Latvijas Universitātes
Matemātikas un informātikas institūts

LATVIJAS UNIVERSITĀTE

78. Latvijas Universitātes starptautiskā zinātniskā konference

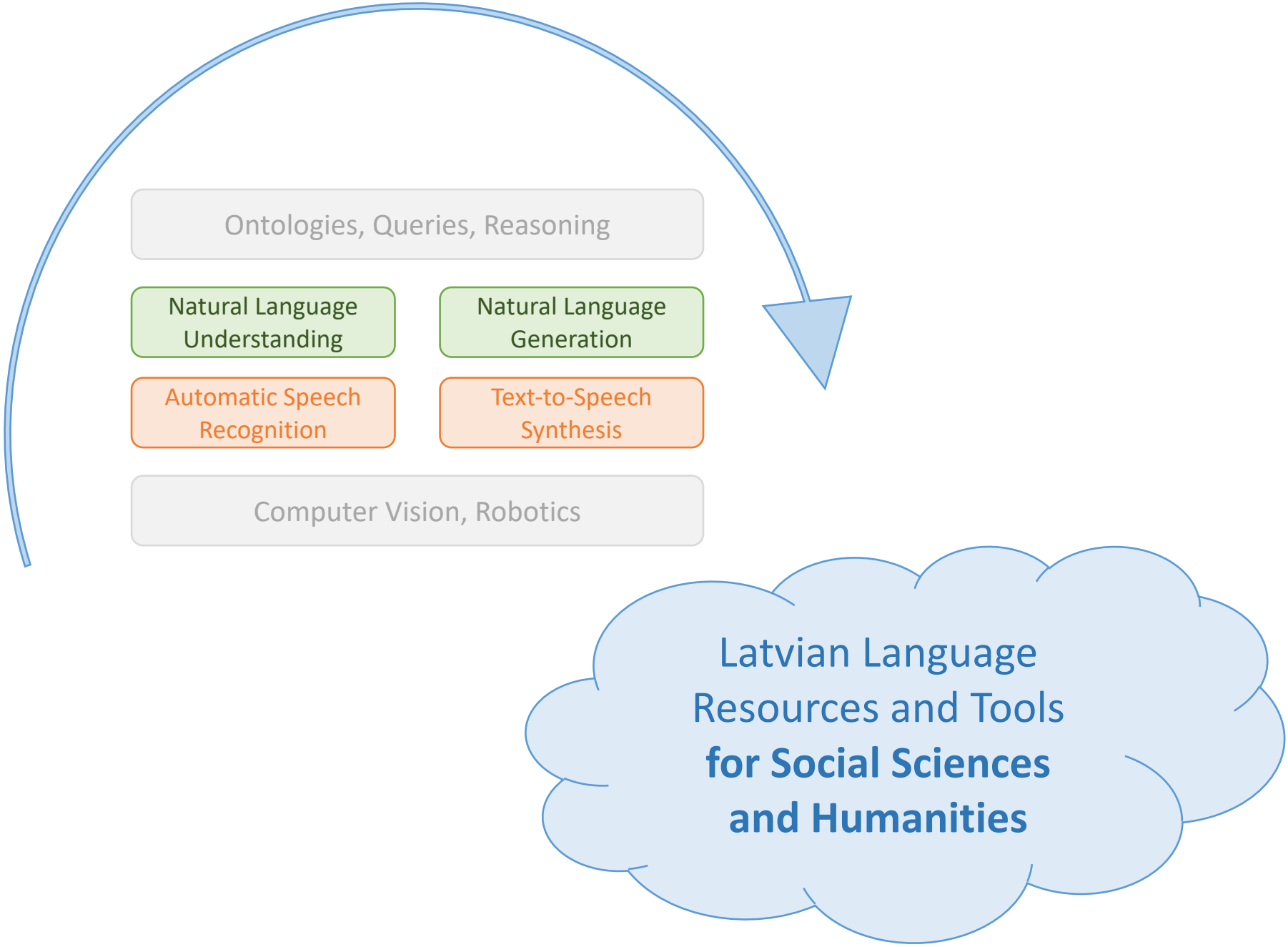
Latvian Language Resources and Tools in the CLARIN infrastructure

