

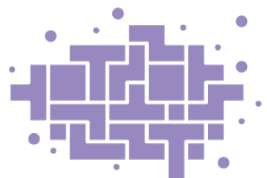


Latvijas Universitātes
Matemātikas un informātikas institūts



Jaunākie LU MII Mākslīgā intelekta laboratorijas valodas resursi izglītībai un pētniecībai

Ilze Auziņa, Pēteris Paikens, Roberts Dargis



FLPP
FUNDAMENTĀLIE UN
LIETISĶIE PĒTĪJUMU
PROJEKTI

CLARIN-LV konference «Latviešu valodas digitālie resursi un rīki izglītībai»
2021. gada 2. jūnijs





Latviešu valodas **runas atpazīšana** (runa→teksts) un **runas sintēze** (teksts→runa).

Teksta sintaktiskā un semantiskā analīze

strukturētas informācijas izguvei, jautājumu atbildēšanai, cilvēka-datora komunikācijai.

Daudzvalodu **tekstrade** (teksta sintēze) no abstraktām nozīmes reprezentācijām un no datiem.

Zināšanu attēlošana kontrolētā dabiskā valodā.

Anotētu **valodas resursu** izveide NLP un ML vajadzībām: mašīnlasāmas **vārdnīcas**, teksta un runas **korpusi**.

Dziļās mašīnmācīšanās un **stimulētās mašīnmācīšanās** lietojumi **robotikā**, datorredzē, valodas apguvē.



Vietne *korpus.lv*



Latviešu valodas tekstu un runas korpusi

Kas ir korpusi?

Valodas korpusi ir strukturēti tekstu vai atšifrētu runas ierakstu kopumi, kas paredzēti lingvistiskai analīzei un valodas tehnoloģiju izstrādei. Korpusa dati bieži satur strukturālu, morfoloģisku, sintaktisku, semantisku vai cita veida marķējumu. Valodas korpusā 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 apguvē –, valodas tehnoloģiju izstrādē un citur.

LVK2018

2016–2018, 10 milj. vārdlietojumu

Līdzsvarotais mūsdienu latviešu valodas tekstu korpusi

[Vairāk informācijas](#)

[Meklēt korpusā](#)

LaVA

2018–2021, 100 000 vārdlietojumu (plānotais apjoms)

Latviešu valodas apguvēju korpusi

[Vairāk informācijas](#)

[Meklēt korpusā](#)

FullStack

2017–2019, 12 691 teikums

Daudzslāņu valodas resursu kopa

[Vairāk informācijas](#)

[Meklēt korpusā](#)

ITKC

2013, 100 stundas

Latviešu valodas runas atpazīšanas korpusi

[Vairāk informācijas](#)

[Meklēt korpusā](#)

LVK2013

2007–2013, 4,5 milj. vārdlietojumu

Līdzsvarotais mūsdienu latviešu valodas tekstu korpusi

[Vairāk informācijas](#)

[Meklēt korpusā](#)

UDLV

2015–2019, 13 643 teikumi (v2.5)

Latviešu valodas universālo atkarību korpusi

[Vairāk informācijas](#)

[Meklēt korpusā](#)

LAMBA

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

Morfoloģiski marķēts longitudināls bērnu runas korpusi

[Vairāk informācijas](#)

[Meklēt korpusā](#)

LaRko

2014, 8 stundas

Latviešu valodas runas korpusi

[Vairāk informācijas](#)

[Meklēt korpusā](#)

Latviešu valodas apguvēju korpus

Korpusā iekļauti to Latvijas augstākajās mācību iestādes studējošo ārvalstnieku darbi, kuri latviešu valodu apgūst kā svešvalodu pirmo vai otro semestri. Teksti ir automātiski morfoloģiski marķēti, un tajos ir manuāli marķētas valodas apguvēju kļūdas. Korpusa apjoms ir gandrīz 1000 studentu esejas (kopumā korpusā – vairāk nekā 180 tūkst. vārdlietojumu).

Korpusā ikviens interesents var noskaidrot, piemēram, kas latviešu valodas apguvējam sagādā vislielākās grūtības, kuri ir visbiežāk lietotie vārdi un vārdu formas. Korpusā ir iespējams meklēt, izmantojot metadatus, t. i., informāciju par valodas apguvēja dzimumu, vecumu, dzimto valodu, svešvalodu prasmi.

