

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 Bārzdīņš

Guna Rābante-Bušā

Roberts Darģis

Didzis Goško

Normunds Grūzītis

Inga Kaija

Kristīne Levāne-Petrova

Gunta Nešpore-Bērzkalne

Pēteris Paikens

Ilmārs Poikāns

Kristīne Pokratiece

Lauma Pretkalniņa

Laura Rituma

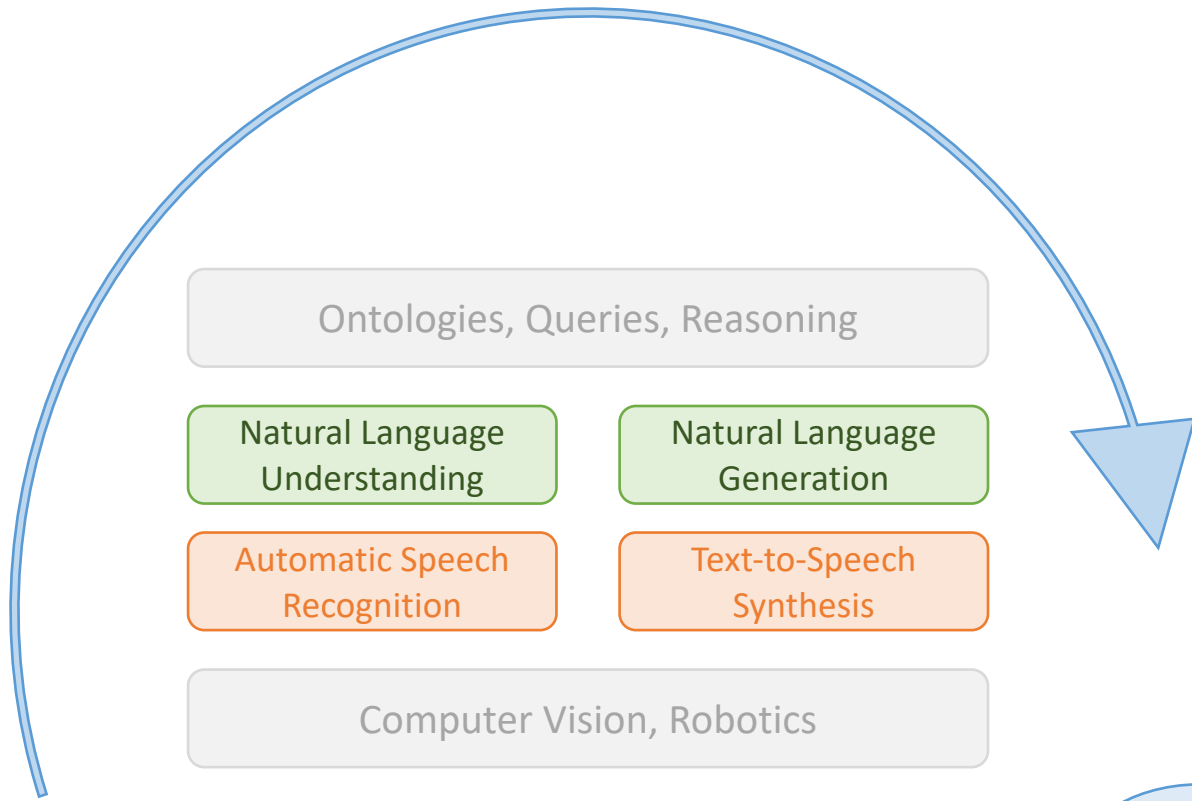
Inguna Skadiņa

Andrejs Spektors

Baiba Valkovska

Ilze Ziņģe

Artūrs Znotiņš



**Latvian Language
Resources and Tools
for Social Sciences
and Humanities**

<http://korpuss.lv>

Korpuss | AiLab

www.korpuss.lv

Latviešu valodas teksta un runas korpusi

Kas ir korpuss?