Normunds Grūzītis & Inguna Skadiņa

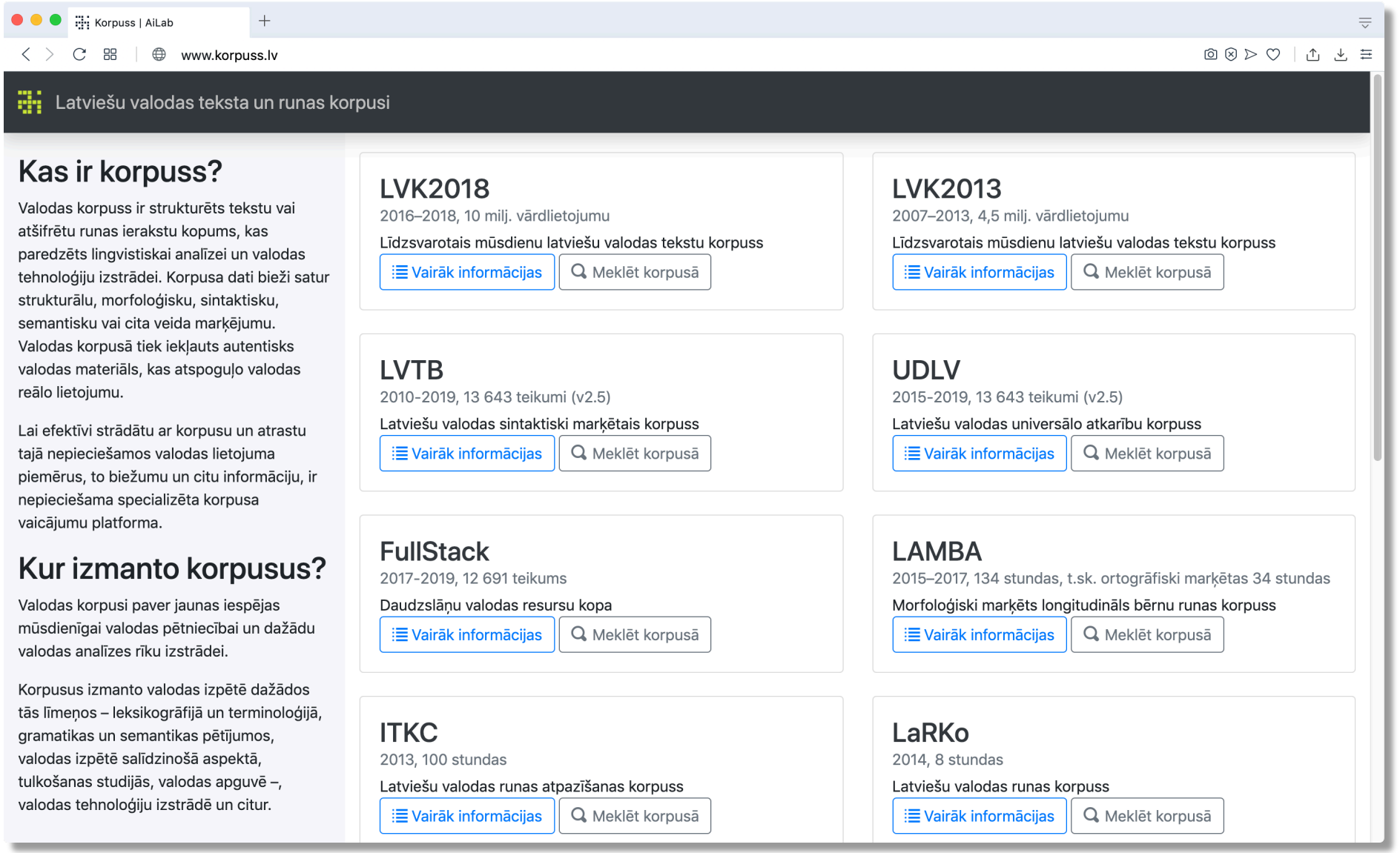
Latviešu valodas digitālie resursi un rīki vienotā pētniecības infrastruktūrā
2020. gada 5. martā

