

Latvian WordNet

Pēteris Paikens, Agute Klints, Ilze Lokmane,
Lauma Pretkalniņa, Laura Rituma, Madara Stāde, Laine Strankale

[peteris@ailab.lv]



Latvian WordNet

- Release of the first WordNet for Latvian
 - "1.0" release now, a work-in-progress was presented at LREC 2022
- Wrapping up a 3-year project 2020 – 2022
- Fully manually curated for most frequent words in Latvian
- 11 000 senses from 7 600 words linked in 6 500 synsets

Motivation

- Resource for building Natural Language Processing systems
 - Natural language understanding, semantic parsing
 - Word sense disambiguation
 - Natural language generation, especially in context of multilingual systems
- Facilitate Latvian linguistic research of semantics and lexicography
- Improvement of Tēzauris.lv online dictionary
- Bringing Latvian to multilingual tools and resources

Latvian semantic resources before this project

- "FullStack" annotated corpus with multiple layers
 - FrameNet
 - PropBank for the frame target words
 - Morphology, syntax, named entities, coreference
- Digital dictionary Tēzaurs.lv that unites most dictionaries of Latvian
- A relatively old synonym dictionary

- No WordNet
- No valence dictionaries

Goals and priorities

- High quality core WordNet
 - Start with most frequently used words (nouns, verbs, adjectives, adverbs)
 - Basis for bootstrapping for larger coverage
 - Usable for both linguists and end-users
- Based on corpus evidence to reflect the lexical system of Latvian
 - Existing Latvian dictionaries were based on linguistic intuition
 - Avoid copying English concepts and hierarchy
- Mapping to other WordNets
 - Princeton WordNet
 - Open Multilingual WordNet

Data sources for building Latvian WordNet

- Tēzaurus.lv online dictionary
 - Largest dictionary for Latvian, 384 000 entries
 - Fully electronic, structured entries
 - Summarized many dictionaries, so a lack of consistency
- Latvian National Corpora collection korpuss.lv
 - Balanced contemporary corpus of Latvian (10M tokens)
 - Common Crawl web corpus (400M tokens)
 - Blogs and parliament debates

Methodology


- Rebuild word sense inventory for the word based on corpus evidence
 - Consistent principles and granularity
 - Two level hierarchy – senses and subsenses
- Assign corpus examples (10-20 per sense)
 - Validation of sense inventory and sense separation/overlap
- Link synsets and semantic relations
 - similarity, hyponymy, antonymy, meronymy, gradation sets
- Map to Princeton WordNet
 - Not only full equivalence, but also interlingual hyponymy if no exact match found

Sense annotation view

Show all subsenses Show all examples

 spēlēt


 spēlēt 2nd conjugation verb; transitive  

1.  Veikt noteiktu darbību kopumu (spēli), kam ir sacensības pazīmes un ar ko cenšas sasniegt vēlamo rezultātu, izmantojot prasmes, iemaņas, arī apstākļu nejaušu sakritību; gūt prieku, izklaidēties.

✓ Examples *Dižistabā vīri laikam spēlēja kārtis.*

✓ Translations *play*


Show subsenses

1.1.  Veikt darbību kopumu (spēli) ar mērķi sagādāt prieku, izklaidēšanos, kam parasti raksturīga iztēlē radīta situācija un darbības objekti, kādu norišu, cilvēku, dzīvnieku u. tml. atdarināšana; rotaļāties.

✓ Examples *Nu vairs neviens neiedrošinās, piemēram, paslēpes spēlējot, turēt acis vaļā un blēdīties.*