Valodas korpuss ir strukturēts tekstu vai atšifrētu runas ierakstu kopums, kas paredzēts lingvistiskai analīzei un valodas tehnoloģiju izstrādei. Korpusa dati bieži satur strukturālu, morfoloģisku, sintaktisku, semantisku vai cita veida marķējumu. Valodas korpūsā tiek iekļauts autentisks valodas materiāls, kas atspoguļo valodas reālo lietojumu.

Lai efektīvi strādātu ar korpusu un atrastu tajā nepieciešamos valodas lietojuma piemērus, to biežumu un citu informāciju, ir nepieciešama specializēta korpusa vaicājumu platforma.

Kur izmanto korpusus?

Valodas korpusi paver jaunas iespējas mūsdienīgai valodas pētniecībai un dažādu valodas analīzes rīku izstrādei.

Korpusus izmanto valodas izpētē dažādos tās līmeņos – leksikogrāfijā un terminoloģijā, gramatikas un semantikas pētījumos, valodas izpētē salīdzinošā aspektā, tulkošanas studijās, valodas apgūvē –, valodas tehnoloģiju izstrādē un citur.

LVK2018

2016–2018, 10 milj. vārdlietojumu

Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss

[Vairāk informācijas](#)

LVK2013

2007–2013, 4,5 milj. vārdlietojumu

Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss

[Vairāk informācijas](#)

LVTB

2010–2019, 13 643 teikumi (v2.5)

Latviešu valodas sintaktiski marķētais korpuss

[Vairāk informācijas](#)

UDLV

2015–2019, 13 643 teikumi (v2.5)

Latviešu valodas universālo atkarību korpuss

[Vairāk informācijas](#)

FullStack

2017–2019, 12 691 teikums

Daudzslāņu valodas resursu kopa

[Vairāk informācijas](#)

LAMBA

2015–2017, 134 stundas, t.sk. ortogrāfiski marķētas 34 stundas

Morfoloģiski marķēts longitudināls bērnu runas korpuss

[Vairāk informācijas](#)

ITKC

2013, 100 stundas

Latviešu valodas runas atpazīšanas korpuss

[Vairāk informācijas](#)

LaRKO

2014, 8 stundas

Latviešu valodas runas korpuss

[Vairāk informācijas](#)

<http://korpuss.lv/id/LVK2018>

10M words

LVK2018 | AiLab Concordance

nosketch.korpuss.lv/run.cgi/viewatrrsx

NoSketch Engine LVK2018 (beta) defaults

Home Search Word list Corpus info My jobs User guide

Save Make subcorpus View options KWIC Sentence Sort Left Right Node References Shuffle Sample Filter Sub-hits 1st hit in doc Frequency Node tags Node forms Doc IDs Collocations Visualize Menu position

Query **vārd.*** 8,082 (657.65 per million) ?

