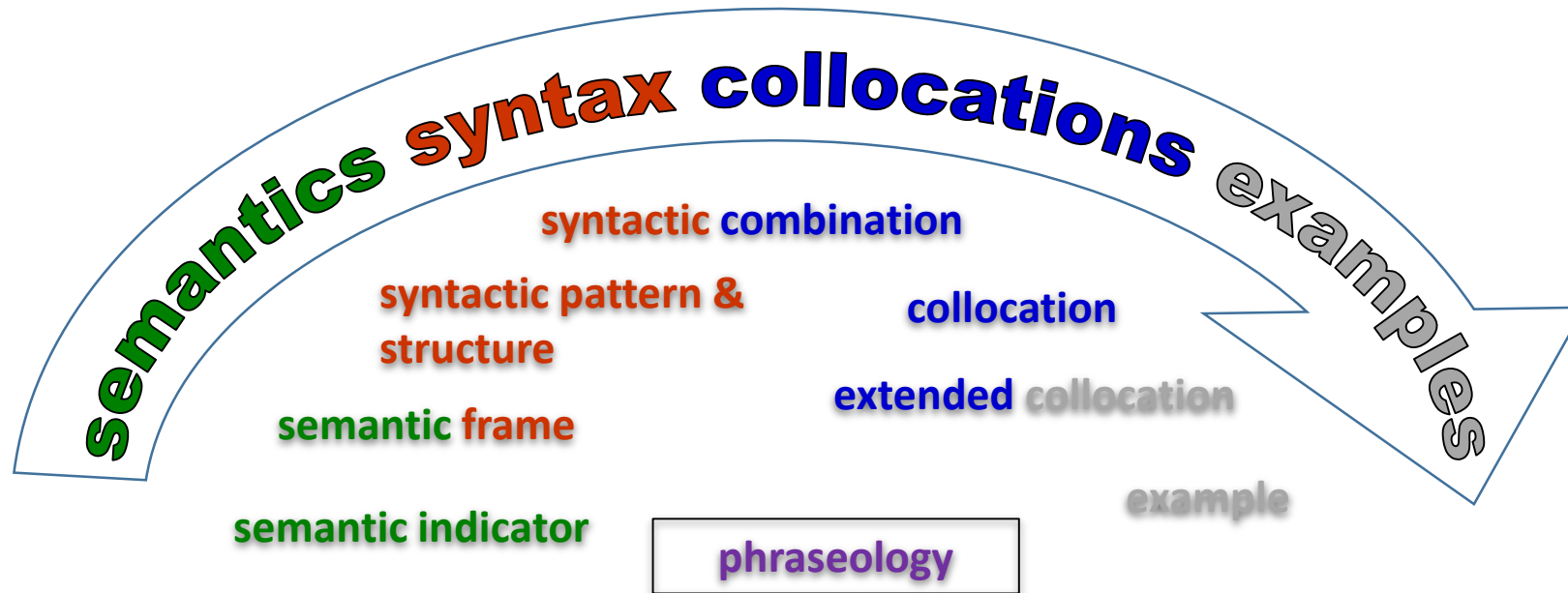
- multi-word unit

VI. PHRASEOLOGY

- phraseological units



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>

Two aspects of subsequent work

Automation (of lexicographic work)

user: Simon Krek corpus: Fida PLUS 620m (SLD sketch grammar)

Concordance
Word List
Word Sketch
Thesaurus
Sketch-Diff
Help on sketch-diff

Save
Change options
Turn on clustering
More data
Less data
Turn off clustering

ljubiti (gлагол) Fida PLUS 620m (SLD sketch grammar) freq = 39356

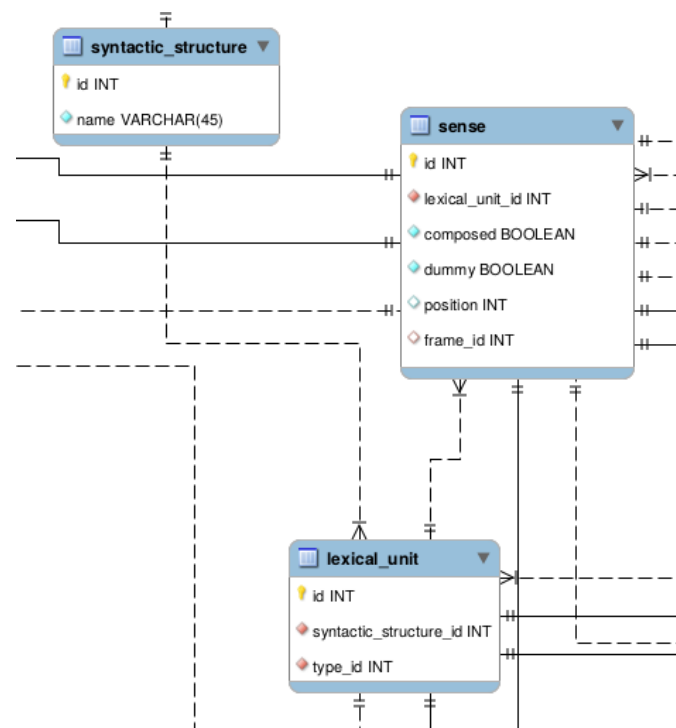
kako-kdaj?	7198	5.3
<input checked="" type="checkbox"/> strastno	134	64.24
<input checked="" type="checkbox"/> neizmerno	150	63.12
<input checked="" type="checkbox"/> neskončno	93	51.73
<input type="checkbox"/> nesmrtno	36	51.49
<input type="checkbox"/> znova	290	49.92
<input type="checkbox"/> vedno	562	47.95
<input type="checkbox"/> nadse	145	47.83
<input type="checkbox"/> brezpogojno	52	47.35
<input type="checkbox"/> iskreno	83	45.13
<input type="checkbox"/> noro	40	42.6
<input type="checkbox"/> ponovno	157	41.24
<input type="checkbox"/> srčno	41	40.47
<input type="checkbox"/> zelo	431	40.35
<input type="checkbox"/> tako	566	38.53
<input type="checkbox"/> resnično	58	37.63
<input type="checkbox"/> preprosto	95	36.85
<input type="checkbox"/> brezmejno	18	36.81
<input type="checkbox"/> bolj	309	36.52