✓ Related senses *rotaļāties*

✓ Translations *play*

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

✓ Examples *Norberts teātra izrādē spēlē sievieti*

✓ Related senses *tēlot*







✓ Multiword Expressions *Spēlēt kumēdiņus. Spēlēt teātri.*

✓ Translations *act, play, represent, roleplay, playact*

 Show other dictionaries

Query lemma word form

Corpus

- *Un pa ielu **spēlēdami**, dziedādami nāk jauni puisi un viņu vidū liela, skaista meita — melnie, vilņainie mati pār pleciem, seja balta, lūpas sarkanās kā tā saule.* 
- *Dižistabā vīri laikam **spēlēja** kārtis.* 
- *Bet nu papucis istabā sāka kliegt par blēdīšanos un piespēlēšanu, un Voldis blāva, ka pats nemāk **spēlēt** un blēdis, gāzās kresli, šļūkāja soļi, kaut kas mīksti būksķēja.* 
- *Mammucis ar Loniju plunčīnājās virtuvē, Billei pieteikts nekur vairs nekustēt — noplēšīšoties līdz kaulam, kā tad uz Rīgu braukšot, bet vīri atkal vienā mierā dzēra alu, tikai kārtis vairs **nespēlēja**, līdz kamēr vecāmāte negribīgi ieminējās, ka nu vajadzētu kādam patecēt pēc Odaļas.* 
- *Tur Jančelis un Bille **spēlēja** "pašiem savu māju", kad citu nebija sētā.* 
- *Nu varētu **spēlēt** nekāpšanu uz strīpām visu ceļu, bet acis, asaru aizvilktas, neļāvās saskatīt, kur strīpas, kur nē.* 

...

bērnelis₁, bērņuks₁, ķipars₂, ute₃, bērns₁, knīpa₁, kverpis₁

bērnelis₁ Bērns.

bērņuks₁ Bērns.

ķipars₂ humoristiska ekspresīvā nokrāsa Bērns.

ute₃ sarunvaloda Bērns.

bērns₁ zēns vai meitene (aptuveni līdz 14 gadu vecumam).

knīpa₁ Maza meitene, mazs zēns.

kverpis₁ Bērns.

SYNONYM DICTIONARY

bērns - mazulis, mazais, bērņuks, bērnelis, ķipars

TRANSLATIONS

bērns - preadolescent, wean, bairn, youngling, babe, child, children, tad, kid, trick, baby, infant, fruit of the womb

LINKS

EXTERNAL LINKS:

(n) child, kid, youngster, minor, shaver, nipper, small_fry, tiddler, tike, tyke, fry, nestling

a young person of either sex; "she writes books for children"; "they're just kids"; "tiddler" is a British term for youngster"

HYPONYMS:

knīveris₁, mazpuika₁, puika₁, puišāns₁, puiškins₁, zēns₁, zeņķis₁, zeperis₁, puisis_{1,2}

knīveris₁ Zēns, puisēns.

mazpuika₁ Zēns.

puika₁ Zēns.

puišāns₁ Zēns.

puiškins₁ Zēns.

zēns₁ Vīriešu dzimuma bērns (aptuveni līdz 11 gadiem); arī pusaudzis.

zeņķis₁ Zēns, aptuveni skolas vecumā; arī pusaudzis.

Link annotation view

meitene

Only senses With subsenses English English*

SENSES

jaunmeitene₁

Jauniete.

Ganu meitene₁

ganumeita.

zemiņkmeitene₁

Meitene, kas aug zemiņku ģimenē; arī zemiņkmeita.

zvejniņkmeitene₁

SYNSETS

jaunekle₁, jauniete₁, meitene₂, skuķis_{1,1}, meitēns₂, jaunmeita₁, mamzele₁

jaunekle₁ Jauniete.

jauniete₁ Sieviete vecumā starp pusaudzes un brieduma gadiem.

meitene₂ Jauniete.

skuķis_{1,1} Nepieredzējusi, arī nenopietna jauniete.

meitēns₂ Meitene (2).

jaunmeita₁ Jauna meitene.

mamzele₁ novecojis Jaunkundze.

LINKS

EXTERNAL LINKS:

(n) girl, miss, missy, young_lady, young_woman, fille

a young woman; "a young lady of 18"

HYPERONYMS:

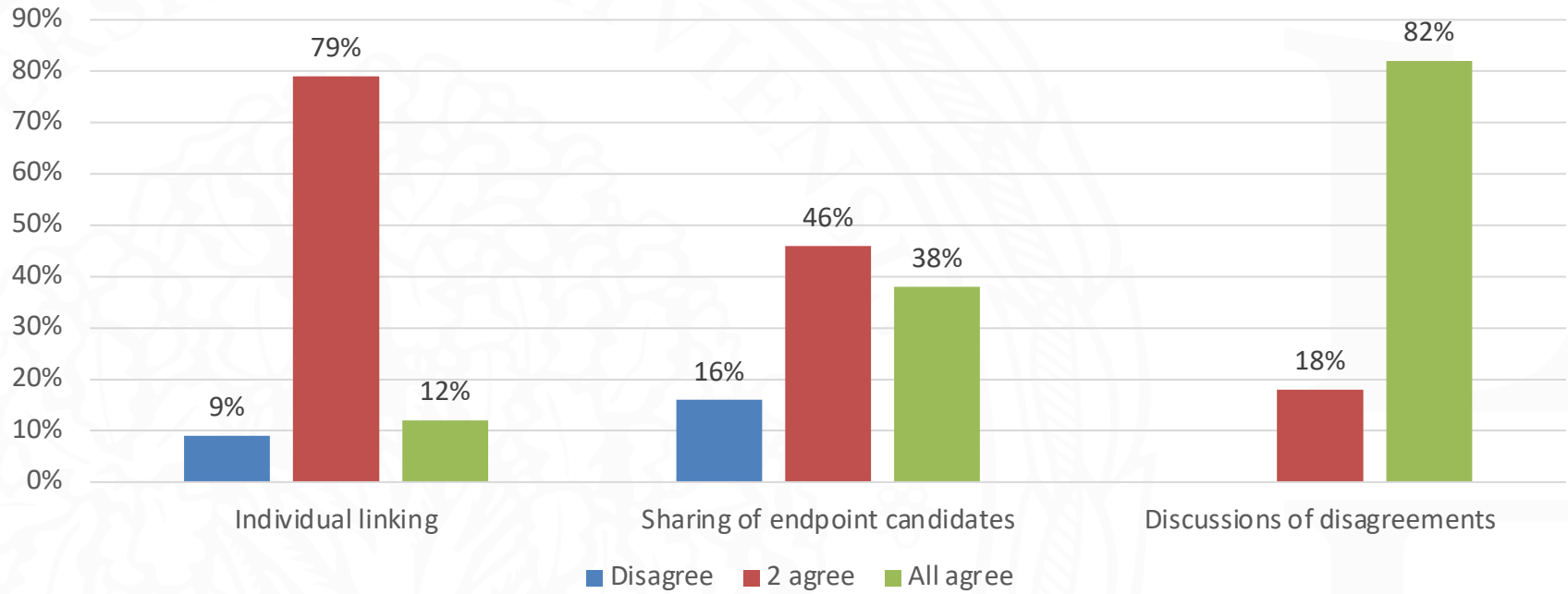
sieva₂, sievietē₁

sieva₂ Sieviete.

sievietē₁ Cilvēku dzimuma būtne, kuras organisma morfoloģiskās un fizioloģiskās īpašības ir piemērotas bērnu dzemdēšanai; pieaugusi šāda cilvēku

Evaluation of semantic linking

Three experiments on links from 20 words (85 senses)



IAA experiment for interlingual links

degsme₁

Dedzīga, kaisla, kvēla aizraušanās; aizrautība, dedzība.

Nē

Vajag vairāk info

Šaurāks/plašāks

No match

Need more data

Broader or narrower

(n) ardor, ardour, elan, zeal

#1 a feeling of strong eagerness (usually in favor of a person or cause); "they were imbued with a revolutionary ardor"; "he felt a kind of religious zeal"

(n) ardor, ardour, fervor, fervour, fervency, fire, fervidness

#2 feelings of great warmth and intensity; "he spoke with great ardor"