Artificial Intelligence Laboratory

Everita Andronova
Ilze Auziņa
Guntis Bārzdiņš
Guna Rābante-Buša
Roberts Darģis
Didzis Goško
Normunds Grūzītis
Inga Kaija
Kristīne Levāne-Petrova
Gunta Nešpore-Bērzkalne
Pēteris Paikens
Ilmārs Poikāns
Kristīne Pokratniece
Lauma Pretkalniņa
Laura Rituma
Inguna Skadiņa
Andrejs Spektors
Baiba Valkovska
Ilze Ziņģe
Artūrs Znotiņš

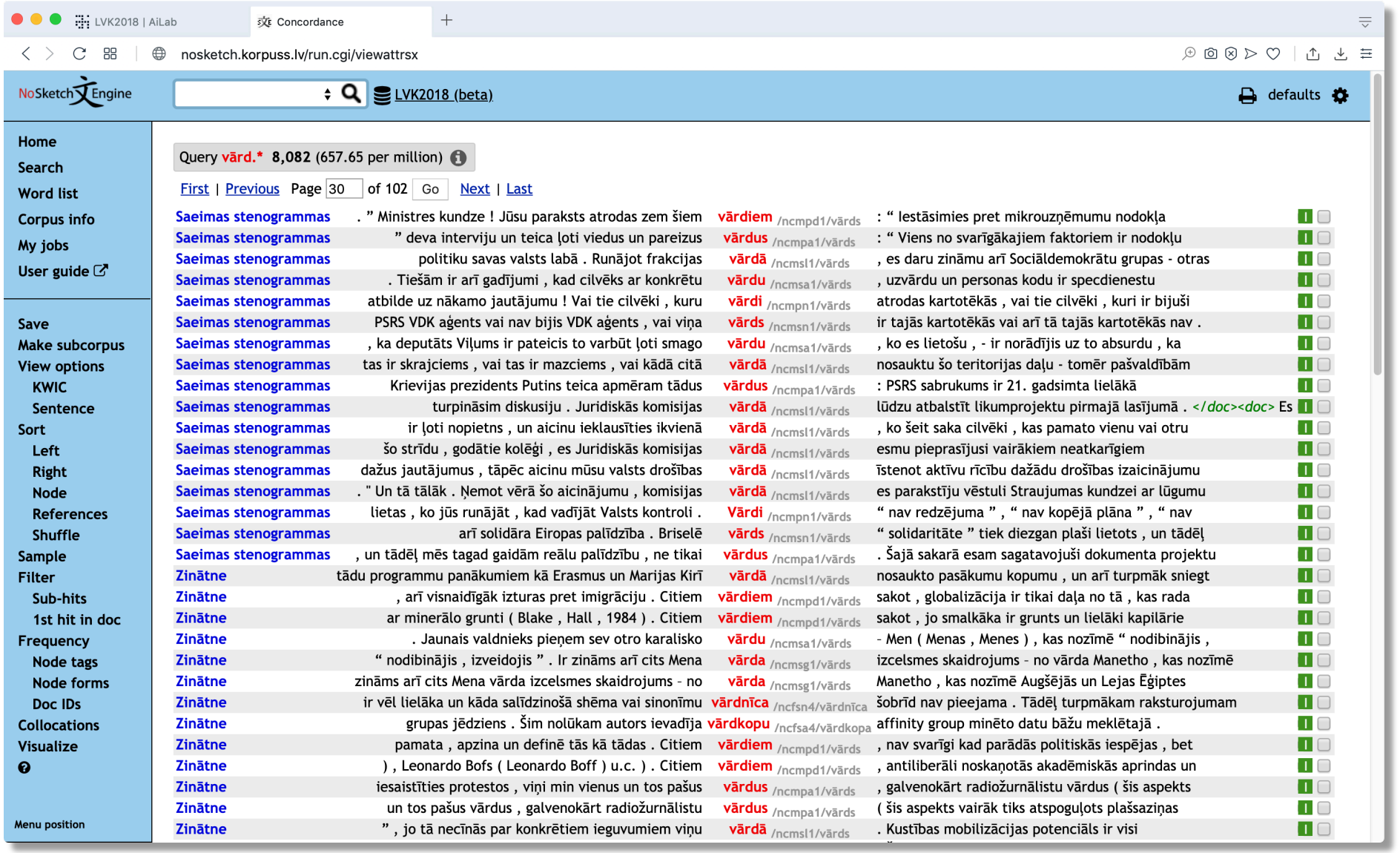

Ontologies, Queries, Reasoning
Natural Language Understanding
Natural Language Generation
Automatic Speech Recognition
Text-to-Speech Synthesis
Computer Vision, Robotics
Latvian Language Resources and Tools
for Social Sciences and Humanities