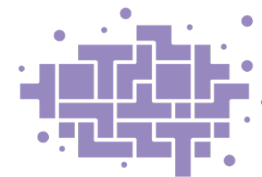
Korpusa veidotāji: Ilze Auziņa, Inga Kaija, Kristīne Levāne-Petrova, Kristīne Pokratniece un Roberts Dargis.

Sadarbības partneri



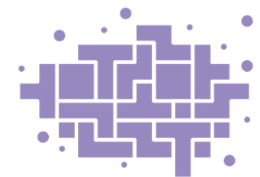
Finansējuma avots: FLPP projekts "Latviešu valodas apguvēju korpusa izveide: metodes rīki un izmantojums" (Nr. lzp-2018/1-0527).

LaVA: korpusa raksturojums



- *Latviešu valodas apguvēju korpusa* (LaVA) apjoms:
 - vairāk nekā 1000 studentu darbu,
 - vairāk nekā 180 000 vārdlietojumu.
- Latvijas augstskolās studējošo ārvalstu studentu darbi, kas savās pamatstudijās vai papildus tām apgūst latviešu valodu.
- Latviešu valodas prasmes līmenis – A1 (A2).
- Temati: *Es un mana ģimene, Mana ikdiena, Manas studijas* u.tml.
- Dati ir morfoloģiski marķēti, tekstos ir marķētas valodas apguvēju pieļautās kļūdas.

LaVA: vienkāršā meklēša korpusā



Meklēt korpusā

Tekstvienība

Teksta tips

- ☒ Oriģinālais teksts
- ☒ Labotais teksts

Kļūdu tips

- ☒ Ar kļūdām
- ☒ Bez kļūdām

Dzimtā valoda

Citas valodas

Dzimums

Meklēt

Palīdzība

Atrasti 169 teksta vienumi

Teksts ID 11. Vienums nr. 24

un man ir divdesmit gadi . Es esmu no Itālijas un es runāju vācu , italū , angļu
, un man ir divdesmit gadi . Es esmu no Itālijas un es runāju vācu , itāļu , angļu

valodā
valodā

. Es studeju medicīnu Rīgas Stradiņa universitatē . Mana nedēļa
. Es studēju medicīnu Rīgas Stradiņa universitatē . Mana nedēļa

Teksts ID 17. Vienums nr. 46

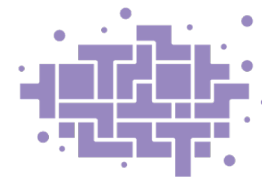
Rīgā . Primajā gada es macojs biokīmija , anatomija un histologigai , un Latviski
Rīga . Pirmajā gadā es mācos bioķīmiju , anatomiju un histoloģiju , un latviešu

valodai
valodu

. Visi kursi ir viegli izņemot Latvian valodai . Latvian valoda ir
. Visi kursi ir viegli , izņemot latviešu valodu . Latviešu valoda ir

<http://lava.korpuss.lv> → Meklēšana korpusā

LaVA datu kvantitatīvā analīze



Izmantojot *SketchEngine* korpusa pārlūkprogrammu, iespējams veikt gan datu kvantitatīvo, gan kvalitatīvo analīzi.

Kvantitatīvā pieeja

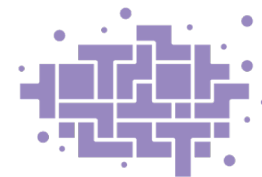
- ļauj noskaidrot kāda vārda, vārdformas vai konstrukcijas lietojuma tendences, vai gluži pretēji – to, ka kāds vārds vai vārdforma netiek lietoti pietiekami;
- ļauj noteikt izplatītākās valodas apguvēju pieļautās kļūdas;
- nodrošina objektīvu valodas apguvēju datu analīzi, aplūkojot tos no dažādiem aspektiem un veicot atkārtotu analīzi.

Biežāk lietotie vārdi

	Lemma	Absolute Frequency		Per Million	
1	es	15,176		67,310.081	...
2	būt	10,456		46,375.475	...
3	un	7,687		34,094.135	...
4	mana	3,815		16,920.661	...
5	viņš	3,722		16,508.179	...
6	patikt	3,689		16,361.814	...
7	mans	3,368		14,938.083	...
8	viņa	2,694		11,948.692	...
9	dzīvot	2,349		10,418.515	...
10	gads	2,212		9,810.879	...