veznik	1903	1.6
<input type="checkbox"/> in	221	43.18
<input type="checkbox"/> a	131	38.35
<input type="checkbox"/> vendar	130	37.95
<input type="checkbox"/> čeprav	70	34.25
<input type="checkbox"/> ker	107	33.38
<input type="checkbox"/> toda	56	32.63
<input type="checkbox"/> če	101	31.75
<input type="checkbox"/> zato	73	30.73
<input type="checkbox"/> kakor	34	30.44
<input type="checkbox"/> da	299	28.95
<input type="checkbox"/> ali	54	28.58
<input type="checkbox"/> ampak	45	27.18
<input type="checkbox"/> ko	81	26.6
<input type="checkbox"/> kar	50	23.55
<input type="checkbox"/> kadar	16	22.89
<input type="checkbox"/> dokler	17	22.12
<input type="checkbox"/> kajti	20	21.81
<input type="checkbox"/> saj	50	20.98

predlog	134	0.0
<input type="checkbox"/> kot	46	26.26
<input type="checkbox"/> od	23	18.13

predi-za	1874	0.5
<input type="checkbox"/> z	756	32.52
<input type="checkbox"/> kot	122	21.28
<input type="checkbox"/> zaradi	71	20.56
<input type="checkbox"/> brez	39	17.48
<input type="checkbox"/> ob	55	15.15
<input type="checkbox"/> do	70	15.08
<input type="checkbox"/> iz	74	14.66
<input type="checkbox"/> med	50	13.82
<input type="checkbox"/> v	268	13.65
<input type="checkbox"/> na	170	13.0
<input type="checkbox"/> od	39	9.82
<input type="checkbox"/> po	35	6.78
<input type="checkbox"/> pri	23	5.95

Data model (for lexicographic data)



I. LEMMA

- headword
- part-of-speech

svitati se (to dawn)

verb

II. SENSE

- indicator

1. *daniti se (day)*

2. *dojemati (understand)*

- semantic

unary
relations &
constructions

na DAN.
hajati sonce

če se ČLOVEKU začne svitati o nekem
DOGAJANJU. začne dojemati. kar
prej ni vedel. ali pa je bilo to pred
njim skrito

gramrels

III. SYNTAX

word
sketches

GDEX

V. EXAMPLES

- lable

- structure
- pattern

- synt. combin.

- collocation

- example

only in 3rd pers.

gbz Inf-GBZ

kaj se svita
(sth is dawning)

[začeti. pričeti] se svitati

Preden se začne zjutraj
svitati. je najtemnejša noč.

Na vzhodu se je že svital
dan. ko sta se poslovila.

rbz GBZ

komu se svita o čem
(sth is dawning to sb about sth)

[počasi. malo. malce] se svita

Počasi se mi je začelo svitati.
zakaj Jasni oči tako žarijo.

Petru se pričinja svitati o nekdanji
zvezi ned Chadom in Heather.

- multi-word unit

VI. PHRASEOLOGY • phraseological units

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)

user: Simon Krek corpus: Fida PLUS 620m (SLD sketch grammar)

Concordance
 Word List
 Word Sketch
 Thesaurus
 Sketch-Diff
 Response history

Save
 Change options
 Turn on clustering
 More data
 Less data
 Switch menu position