http://korpuss.lv



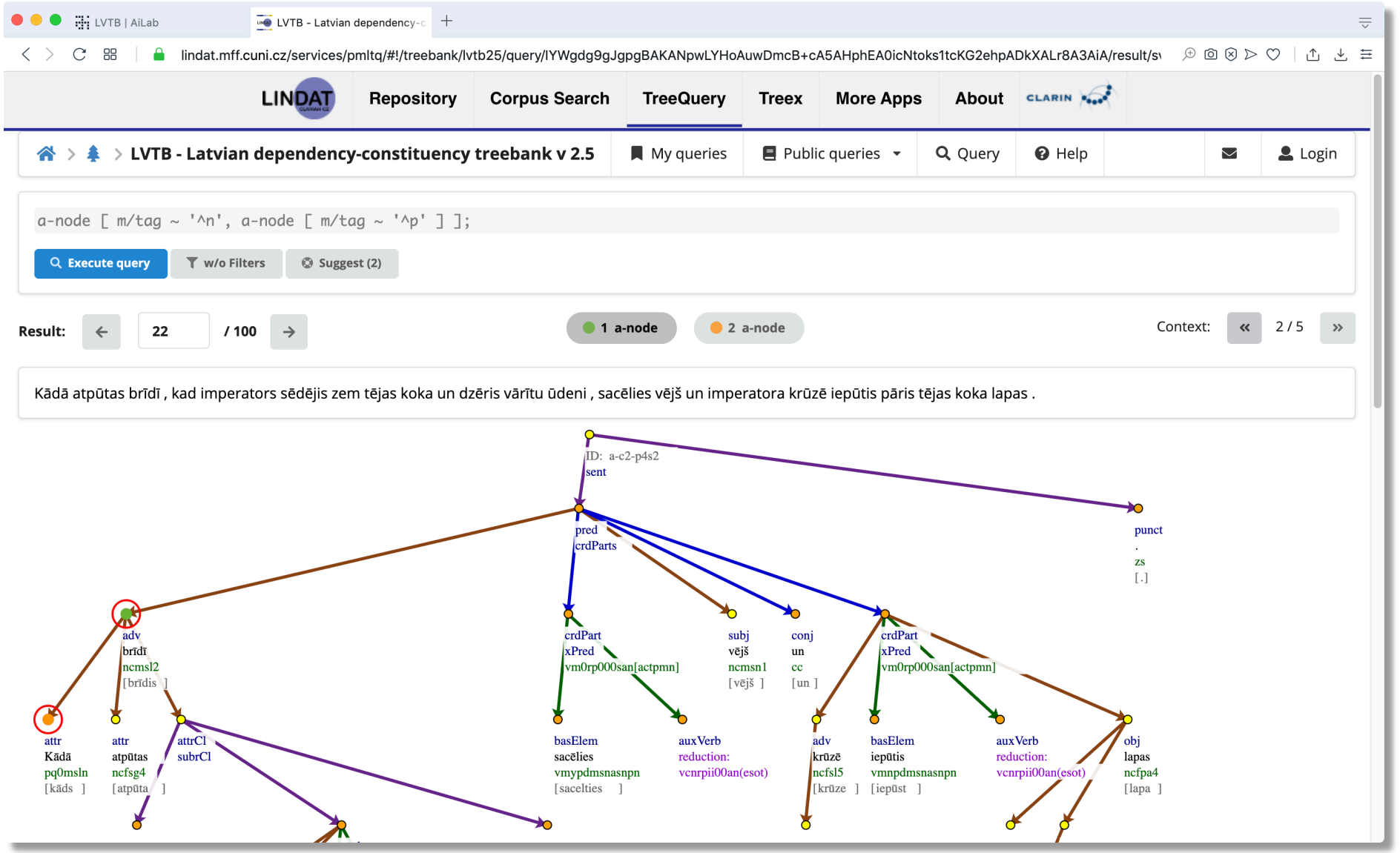
http://korpuss.lv/id/LVK2018

10M words



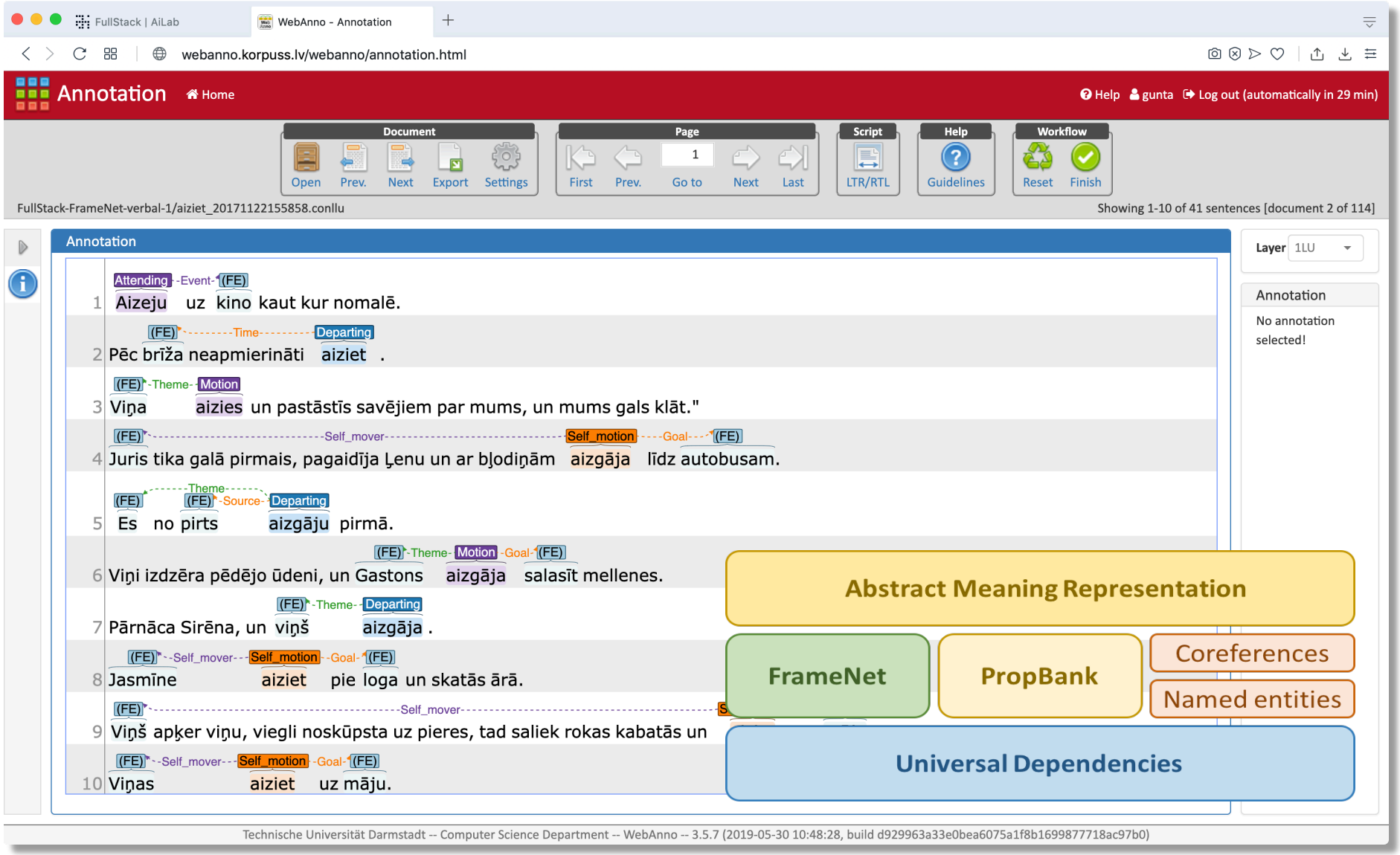
http://korpuss.lv/id/LVTB

13.6K sentences



http://korpuss.lv/id/FullStack

UD 12.7K sentences
FrameNet 20.9K annosets
PropBank 19.7K annosets
AMR 12.7Ksentences



Universal Dependencies

Afrikaans	1	49K
Amharic	1	10K
Ancient Greek	2	416K
Arabic	3	1,042K
Armenian	1	52K