	Lemma	Absolute Frequency		Per Million	
11	Rīga	2,134		9,464.926	...
12	universitāte	1,935		8,582.301	...
13	ar	1,924		8,533.513	...
14	studēt	1,772		7,859.348	...
15	saukt	1,718		7,619.842	...
16	mēs	1,690		7,495.653	...
17	uz	1,512		6,706.17	...
18	bet	1,465		6,497.711	...
19	no	1,458		6,466.664	...
20	ļoti	1,419		6,293.688	...

Biežāk lietotie lietvārdi



10 biežāk lietotie lietvārdi: *gads, Rīga, universitāte, ģimene, brālis, draugs, māsa, māja, tēvs, medicīna*

1. dekl.	2. dekl.	3. dekl.	4. dekl.	5. dekl.	6. dekl.	Nelokāmie
gads	brālis	alus	Rīga	universitāte	brokastis	kino
draugs	dzīvoklis	Centrāltirgus	medicīna	ģimene	zivs	suši
tēvs	kaķis	tirgus	māja	māte	valsts	Mateo
Stradiņš	auglis	medus	Vācija	draudzene	nakts	Frančesko
laiks	suns		māsa	minūte	pusnakts	Oto
vārds	latvietis		nodarbība	šokolāde	Daugavpils	Hindi
students	dārzenis		valoda	studente	sirds	Andre
bērns	kartupelis		nedēļa	zāle	Ventspils	Kivi
ēdiens	Jānis		vakariņas	ārste	acs	Oslo
ārsts	ūdens		Latvija	ballīte	debess	Miko

Tīmeklis 2020 (*CommonCrawl*-LV)



- *Common Crawl* fonds ir bezpeļņas organizācija, kas nodarbojas ar tīmekļa lapu apstrādi un arhivēšanu visā pasaulē.
- *CommonCrawl*-LV ir latviešu valodas tīmekļa tekstu korpuss, kas izgūts no *CommonCrawl* arhīva.
- Korpuss ir automātiski morfoloģiski marķēts un indeksēts *noSketchEngine* rīkā.
- Korpuss satur 5,5 miljardus tekstvienību no 12 miljoniem tīmekļa lapu.
- Pieejams: <https://ej.uz/Timeklis2020>

Tīmeklis 2020 (*CommonCrawl*-LV)

***mazgātava** / mazgātuve*

2990

1701

āvošais pie bāra letes... vai cilvēks, kuru esi satīcis trijos naktī veļas	mazgātavā	, kurā ir pašapkalpošanās. Tas ir stāsts par atturīgās Albertas un ne
ims graužošs, nostāvu rindā ~ 1,5h, tieku līdz iebraukšanai un VISS!	Mazgātavai	3 līnijas, 2 tukšas, bet neviens neko nedara. Piegāju pie maiņas va
iv labojams. Un kāpēc gan man būtu kas jālabo? Ne tāpēc dzenu uz	mazgātuvī	. Saimnieks solīja risināt, bet tā arī neatzvanīja. Ja gribat nesabojāt
bremzēm, priekšā un apauguši ar bremžu putekļiem - diezgan cītīgi	mazgātuvē	tos diskus pūtu, bet tumši brūnā kārtā tikai pa pusei nogāja, tos nāk
; valdes loceklis. Rīgā jau darbojas viena LaserWash auto robotizētā	mazgātava	- tā atrodas Auto Rings centrā, Senču ielā 2F. " Esam gandarīti, ka

***pašbilde** / pašfoto / selfijs / pašņiņš*

1151

83

3519

77

(Tīmeklis 2020)

0

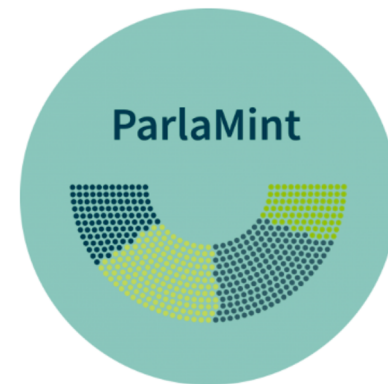
0

24

0

(LVK2018)