First | Previous Page 30 of 102 Go Next | Last

Saeimas stenogrammas	. " Ministres kundze ! Jūsu paraksts atrodas zem šiem	vārdiem	/ncmpd1/vārds	: " lestāsimies pret mikrouzņēmumu nodokļa	<input type="checkbox"/>
Saeimas stenogrammas	" deva interviju un teica ļoti viedus un pareizus	vārdus	/ncmpa1/vārds	: " Viens no svarīgākajiem faktoriem ir nodokļu	<input type="checkbox"/>
Saeimas stenogrammas	politiku savas valsts labā . Runājot frakcijas	vārdā	/ncmsl1/vārds	, es daru zināmu arī Sociāldemokrātu grupas - otras	<input type="checkbox"/>
Saeimas stenogrammas	. Tiešām ir arī gadījumi , kad cilvēks ar konkrētu	vārdu	/ncmsa1/vārds	, uzvārdu un personas kodu ir specdienestu	<input type="checkbox"/>
Saeimas stenogrammas	atbilde uz nākamo jautājumu ! Vai tie cilvēki , kuru	vārdi	/ncmpn1/vārds	atrodas kartotēkās , vai tie cilvēki , kuri ir bijuši	<input type="checkbox"/>
Saeimas stenogrammas	PSRS VDK aģents vai nav bijis VDK aģents , vai viņa	vārds	/ncmsn1/vārds	ir tajās kartotēkās vai arī tā tajās kartotēkās nav .	<input type="checkbox"/>
Saeimas stenogrammas	, ka deputāts Viļums ir pateicis to vārbūt ļoti smago	vārdu	/ncmsa1/vārds	, ko es lietošu , - ir norādījis uz to absurdu , ka	<input type="checkbox"/>
Saeimas stenogrammas	tas ir skrajciems , vai tas ir mazciems , vai kādā citā	vārdā	/ncmsl1/vārds	nosauktu šo teritorijas daļu - tomēr pašvaldībām	<input type="checkbox"/>
Saeimas stenogrammas	Krievijas prezidents Putins teica apmēram tādus	vārdus	/ncmpa1/vārds	: PSRS sabrukums ir 21. gadsimta lielākā	<input type="checkbox"/>
Saeimas stenogrammas	turpināsim diskusiju . Juridiskās komisijas	vārdā	/ncmsl1/vārds	lūdzu atbalstīt likumprojektu pirmajā lasījumā .	<input type="checkbox"/>
Saeimas stenogrammas	ir ļoti nopietns , un aicinu ieklausīties ikvienā	vārdā	/ncmsl1/vārds	, ko šeit saka cilvēki , kas pamato vienu vai otru	<input type="checkbox"/>
Saeimas stenogrammas	šo strīdu , godātie kolēģi , es Juridiskās komisijas	vārdā	/ncmsl1/vārds	esmu pieprasījis vairākiem neatkarīgiem	<input type="checkbox"/>
Saeimas stenogrammas	dažus jautājumus , tāpēc aicinu mūsu valsts drošības	vārdā	/ncmsl1/vārds	īstenot aktīvu rīcību dažādu drošības izaicinājumu	<input type="checkbox"/>
Saeimas stenogrammas	. " Un tā tālāk . Ņemot vērā šo aicinājumu , komisijas	vārdā	/ncmsl1/vārds	es parakstīju vēstuli Straujumas kundzei ar lūgumu	<input type="checkbox"/>
Saeimas stenogrammas	lietas , ko jūs runājat , kad vadījāt Valsts kontroli .	vārdi	/ncmpn1/vārds	" nav redzējuma " , " nav kopējā plāna " , " nav	<input type="checkbox"/>
Saeimas stenogrammas	arī solidāra Eiropas palīdzība . Briselē	vārds	/ncmsn1/vārds	" solidaritāte " tiek diezgan plaši lietots , un tādēļ	<input type="checkbox"/>
Saeimas stenogrammas	, un tādēļ mēs tagad gaidām reālu palīdzību , ne tikai	vārdus	/ncmpa1/vārds	. Šajā sakarā esam sagatavojuši dokumenta projektu	<input type="checkbox"/>
Zinātne	tādu programmu panākumiem kā Erasmus un Marijas Kirī	vārdā	/ncmsl1/vārds	nosauktu pasākumu kopumu , un arī turpmāk sniegt	<input type="checkbox"/>
Zinātne	, arī visnaidīgāk izturas pret imigrāciju . Citiem	vārdiem	/ncmpd1/vārds	sakot , globalizācija ir tikai daļa no tā , kas rada	<input type="checkbox"/>
Zinātne	ar minerālo grunti (Blake , Hall , 1984) . Citiem	vārdiem	/ncmpd1/vārds	sakot , jo smalkāka ir grunts un lielāki kapilārie	<input type="checkbox"/>
Zinātne	. Jaunais valdnieks pieņem sev otro karalisko	vārdu	/ncmsa1/vārds	- Men (Menas , Menes) , kas nozīmē " nodibinājis ,	<input type="checkbox"/>
Zinātne	" nodibinājis , izveidojis " . Ir zināms arī cits Mena	vārda	/ncmsg1/vārds	izcelsmes skaidrojums - no vārda Manetho , kas nozīmē	<input type="checkbox"/>
Zinātne	zināms arī cits Mena vārda izcelsmes skaidrojums - no	vārda	/ncmsg1/vārds	Manetho , kas nozīmē Augšējās un Lejas Ēģiptes	<input type="checkbox"/>
Zinātne	ir vēl lielāka un kāda salīdzinošā shēma vai sinonīmu	vārdnīca	/ncfsn4/vārdnīca	šobrīd nav pieejama . Tādēļ turpmākam raksturojumam	<input type="checkbox"/>
Zinātne	grupas jēdziens . Šim nolūkam autors ievadīja	vārdkopu	/ncfsa4/vārdkopa	affinity group minēto datu bāzu meklētajā .	<input type="checkbox"/>
Zinātne	pamata , apzina un definē tās kā tādas . Citiem	vārdiem	/ncmpd1/vārds	, nav svarīgi kad parādās politiskās iespējas , bet	<input type="checkbox"/>
Zinātne	, Leonardo Bofs (Leonardo Boff) u.c.) . Citiem	vārdiem	/ncmpd1/vārds	, antiliberāli noskaņotās akadēmiskās aprindas un	<input type="checkbox"/>
Zinātne	iesaistīties protestos , viņi min vienus un tos pašus	vārdus	/ncmpa1/vārds	, galvenokārt radiožurnālistu vārdus (šis aspekts	<input type="checkbox"/>
Zinātne	un tos pašus vārdus , galvenokārt radiožurnālistu	vārdus	/ncmpa1/vārds	(šis aspekts vairāk tiks atspoguļots plašsaziņas	<input type="checkbox"/>
Zinātne	, jo tā necinās par konkrētiem ieguvumiem viņu	vārdā	/ncmsl1/vārds	. Kustības mobilizācijas potenciāls ir visi	<input type="checkbox"/>