Bambara	1	13K
Basque	1	121K
Belarusian	1	13K
Breton	1	10K
Bulgarian	1	156K
Cantonese	1	13K
Catalan	1	531K
Chinese	5	285K
Croatian	1	199K
Czech	5	2,222K
Danish	2	100K
Dutch	2	306K
English	7	620K
Erzya	1	15K
Estonian	2	465K
Finnish	3	377K
French	8	1,157K
Galician	2	164K
German	4	3,753K
Gothic	1	55K
Greek	1	63K
Hebrew	1	161K
Hindi	2	375K
Hungarian	1	42K
Indonesian	2	141K
Irish	1	40K
Italian	6	811K
Japanese	5	1,688K
Kazakh	1	10K
Korean	5	446K
Kurmanji	1	10K
Latin	3	582K
Latvian	1	220K
Lithuanian	2	75K
Livvi	1	1K
Maltese	1	44K
Marathi	1	3K
Mbya Guarani	2	13K
Moksha	1	<1K
Naija	1	12K
North Sami	1	26K
Norwegian	3	666K

HOME PARTNERS ANNOTATION REPOSITORIES SUGGEST A FRAME EVENTS

GLOBAL FRAMENET

the hub for collaboration in the development of frame-based language resources and applications

"Meanings are relativized to scenes"