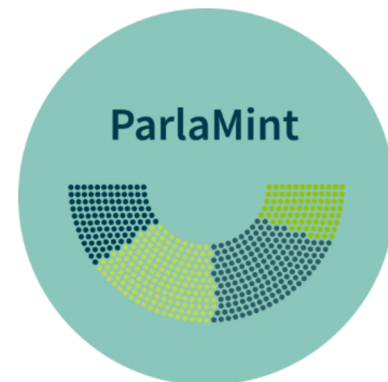
ParlaMint



- ParlaMint ir salīdzināmais parlamentāro datu korpus, kas ietver debates no **18 dažādiem parlamentiem**.
- Korpusa Latvijas daļā ir iekļauti dati no 2015. gada oktobra līdz 2020. gada jūlijam. Dažām valstīm pieejami arī senāki dati.
- Korpus ir morfoloģiski un sintaktiski marķēts, tajā ir marķētas arī nosauktās vienības (named entities).
- Dati:
<https://www.clarin.si/repository/xmlui/handle/11356/1405>
- noSketchEngine: <https://www.clarin.si/noske/parlamint.cgi/>

ParlaMint

Datu lejupielāde



Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.0

“ Please use the following text to cite this item or export to a predefined format:

BIBTEX

CMDI

Erjavec, Tomaž; et al., 2021, *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.0*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1405>.



 This resource is also integrated in following services:

 Share:  

KonText



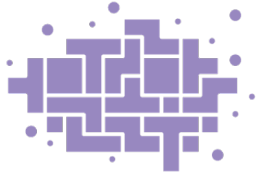
noSketch



Name	ParlaMint-LV.ana.tgz
Size	371 MB
Format	Unknown
Description	Latvian corpus
MD5	f8e58a40ae265531c41b7ef4dfef8f3f



 Download file



Kas ir Tēzaurs: tezaurs.lv



Pavasara versija 2021

328 944 vārdi
48 692 frāzes

Apkaime
tezaurācija
tēzaurācija
tezaurēšana
tēzaurēšana
tezaurismoze
tezaurizācija
tezaurizēt
tēzaurs
tēze
tēzgāt
tēznes
Težu
tēžveida

tēzaurs

tēzaurs vīriešu dzimtes 1. deklinācijas lietvārds +

1. Vārdnīca, kas aptver visus kādas valodas vārdus; ļoti liela, plaša vārdnīca.

2. Plaša sinonīmu, antonīmu un citādi semantiski saistītu vārdu vārdnīca.

AVOTI: MLVV, ODE

KORPUSA PIEMĒRI: +

Dalīties

Ziņot

Mašīnlasāma vārdnīca

Reizi ceturksnī tiek publicēta jauna, papildināta versija

ķīķerēt

ķīķerēt 2. konjugācijas darbības vārds; sarunvaloda, intransitīvs +

1. Vērot, skatīties, arī censties ko saskatīt (parasti slepus).

// apvidvārds Uzmanīgi skatīties.

2. apvidvārds Skatīties ar tālskati.

AVOTI: LLV, ME

KORPUSA PIEMĒRI: +

Īstenības izteiksme:

	Tagadne		Pagātne		Nākotne	
	Vsk.	Dsk.	Vsk.	Dsk.	Vsk.	Dsk.
1. pers.	ķīķerēju	ķīķerējam	ķīķerēju	ķīķerējām	ķīķerēšu	ķīķerēsim
2. pers.	ķīķerē	ķīķerējat	ķīķerēji	ķīķerējāt	ķīķerēsi	ķīķerēsiet, ķīķerēsīt
3. pers.		ķīķerē		ķīķerēja		ķīķerēs

Pavēles izteiksme: ķīķerē (vsk. 2. pers.), ķīķerējiet (dsk. 2. pers.)

Atstāstījuma izteiksme: ķīķerējot (tag.), ķīķerēšot (nāk.)

Vēlējuma izteiksme: ķīķerētu

Vajadzības izteiksme: jāķīķerē

« Uz tiem ilgi un pamatīgi ķīķerēja Rūķis Ceļvedis. »

Vladimirs Kaijaks, Melnais kamols-amols-mols. Rīga, Zvaigzne ABC, 2003. (Daiļliteratūra)

« Viņa rāva Billi prom, sacīdama: — Nav ko ķīķerēt. »

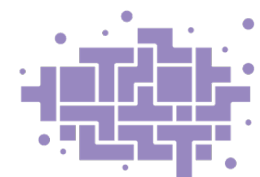
Vizma Belševica, Bille. Rīga, Jumava, 1995. (Daiļliteratūra)

« Pat vakaros ķīķerēju kā vecs ērms. »

Gundega Repše, Stāsti par mācekļiem (Jūras velni). Rīga, Dienas Grāmata, 2009. (Daiļliteratūra)

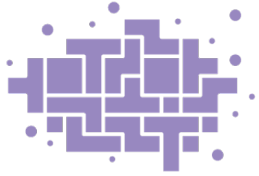
Piemēri no latviešu valodas tekstu korpusa ir atlasīti automātiski un var būt neprecīzi.