<http://korpuss.lv/id/LVTB>

13.6K sentences

LVTB | AiLab

LVTB - Latvian dependency-constituency treebank v 2.5

lindat.mff.cuni.cz/services/pmltq/#!/treebank/lvtb25/query/IYWgdg9gJgpgBAKANpwLYHoAuwDmcB+cA5AHphEA0icNtoks1tcKG2ehpADkXALr8A3AiA/result/sv

LINDAT Repository Corpus Search **TreeQuery** Treex More Apps About CLARIN

Home > LVTB - Latvian dependency-constituency treebank v 2.5 My queries Public queries Query Help Login

a-node [m/tag ~ '^n', a-node [m/tag ~ '^p']];

Execute query w/o Filters Suggest (2)

Result: 22 / 100 1 a-node 2 a-node Context: 2 / 5

Kādā atpūtas brīdī , kad imperators sēdējis zem tējas koka un dzēris vārītu ūdeni , sacēlies vējš un imperatora krūzē iepūtis pāris tējas koka lapas .

The diagram shows a dependency-constituency tree for the sentence. The root node is 'sent' (ID: a-c2-p4s2). The tree structure is as follows:

- sent (ID: a-c2-p4s2) branches into:
 - adv (ID: a-c2-p4s2) branches into:
 - attr (ID: a-c2-p4s2) branches into: Kādā [pq0msln [kāds]]
 - attr (ID: a-c2-p4s2) branches into: atpūtas [ncfsg4 [atpūta]]
 - attrCl (ID: a-c2-p4s2) branches into: brīdī [brīdis]]
 - pred (ID: a-c2-p4s2) branches into:
 - crdPart (ID: a-c2-p4s2) branches into:
 - xPred (ID: a-c2-p4s2) branches into:
 - basElem (ID: a-c2-p4s2) branches into: sacēlies [vmypdmsnasn [sacelties]]
 - auxVerb (ID: a-c2-p4s2) branches into: reduction: vcnrpii00an(esot)
 - suby (ID: a-c2-p4s2) branches into: vējš [ncmsn1 [vējš]]
 - conj (ID: a-c2-p4s2) branches into: un [cc [un]]
 - crdPart (ID: a-c2-p4s2) branches into:
 - xPred (ID: a-c2-p4s2) branches into:
 - adv (ID: a-c2-p4s2) branches into: krūzē [ncfsl5 [krūze]]
 - basElem (ID: a-c2-p4s2) branches into: iepūtis [vmnpdmsnasn [iepūst]]
 - auxVerb (ID: a-c2-p4s2) branches into: reduction: vcnrpii00an(esot)
 - obj (ID: a-c2-p4s2) branches into: lapas [ncfpa4 [lapa]]
 - punct (ID: a-c2-p4s2) branches into: . [zs [.]]