— Fillmore

FrameNet 101

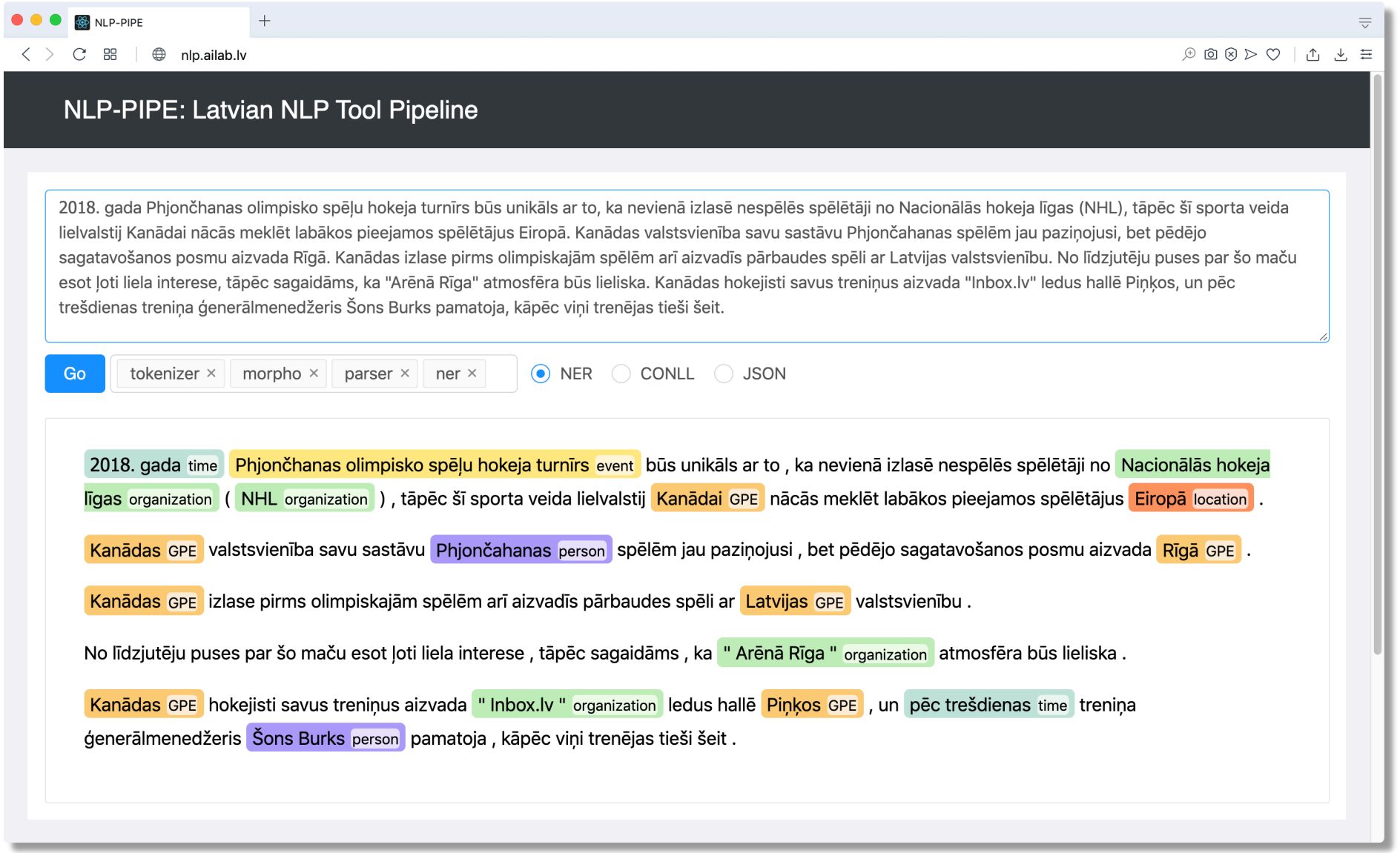
FrameNet is the computational application of the theory of Frame Semantics. In FrameNet, knowledge about the semantics of lexical items is modeled against the background semantic frames which they evoke. A frame is a sort of scene, a set of concepts related to each other in such a way that the presence of one of them makes all the other concepts readily available. The frames are also connected to each other by several types of formally defined relations, so they can be thought of as a network of nodes linked by different types of arcs.

The FrameNet project started in 1997, at the International Computer Science Institute in Berkeley, California, under the leadership of Prof. Charles Fillmore. During the past two decades, FrameNets have been built for many other languages, including Japanese, German, Spanish, Swedish, Brazilian Portuguese, Chinese, French, Hebrew, Korean and Latvian. FrameNet is applied in several areas of Natural Language Understanding. Global FrameNet is an effort to bring together all existing - and yet to be created - FrameNets in a common multilingual setting, focused on the development of collaborative