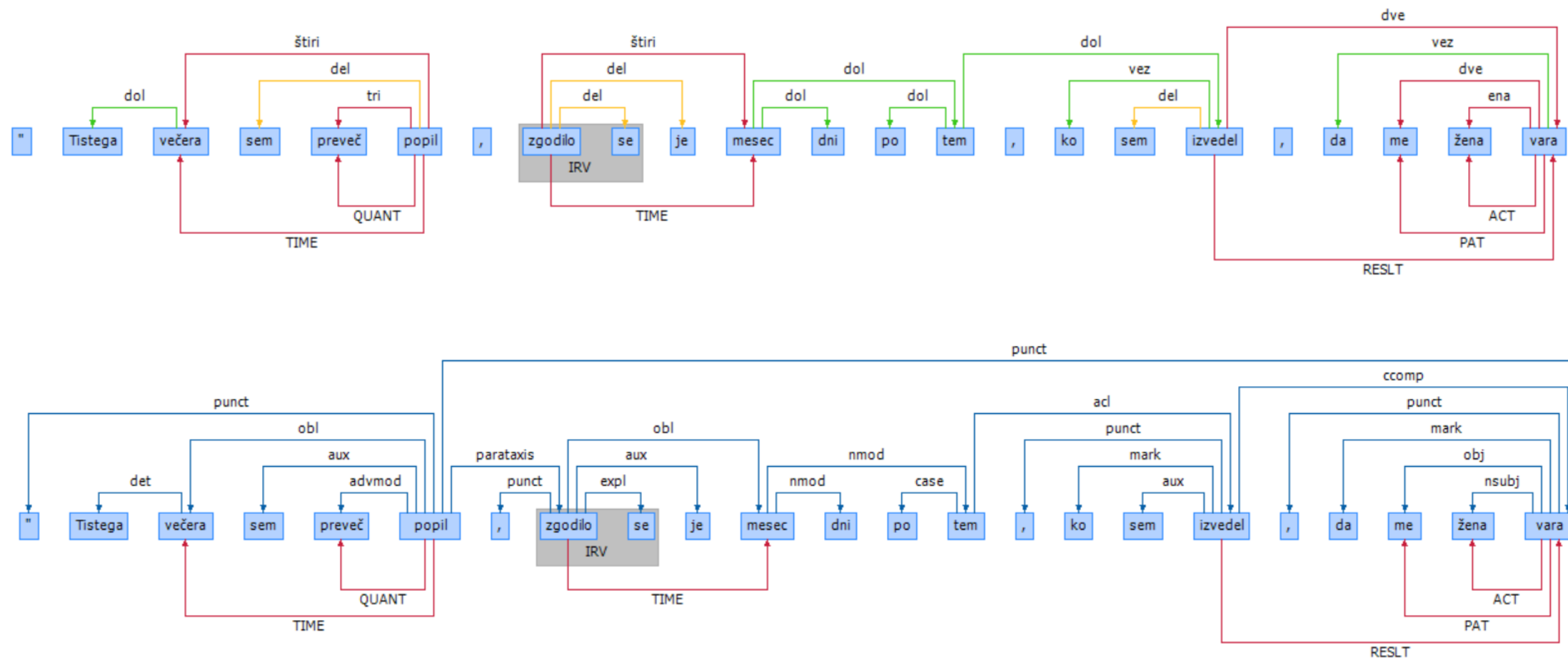
ljubiti (glagol) Fida PLUS 6

	kako-kdaj?	7198	5.3
<input checked="" type="checkbox"/>	strastno	134	64.24
<input checked="" type="checkbox"/>	neizmerno	150	63.12
<input checked="" type="checkbox"/>	neskončno	93	51.73
<input type="checkbox"/>	nesmrtno	36	51.49
<input type="checkbox"/>	znova	290	49.92
<input type="checkbox"/>	vedno	562	47.95
<input type="checkbox"/>	nad vse	145	47.83
<input type="checkbox"/>	brezpogojno	52	47.35
<input type="checkbox"/>	iskreno	83	45.13
<input type="checkbox"/>	noro	40	42.6