<http://korpuss.lv/id/FullStack>

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

Annotation

Home Help gunta Log out (automatically in 29 min)

Document: Open Prev. Next Export Settings

Page: First Prev. 1 Go to Next Last

Script: LTR/RTL

Help: Guidelines

Workflow: Reset Finish

FullStack-FrameNet-verbal-1/aiziet_20171122155858.conllu Showing 1-10 of 41 sentences [document 2 of 114]

Annotation

Layer 1LU

Annotation
No annotation selected!

1 **Attending** -Event-(FE)
1 Aizeju uz kino kaut kur nomalē.

2 (FE)---Time---Departing
2 Pēc brīža neapmierināti aiziet .

3 (FE)-Theme-Motion
3 Viņa aizies un pastāstīs savējiem par mums, un mums gals klāt."

4 (FE)---Self_mover---Self_motion---Goal---(FE)
4 Juris tika galā pirmais, pagaidīja Ļenu un ar bļodiņām aizgāja līdz autobusam.

5 (FE)---Theme---(FE)-Source-Departing
5 Es no pirts aizgāju pirmā.

6 (FE)-Theme-Motion-Goal-(FE)
6 Viņi izdzēra pēdējo ūdeni, un Gastons aizgāja salasīt mellenes.

7 (FE)-Theme-Departing
7 Pārnāca Sirēna, un viņš aizgāja .

8 (FE)-Self_mover---Self_motion---Goal---(FE)
8 Jasmīne aiziet pie loga un skatās ārā.

9 (FE)---Self_mover---
9 Viņš apķer viņu, viegli noskūpst uz pieres, tad saliek rokas kabatās un

10 (FE)-Self_mover---Self_motion---Goal---(FE)
10 Viņas aiziet uz māju.

Abstract Meaning Representation

FrameNet PropBank

Coreferences
Named entities

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



“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>



NLP-PIPE: Latvian NLP Tool Pipeline

2018. gada Phjončhanas olimpisko spēļu hokeja turnīrs būs unikāls ar to, ka nevienā izlasē nespēlēs spēlētāji no Nacionālās hokeja līgas (NHL), tāpēc šī sporta veida lielvalstij Kanādai nācās meklēt labākos pieejamos spēlētājus Eiropā. Kanādas valstsvienība savu sastāvu Phjončhanas spēlēm jau paziņojusi, bet pēdējo sagatavošanos posmu aizvada Rīgā. Kanādas izlase pirms olimpiskajām spēlēm arī aizvadīs pārbaudes spēli ar Latvijas valstsvienību. No līdzjutēju puses par šo maču esot ļoti liela interese, tāpēc sagaidāms, ka "Arēnā Rīga" atmosfēra būs lieliska. Kanādas hokejisti savus treniņus aizvada "Inbox.lv" ledus hallē Piņķos, un pēc trešdienas treniņa ģenerālmenedžeris Šons Burks pamatoja, kāpēc viņi trenējas tieši šeit.

Go **NER** CONLL JSON

2018. gada **time** Phjončhanas olimpisko spēļu hokeja **event** būs unikāls ar to, ka nevienā izlasē nespēlēs spēlētāji no **Nacionālās hokeja līgas organization** (**NHL organization**), tāpēc šī sporta veida lielvalstij **Kanādai GPE** nācās meklēt labākos pieejamos spēlētājus **Eiropā location** .