Shared Annotation Task

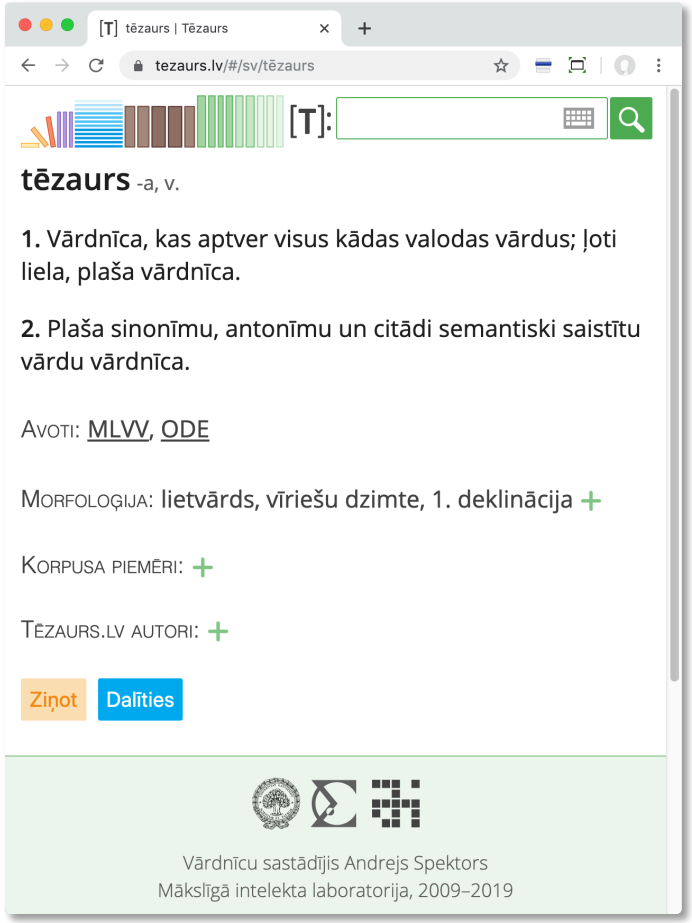
The shared annotation task was devised in part as a means to evaluate the complexity of the work required to align the FrameNets developed for different languages during the past decade and more. By annotating either translations of a given text or comparable texts from the same genre and on the same topic, we aim to assess what kinds of differences must exist between FrameNets for different languages in order to provide an adequate analysis of the lexicon of each language. Moreover, the shared annotation task will generate a collection of texts annotated with frames and LUs for several languages, which can be used in the future, for instance, as training data in a variety of applications.

http://nlp.ailab.lv



Tēzaurs.lv

- The largest open dictionary of Latvian
  - More than **317 000** entries
  - More than **129 000** synonymy relations
  - More than **90 000** inflection tables
- User statistics
  - More than **35 000** visitors per month
  - More than **100 000** queries per month
  - From more than **25** countries
- Open data and API
- Android application, Kindle dictionary
- Ongoing linking to WordNet, FrameNet



LETONIKA

VALSTS PĒTĪJUMU PROGRAMMA NACIONĀLĀ IDENTITĀTE

LATVIEŠU VALODA
VALSTS PĒTĪJUMU PROGRAMMA

KC/2.1.2.1.1/10/01/001
1.2.1.1/16/A/007
1.2.1.1/18/A/003
1.1.1.1/18/A/153
1.1.1.1/16/A/219

Industry-oriented research (ERDF)

- **Speech Recognition**
  - Development of a balanced Latvian speech corpus (2013)
  - Latvian speech recognition for media monitoring (2014–2015)
  - Latvian speech recognition and synthesis for medical applications (2019–2021)
- **Information Extraction and Meaning Representation**
  - Information Extraction for Latvian Media Monitoring (2013–2020)
  - Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian (2016–2019)

NACIONĀLAIS ATTĪSTĪBAS PLĀNS 2020
EIROPAS SAVIENĪBA
Eiropas Savienības struktūrfondi un Kohēzijas fonds
IEGULDĪJUMS TAVĀ NĀKOTNĒ
KOMPETENCES CENTRS
LETA nacionālā informācijas aģentūra
TILDE
aslimnica RĪGAS AUSTRUMU KLĪNISKĀ UNIVERSITĀTES SLIMNĪCA

Grants from Latvian Council of Science

- *Development of a Latvian Language Learner Corpus: Methods, tools and use cases* (lzp-2018/1-0527)
- *Latvian Language Understanding and Generation in Human-Computer Interaction* (lzp-2018/2-0216)
- *Latvian WordNet and Word Sense Disambiguation* (lzp-2019/1-0464)

State Research programmes

Latvian Language Agency contracts

FLPP FUNDAMENTĀLIE UN LIETIŠĶIE PĒTĪJUMU PROJEKTI

Latviešu valodas aģentūra

semti kamols

IMIS

LETONIKA

VALSTS PĒTĪJUMU PROGRAMMA NACIONĀLĀ IDENTITĀTE

LATVIEŠU VALODA VALSTS PĒTĪJUMU PROGRAMMA