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

JOS dependency <-> Universal Dependencies



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>
```

Improvements

- Better at identifying difficult relations (subject and object)
 - **substanca + absorbirati** ($N_{\text{Nominative}} + V$) substance absorbs x
 - **absorbirati + substance** ($V + N_{\text{Acusative}}$) x absorbs substance
- Rendering typical forms of elements in collocations
 - finančna težava → finančne težave (financial + trouble → financial troubles)
 - 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)

Collocations Dictionary of Modern Slovene

The screenshot shows the homepage of the 'Kolokacije 1.0' (Collocations Dictionary of Modern Slovene) website. The header is red with the 'cjvt kolokacije 1.0' logo on the left and 'About | Slovenščina' links on the right. The main content area is also red and features the title 'Kolokacije 1.0' and 'Collocations Dictionary of Modern Slovene'. Below this is a large white search bar with a magnifying glass icon. Under the search bar, it says 'Example entries: bajen, prevajanje'. At the bottom of the red section are social media icons for Facebook and Twitter. Below the red section is a white box containing three statistics: 35.989 headwords, 7.338.801 collocations, and 34.935.880 examples.

Category	Count
headwords	35.989
collocations	7.338.801
examples	34.935.880

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

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.

Explain Combine Exemplify Soundify Streamify Visualize Translate

baboon noun plural: baboons

Sounds, Graphics and Visuals

Sounds
Recorded
baboon ▶

Speech Synthesis
baboon ▶

Graphics



Visuals
Images



<http://www.image-net.org/> - Synset: baboon



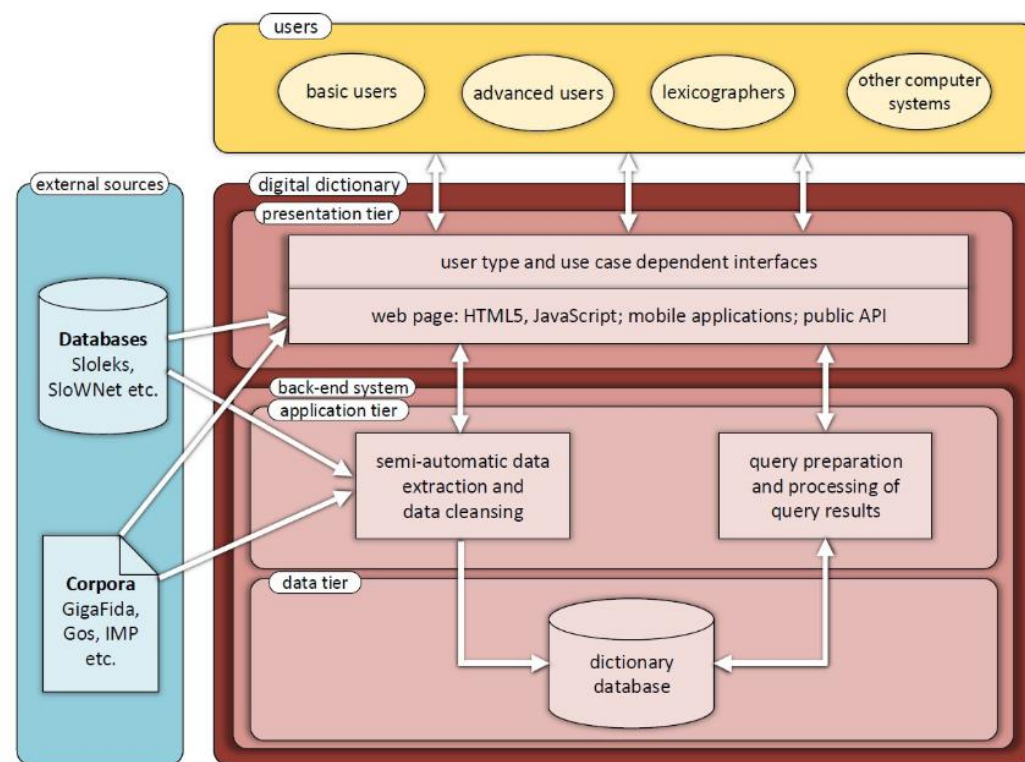
<http://www.image-net.org/> - Synset: chacma, chacma

Videos



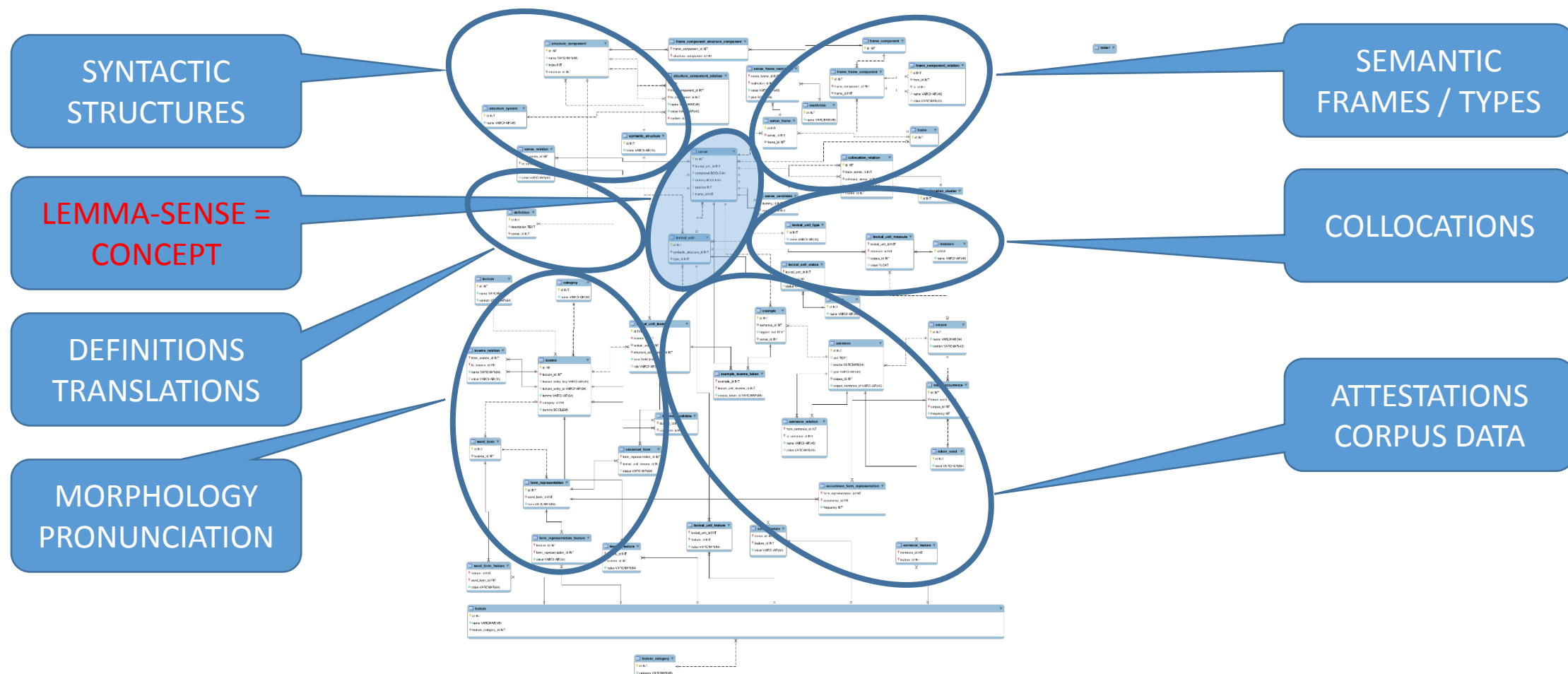
www.youtube.com

Data model (for all types of lexicographic data)

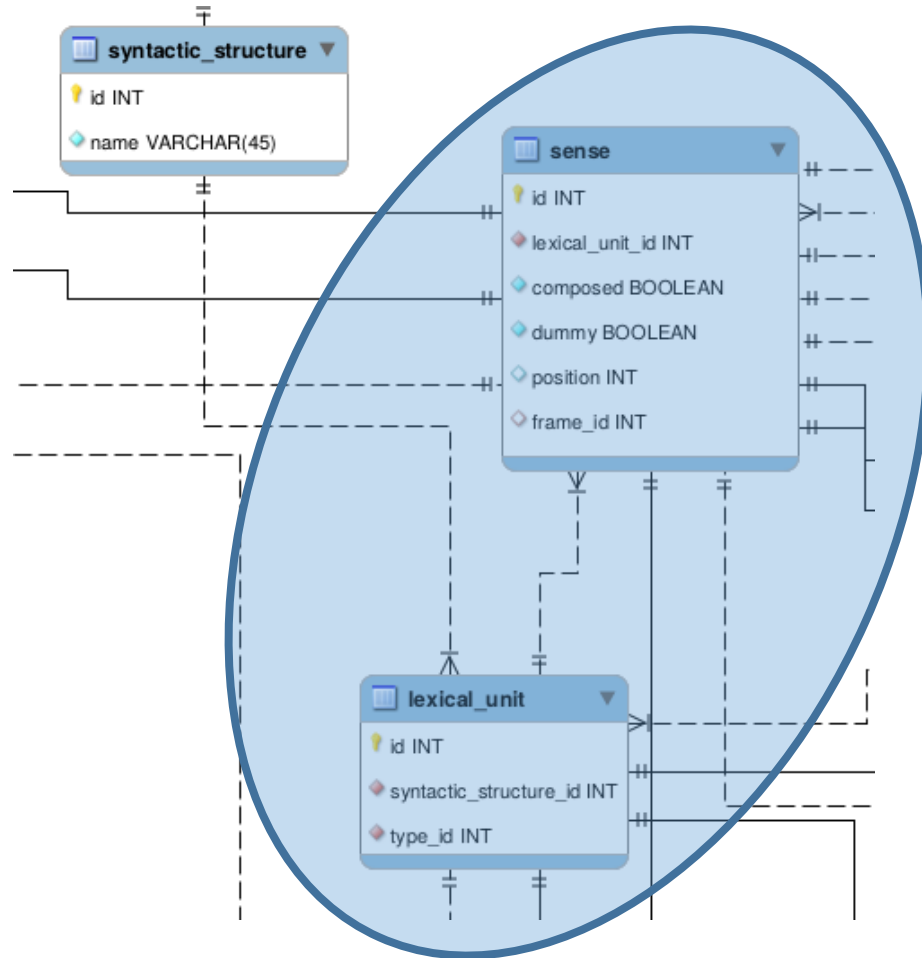


- Dictionary of Modern Slovene: Problems and Solutions (monograph)
- Bojan Klemenc, Marko Robnik-Šikonja, Luka Fürst, Ciril Bohak and Simon Krek: Technological Design of a State-of-the-art Digital Dictionary

Digital Dictionary Database (for Slovene)



Lexical unit + Sense = Concept (stable id)



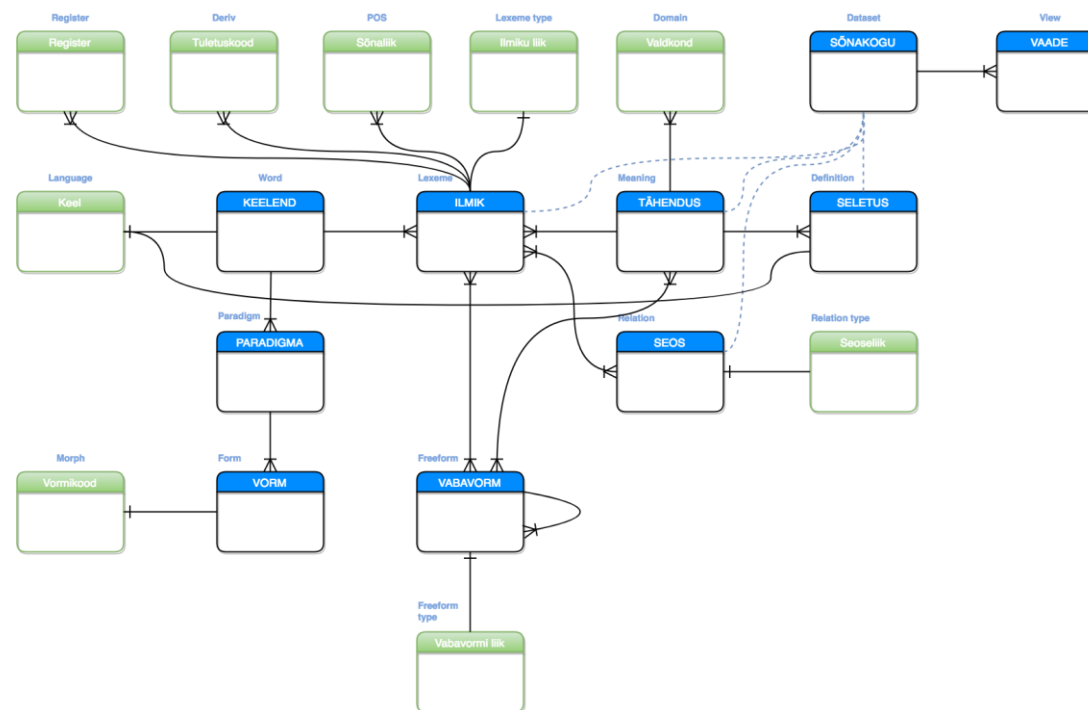
- Lexicographic Data as a Service (LDaaS)
- Available through API service
- Open Access (CC BY-SA)
- In CLARIN.SI repository
- Development of Slovene Language in Digital Environment (2020-2022)

Why? Artificial Intelligence

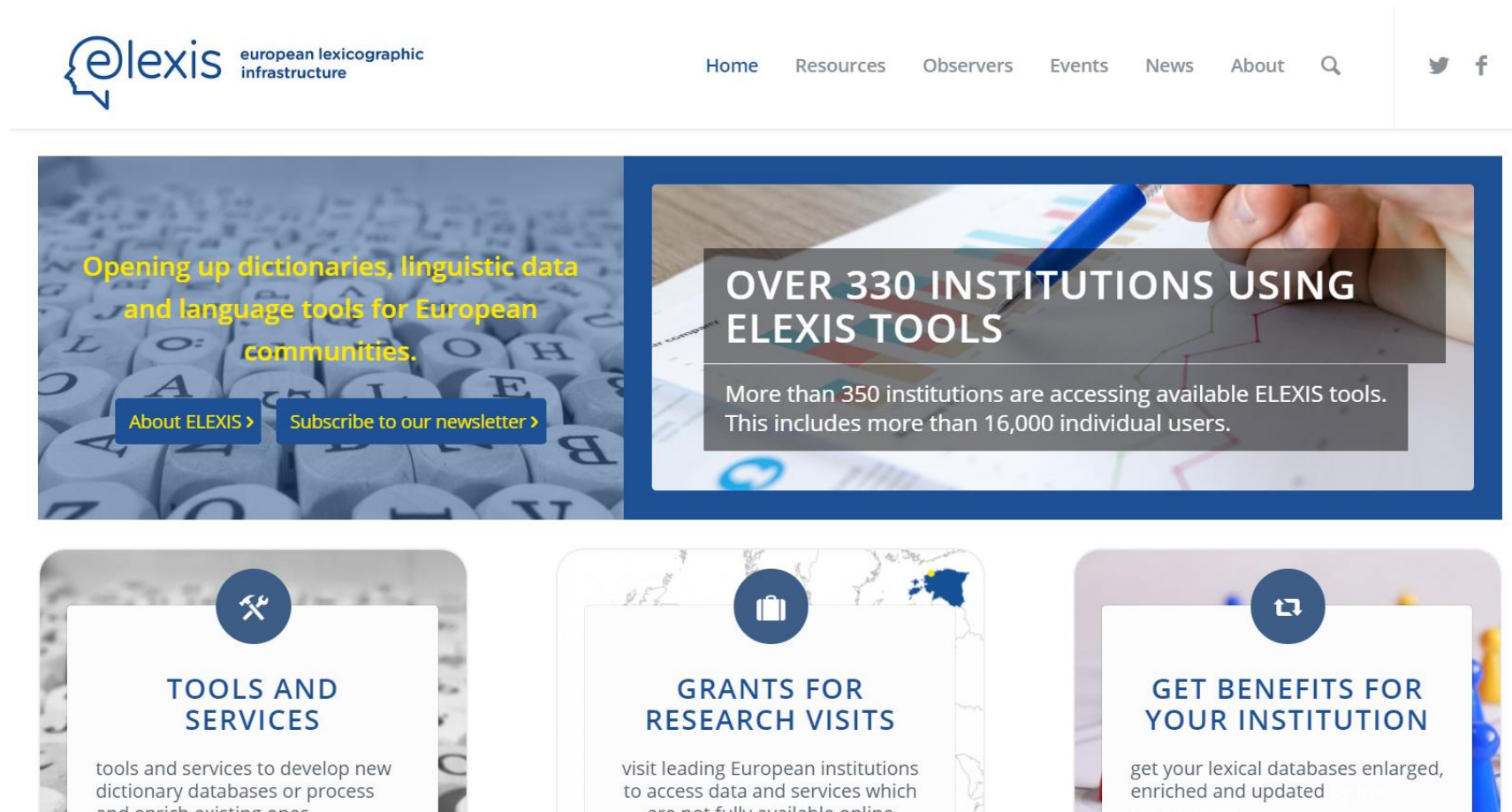
Denmark's National Strategy for Artificial Intelligence (2019)

- 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)



