Automatic Semantic Tagging of Estonian Spatial Adverbials for Valency Pattern Mining

Kertu Saul, Sven Laur,
Kadri Muischnek, Jelena Kallas

28 August 2025

Valency patterns

- Valency pattern = lexical verb + abstract arguments that relate to the verb's meaning

  throw: AGENT PATIENT GOAL
  throw: subj obj to+obl/obl+all
  Anna **threw** the ball to Kevin.
  Anna **viskas** palli Kevinile.

- Large languages manually compile these in dictionaries or other similar resources (i.e. Framenet)
- Small languages need to **automatically mine these from corpora**

Estonian (and why we need semantic tagging)

- ~ 1 million speakers
- Free word order
- Morphologically very rich, **14 nominal cases**
- Cases are **very polyfunctional**: 2-13 functions per case
- Arguments can be in most cases

Lubja-l	kadu-s	eile	õhtu-l	80 protsendi-l	elanike-l	elekter.
Lubja-ADE	disappear-3SG.PST	yesterday	night-ADE	80 percent-ADE	resident.PL-ADE	electricity

“Electricity disappeared for 80 percent of residents in Lubja yesterday night”

Automatic semantic tagging

- Focus on tagging **nominal adverbials with spatial meaning**
- Look at adverbials in **6 “spatial” cases**: allative (onto), adessive (on), ablative (from on), illative (into), inessive (in), elative (from in)
- Differentiate between real spatial usage vs using a "spatial" case for coding other semantic types
- Data from morphosyntactically annotated Estonian Reference Corpus (245 million words)

- Test two methods:
  - **LLMs**
  - **Verb-case patterns**: can adverbials be semantically tagged by only knowing their case and head verb

Method 1: LLMs

- Detecting **physical locations**
- Test-set of **1000 adverbials in spatial cases** + sentence for context
  - 10 tags: physical location, abstract location, event, time, manner, state, owner, reason, dependent, other, error

- **Annotation guide** as main basis of the prompt
- Model: **GPT-4o**
- Accessed through Open-AI’s API
- **Word and sentence as input** from csv file
- Asked if this word in this sentence is a physical location or not
- Zero-shot approach

```
I am a linguist. You are a linguist, who helps detect physical locations. Determine whether the word "{row['form']}" in
the following sentence "{row['sentence']}" is a physical location based on the following categories:

Physical locations:
1. Place names (e.g. Bristol, Sepphoris)
2. Buildings/physical locations of businesses (bankhouse, multimedia studio, club Kuku, computer company)
3. Physical objects, including living beings (first place podium, the Moon, backup device, saddle, cloud)
4. Areas with a definable geographical location (scene of the fire, the North Pole, shoreline, cloud of dust)

Not physical locations:
1. Abstract locations, whose geographical location can't be determined (e.g. Wifi, computer market, airspace, digital
platform).
2. Activities and events (dress rehearsal, recruitment).
3. Living being who's the performer of the action (slippers went for the scrambler).
4. State (run legs to blisters, sit in shit).
5. Manner adverbials (most acutely, hand in hand).
6. Reason adverbials (in case of destruction, in the existence of a processor).
7. Time adverbials (year, morning).
8. Constructions and expressions (despite the attitude, talking about validity).

Word: {row['form']}
In sentence: {row['sentence']}

Answer in the format:
- If a word is a physical location: "{row['form']}|{row['sentence']}|LOC"
- If a word isn't a physical location: "{row['form']}|{row['sentence']}|NONE"

NB! ALWAYS ONLY answer in the form LOC or NONE
```

Results

- **Recall 0.93, Precision 0.78, F-score 0.85**

- **Should have been tagged as physical locations but weren't (FN)**:
  - Organisations
    - *I went to the modelling school yesterday* - physical location
    - *I go to modelling school* - abstract location
  - Brand names (*I'm sitting on an Aeron*)
  - Uncapitalised place names (*I've never been to piibe*)
- **Shouldn't have been tagged as physical locations but were (FP)**:
  - 60% abstract locations
  - 13% typos (went to pdagogic university [sic])
  - 13% events (sit at a sculpting class)
  - 9% constructions (repairs are moving **thanks to** a new machine).

Method 2: verb-case patterns

- **Hypothesis**: there is a significant amount of instances where all of a verb's dependents in a specific case belong to the same semantic class
  - travel **to**: Africa ✓, him ✗, a CD ✗, pieces ✗
  - listen **to**: Africa ✓, him ✓, a CD ✓, pieces ✓

- These patterns could be used to **semantically annotate all of a verb's dependents in said case**
  - travel to: *Africa*, *piibe*, *an area* - all locations

- Words with **one semantic type across all patterns** could be annotated with that type across **the entire corpus**
  - area: location, location, location - ✓
  - Africa: location, organization, object - ✗