Vairāk piemēru ...



Latviešu valodas leksiskais tīkls (WordNet)

- Tēzaura pārstrādāšana, koncentrējoties uz nozīmēm
 - Korpusa datos balstīts nozīmju sadalījums
 - Nozīmēm piemeklēti ilustratīvi lietojuma piemēri
 - Nozīmes ar skaidrākām robežām, kas nepārklājas
- Jauna platforma redigēšanai ar strukturētākiem datiem
- Uzlabota meklēšana – vārdformas, vārdu līdzība
- Leksiskais tīkls – formālas saites starp nozīmēm
- Nozīmju piekārtojums angļu valodas *Princeton WordNet* datiem



"Sinonīmu kopa"

☒ Rādīt visas saites

Sinonīmu kopa



spēlēt₂, tēlot₁

spēlēt₂ Atveidot, tēlot (lomu drāmas daiļdarbā vai filmā); īstenot skatītāju priekšā (iestudētu izrādi).

tēlot₁ Atspoguļot, atveidot (ko) mākslas tēlos, mākslas darbā; atspoguļot, atveidot (ko) - par mākslas darbu.

SINONĪMU VĀRDNĪCA

spēlēt - muzicēt, atskaņot, strinkšķināt, strinkšināt, trinkšķināt, trinkšināt, staipīt, virgāt, lirināt, pļerkstināt, bungot, plinkšķināt, klimpināt, klimšķināt, raudināt, čīgāt, džīgāt, vīlēt, zāgēt, ērgelēt, plēst vaļā, raut vaļā

tēlot - spēlēt, atveidot

TULKOJUMI

spēlēt - play, game, finger, pick, play, act, perform, feature in

SAITES



ĀRĒJĀS SAITES:

[LV plašāka nozīme](#)



(v) act, play, represent

play a role or part; "Gielgud played Hamlet"; "She wants to act Lady Macbeth, but she is too young for the role"; "She played the servant to her husband's master"

[LV plašāka nozīme](#)