ELEXIS (European Lexicographic Infrastructure)



The screenshot shows the ELEXIS website homepage. At the top is the @lexis logo and navigation menu. Below is a large banner with two main sections: one about opening up dictionaries and linguistic data, and another stating that over 330 institutions use ELEXIS tools. Below the banner are three service cards: Tools and Services, Grants for Research Visits, and Get Benefits for Your Institution.

@lexis european lexicographic infrastructure

Home Resources Observers Events News About

Opening up dictionaries, linguistic data and language tools for European communities.

About ELEXIS > Subscribe to our newsletter >

OVER 330 INSTITUTIONS USING ELEXIS TOOLS

More than 350 institutions are accessing available ELEXIS tools. This includes more than 16,000 individual users.

TOOLS AND SERVICES

tools and services to develop new dictionary databases or process and enrich existing ones

GRANTS FOR RESEARCH VISITS

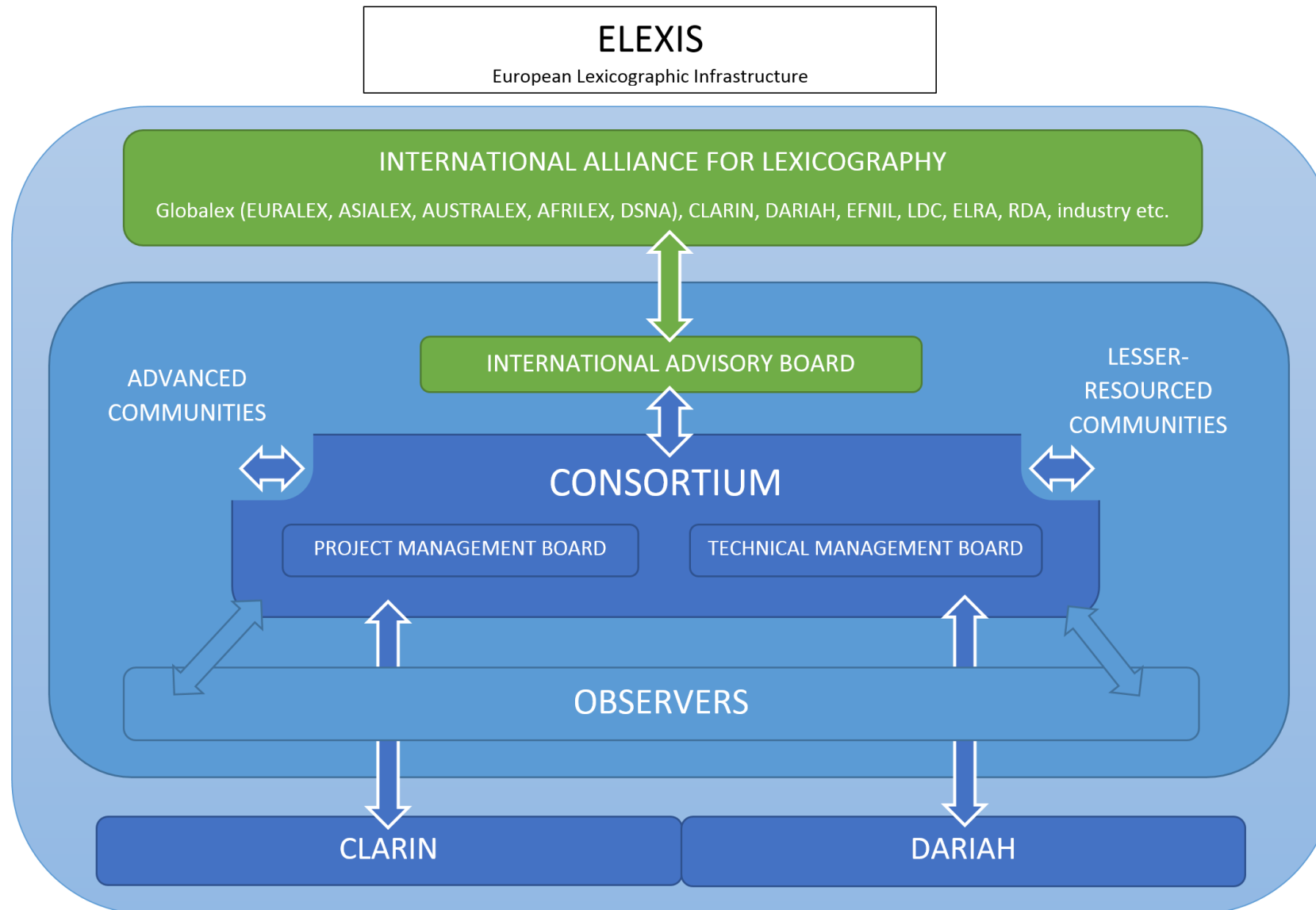
visit leading European institutions to access data and services which are not fully available online

GET BENEFITS FOR YOUR INSTITUTION

get your lexical databases enlarged, enriched and updated

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

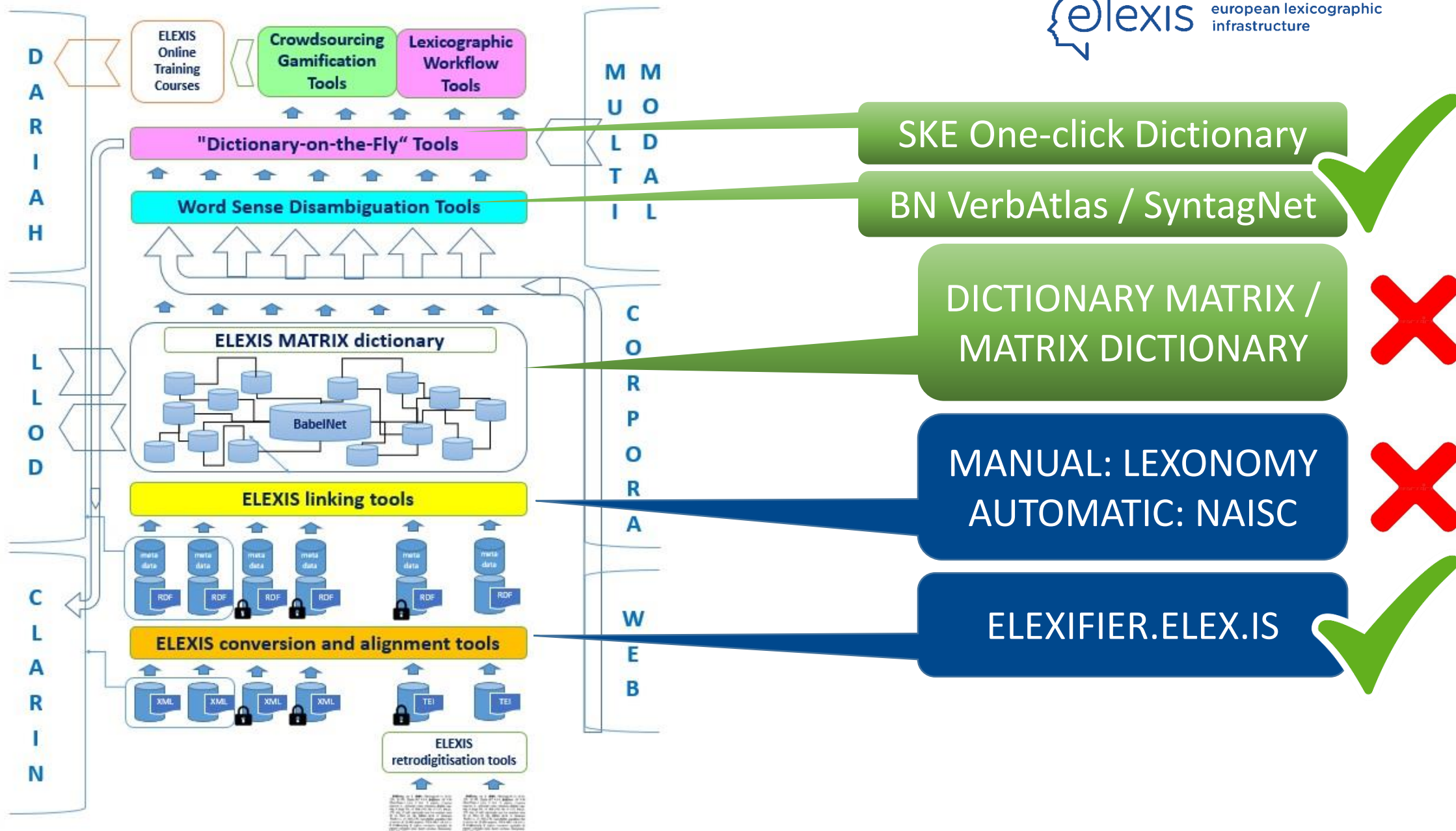


ELEXIS project summary (1)