(n) heat, warmth, passion

#3 the trait of being intensely emotional

(n) gusto, relish, zest, zestfulness

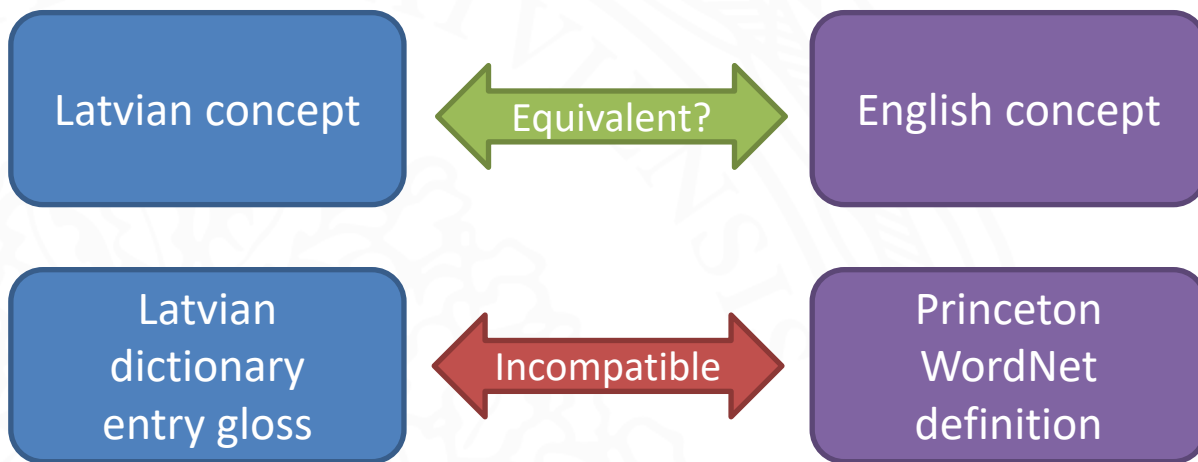
#4 vigorous and enthusiastic enjoyment

(n) soul, soulfulness

#5 deep feeling or emotion

57% interannotator agreement (390/684 words)

Problem with overspecific definitions



- bass guitar "the guitar with six strings that has the lowest pitch"
- apache "a Parisian gangster" vs *apašs* "a French gangster"

Extending Tezaurs.lv as "3 in 1" dictionary

baznīca

baznīca sieviešu dzimtes 4. deklinācijas lietvārds ▼ Locīšana

1. Kristīgo konfesiju reliģiskajam kultam paredzēta celtne.

▼ Piemēri *Bille ieskatās — cauri kokiem saredzama balta baznīca ar zillem un zaļiem torņiem.*

^ Saistītās nozīmes

Sinonīmi ⓘ

dievnams₁ — Baznīca. lūgšanas nams.

Hiponīmi ⓘ

svētnīca₁ — Reliģiska kulta celtne, telpa, arī vieta; arī templis, baznīca, dievnams.

templis₁ — Sakrāla celtne reliģiskajiem rituāliem; arī baznīca, mošeja, pagoda, sinagoga u. tml.


Hiperonīmi ⓘ

celtne₁ — Celtniecības procesā izveidots objekts vai objektu komplekss; arī ēka.

ēka₁ — Celtne, ko parasti izmanto dzīvošanai, saimnieciskām, ražošanas vai sabiedriskām vajadzībām; atsevišķs objektu celtnu kompleksā.

▼ Stabili vārdu savienojumi *let baznīcā.*

^ Tulkojumi

church, church building 

a place for public (especially Christian) worship; "the church was empty"

[Princeton WordNet 3.0]

2. Vienas konfesijas ticīgo organizācija ar noteiktu dogmatiku un kultu.

▼ Piemēri *Viņi iekasēja desmito daļu no visa mužiku nopelnītā un algas dienā saņemtā — kā katoļu baznīca viduslaikos.*

- Explanatory dictionary
 - Existed already
- Synonym dictionary
 - WordNet semantic links
- Translation dictionary
 - Links to Princeton WordNet
 - Sense-to-sense translation
- Widely used in Latvia
 - Students
 - Language learners
 - Translators, linguists

Results and impact

- 11 227 senses from 7 609 words linked in 6 515 synsets
- Served to general public via Tezaurs.lv dictionary platform
- Available as open data via Clarin and on wordnet.ailab.lv
- Basis for ongoing WSD tool development for Latvian
 - 75 247 usage examples linked to senses
- Advanced Latvian lexicography and semantics

Conclusions

- Largest task - review of sense inventory
 - Overlap, consistency, granularity of existing dictionaries is problematic
 - The new corpus-based sense inventory seems much better
 - Very time consuming
- Working topic-by-topic might have been more efficient
- Annotation is highly subjective, teamwork helps
- Many options for future work
 - Extending both manually and with semi-automated methods
 - Additional types of semantic links
 - Multiword expressions
 - Derivation relations

Thank you!

Questions?

peteris@ailab.lv
wordnet.ailab.lv