Kanādas GPE valstsvienība savu sastāvu **Phjončhanas person** spēlēm jau paziņojusi, bet pēdējo sagatavošanos posmu aizvada **Rīgā GPE** .

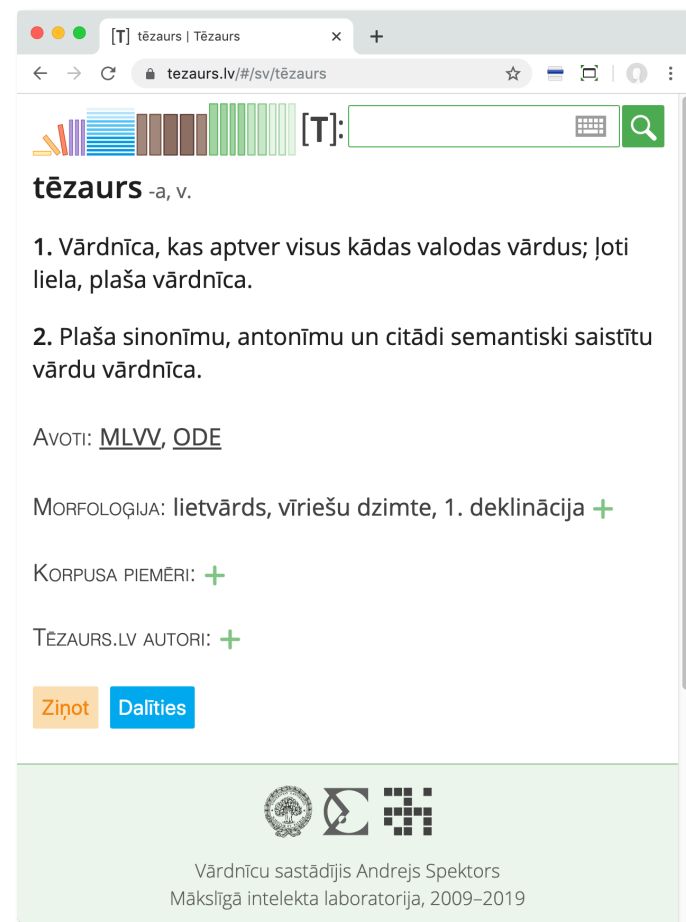
Kanādas GPE izlase pirms olimpiskajām spēlēm arī aizvadīs pārbaudes spēli ar **Latvijas GPE** valstsvienību .

No līdzjutēju puses par šo maču esot ļoti liela interese, tāpēc sagaidāms, ka **" Arēnā Rīga " organization** atmosfēra būs lieliska .

Kanādas GPE hokejisti savus treniņus aizvada **" Inbox.lv " organization** ledus hallē **Piņķos GPE** , un **pēc trešdienas time** treniņa ģenerālmenedžeris **Šons Burks person** pamatoja, kāpēc viņi trenējas tieši šeit .

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



The screenshot shows a web browser window with the URL tezaurs.lv/#/sv/tezaurs. The page displays the definition of the Latvian word "tēzaurs".

tēzaurs -a, v.

1. Vārdnīca, kas aptver visus kādas valodas vārdus; ļoti liela, plaša vārdnīca.
2. Plaša sinonīmu, antonīmu un citādi semantiski saistītu vārdu vārdnīca.

AVOTI: [MLWV](#), [ODE](#)

MORFOLOĢIJA: lietvārds, vīriešu dzimte, 1. deklinācija +

KORPUSA PIEMĒRI: +

TĒZAURS.LV AUTORI: +

Ziņot Daļīties

Vārdnīcu sastādījis Andrejs Spektors
Mākslīgā intelekta laboratorija, 2009–2019

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)



Grants from Latvian Council of Science

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

State Research programmes

Latvian Language Agency contracts



FLPP
FUNDAMENTĀLIE UN
LIETIŠĀJIE PĒTĪJUMU
PROJEKTI



Latviešu valodas
agentūra



VALSTS
PĒTĪJUMU PROGRAMMA
NACIONĀLĀ
IDENTITĀTE