- 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.

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.



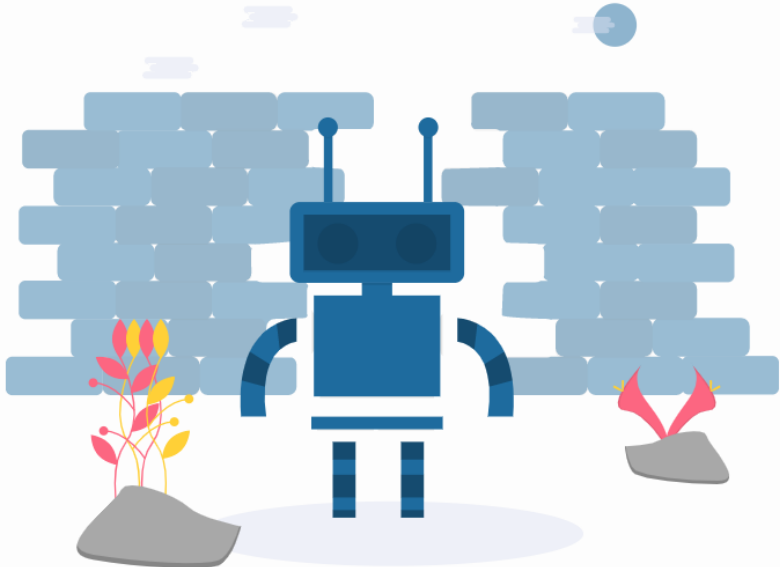
Conversion tool (<https://github.com/elexis-eu/elexifier>)

elexis european lexicographic infrastructure

Contact Support Help [APP](#)

Convert your PDF and XML dictionaries with Elexifier!

Elexifier uses advanced XML parsing and machine learning techniques to help you convert your PDF and XML dictionaries in a standardized machine-readable format.

[Start Here!](#)

“Dictionary-on-the-fly”

DASHBOARD

Gigafida v2.0 (referenčni)

Get more space +

🕒

🔗

?

!

👤

GIGAFIDA V2.0 (REFERENČNI)

Word Sketch
Collocations and word combinations

Thesaurus
Synonyms and similar words

Parallel Concordance
Translation search

N-grams
Multiword expressions (MWEs)

Trends
Diachronic analysis, neologisms

CORPUS INFO

Word Sketch Difference
Compare collocations of two words

Concordance
Examples of use in context

Wordlist
Frequency list

Keywords
Terminology extraction

One-Click Dictionary
Automatic dictionary drafting

MANAGE CORPUS

RECENTLY USED CORPORA

NEW CORPUS

Gigafida v2.0 (referenčni)	Slovenian	1,109,441,592	🗑️
wsd-sentences1	English	2,234	🗑️
English Wikipedia	English	1,356,523,079	🗑️
Araneum Anglicum Maius [2015]	English	888,466,066	🗑️
OEC	English	2,073,319,589	🗑️
Slovenian Web 2015 (slTenTen15, TreeTagger v2)	Slovenian	829,544,337	🗑️