EESTI
KEELE
INSTITUUT

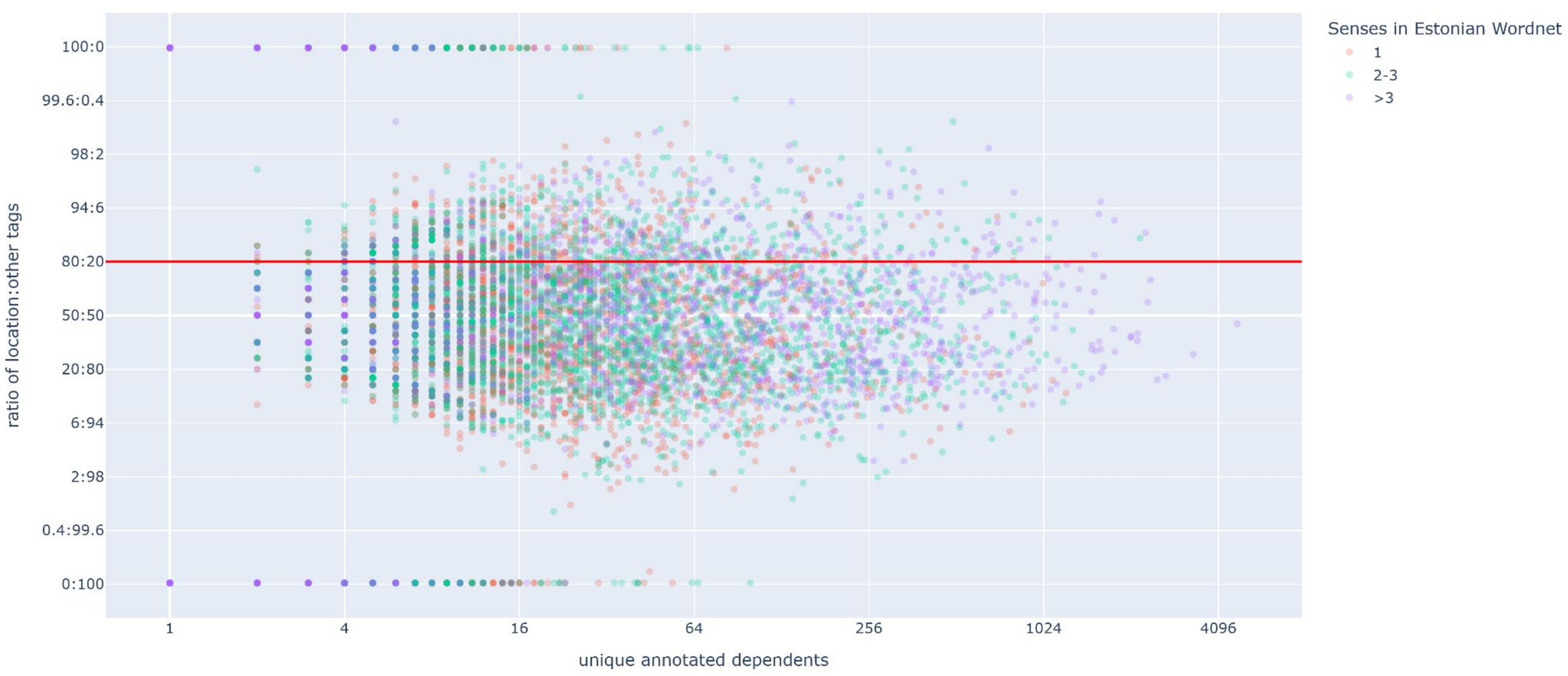
Step 1: preliminary semantic tagging

- Using a **semantically tagged dictionary**
- 128 semantic types of **various specificity** (time, time_month, time_ADV etc)
- Combine into general types
- Create **wordlists** for 5 semantic types: location, time, state, event, not_location
  - only include words with one semantic type

- **Annotate nominal adverbials** in spatial cases in the Estonian Reference Corpus using these wordlists
  - Unique adverbials annotated: 23,979 out of 245,358 aka **9.8%**
  - Repeating adverbials annotated: 2.3M out of 7.8M aka **28.76%**

Step 2: statistics

1. Count per verb + case combination **how many annotated dependents were locations** or had some other tag
    - kuuluma + ill (*belong into*): location = 1165, other_tags = 1347
2. Calculate **relative frequency** for location tag and all other tags
    - kuuluma + ill: location 46,4%, other_tags 53,6%
3. Calculate **logarithmic fold change** for plotting
    - kuuluma + ill: log2(0.464/0.536) = -0.2094
4. Count how many annotated dependents were **unique** words
    - kuuluma + ill: 261


Senses in Estonian Wordnet
1
2-3
>3
ratio of location:other tags
100:0
99.6:0.4
98:2
94:6
80:20
50:50
20:80
6:94
2:98
0.4:99.6
0:100
1
4
16
64
256
1024
4096
unique annotated dependents

Results

- 3992 aka **18.8%** of verb-case patterns only have dependents with the location tag
  - These patterns have **~45000 unannotated dependents** combined which can now be annotated as locations
- Variability of semantic types in a pattern is **not correlated with how many senses a verb has**
- In patterns above the 80:20 ratio line, other tags actually occurred either very peripherally or **were there due to incorrect morphological, syntactic or semantic tagging**
  - Accounts for additional **10% of patterns or ~643000 unannotated dependents**
- Some patterns were systematic mistakes of the Estonian syntactic parser

Conclusions

- **GPT-4o works well** for semantic annotation, even in smaller languages
- Some semantic types have to be **explained in the prompt more than others**
- Around **30%** of verb-case patterns take dependents in a single dominant semantic type
  - Out of these, patterns above the 80:20 ratio line require additional analysis before use in tagging
- **Same method can be applied in other languages** when the language:
  - encodes adverbials with cases and/or adpositions
  - has access to a limited semantic dataset

Thank you for listening!