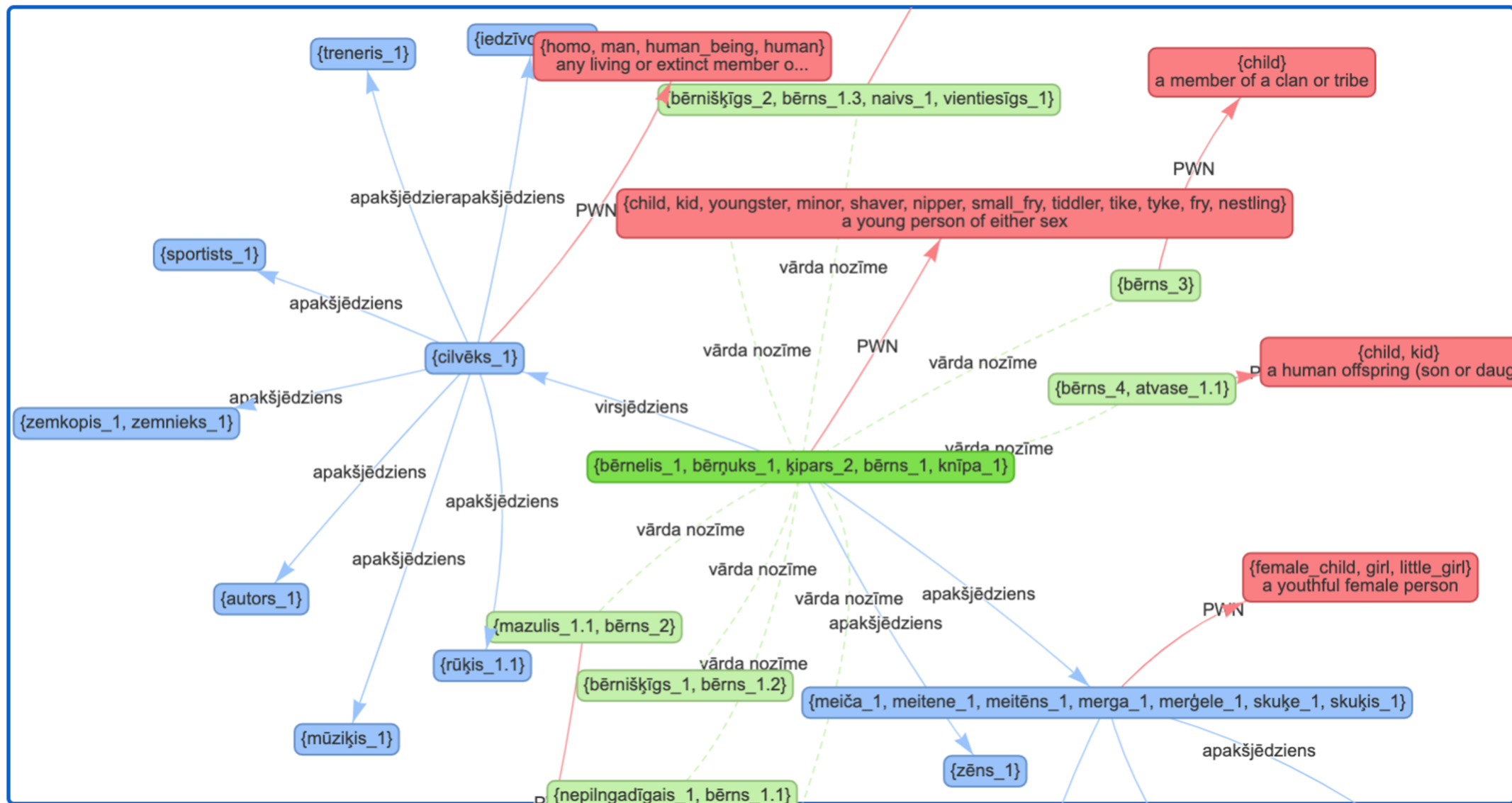
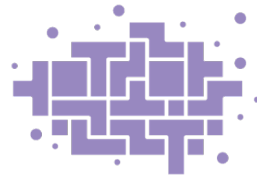
(v) act, play, roleplay, playact

perform on a stage or theater; "She acts in this play"; "He acted in `Julius Caesar'"; "I played in `A Christmas Carol'"

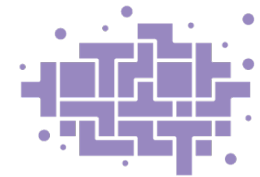
Meklēt

☒ Tikai nozīmes ☐ Ar apakšnozīmēm ☐ Angliski ☐ Angliski*

Latviešu valodas leksiskais tīkls (WordNet)



Latviešu valodas leksiskais tīkls (WordNet)



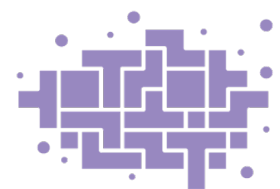
- LŽP projektā galvenais lietojums ir vārdu nozīmju nošķiršana automātiskās teksta analīzes sistēmās
- “Trīs vienā” vārdnīca – skaidrojumi, sinonīmi, tulkojumi
- Šķiet derīgs arī tulkotājiem un valodas apguvējiem
- Tīks publicēts Tezaurs.lv lapā gada otrajā pusē
 - Izstrādes procesā, pašreiz 2500 nozīmju
 - Kā šo sarežģīto informāciju skaidrāk parādīt?
- Nākotnes mērķis – meklēt korpusā nozīmju lietojumus



Ieskaties: <http://ailab.lv>

Seko mums: @AiLab.lv

Paldies par uzmanību!



Projekts "Latviešu valodas apgūvēju korpusa izveide: metodes, rīki un izmantojums" (Nr. Izp-2018/1-0527)

Projekts "Latviešu valodas WordNet un vārdu nozīmju nošķiršana" (Nr. Izp-2019/1-0464)

Valsts pētījumu programma "Latviešu valoda" (Nr. VPP-IZM-2018/2-0002)

Valsts pētījumu programmas "Humanitāro zinātņu digitālie resursi» projekts "Humanitāro zinātņu digitālie resursi: integrācija un attīstība» (Nr. VPP-IZM-DH-2020/1-0001)