English Corpus for SkELL 3.9	English	1,041,138,575	🗑️
Dutch Web 2014 (nlTenTen14)	Dutch	2,253,777,579	🗑️

One click dictionary



VerbAtlas – Semantic Roles & Frames

VerbAtlas

API-DOC DOWNLOAD ABOUT

deceive

SEARCH

[List of Semantic Roles](#) • [List of Frames](#)

DECEIVE

affect • feign • pretend

Make believe with the intent to deceive



lie

Tell an untruth; pretend with intent to deceive



invent • cook up • fabricate

Make up something artificial or untrue



DECEIVE

Cause someone to believe an untruth

AGENT



individual



social group

PATIENT



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

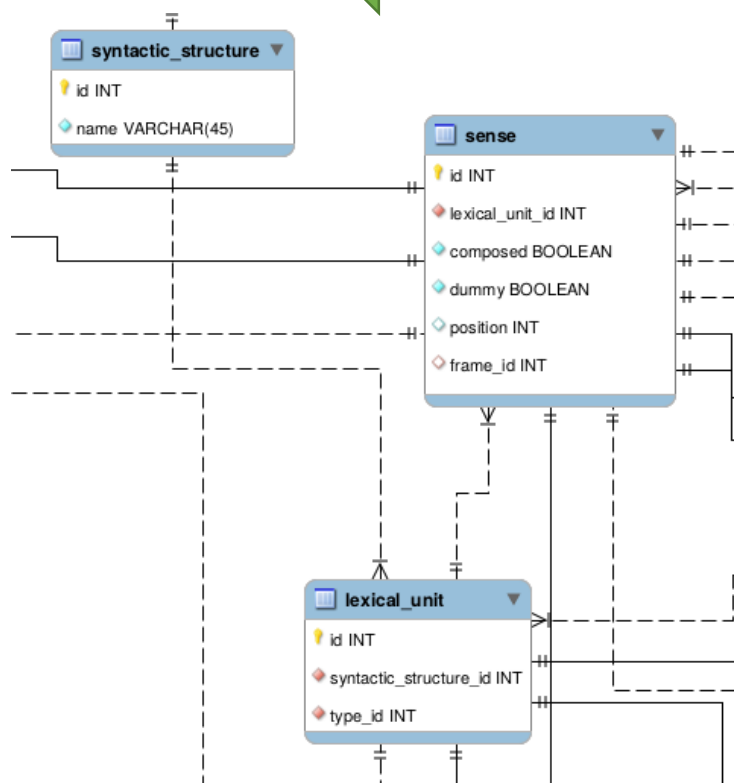
perish

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

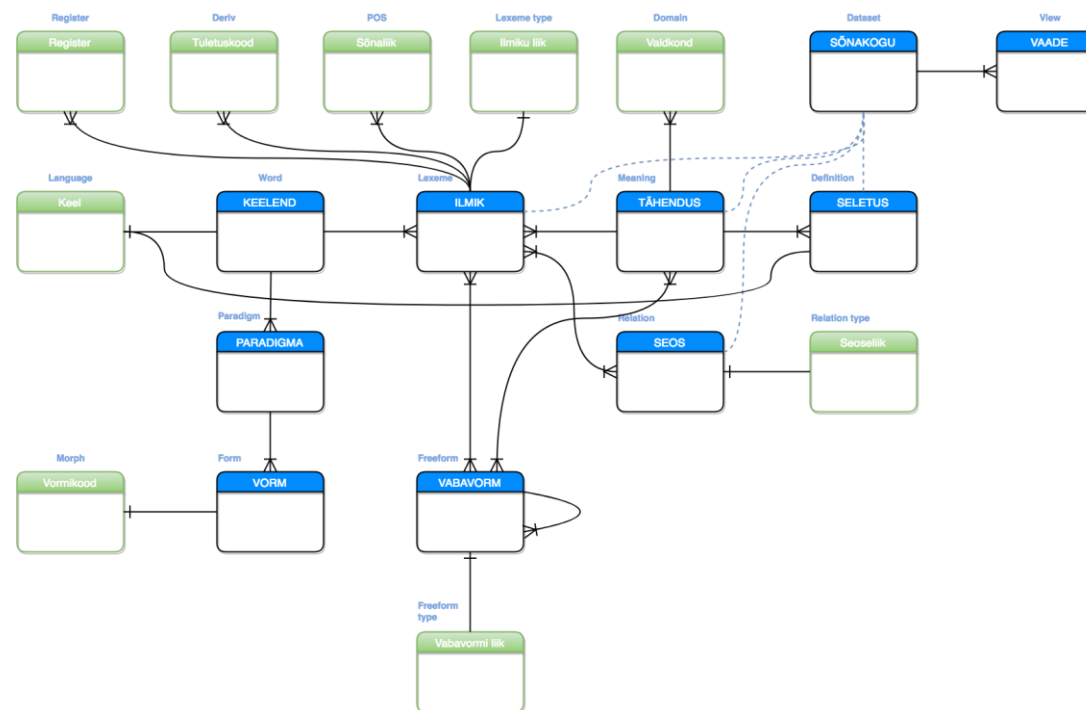
WordNet (license)

Dictionary Matrix – the cross-lingual challenge

Slovenian



Estonian



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>

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


Dictionary Matrix



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.

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	majka
Korean	어머니
Chinese	妈
Hindi	माँ
Telugu	తల్లి
Catalan	mare
Georgian	დედა
Kazakh	ана
Azerbaijani	ana
Cherokee	ᎠᎩ
Swedish	mor
Slovak	matka
Lithuanian	motė
Latvian	māte



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.