EESTI KEELE INSTITUUT

# Using dictionary definitions to identify the semantic profile of an open slot in a construction

Ene Vainik & Heete Sahkai

Workshop “From a dictionary to a constructicon” 19.04.2024
21st Annual Conference of Applied Linguistics (18.–19.4.2024, Tallinn, Institute of the Estonian Language)

# Outline

- Background
- Our goal and the expected benefits of the proposed method
- A case study – the Estonian Translative Inchoative Construction
- Methods and procedures
- Results
- Summary and discussion

# Background

- The properties of a schematic construction, like its meaning and productivity, determine the class of elements that can appear in an open slot of the construction.
- Consequently, identifying the semantic profile of the class of words that appear in an open slot of a construction in corpus data helps to:
    - Identify the meaning of the construction
    - Identify the productivity of the construction
    - Identify the development of the construction
    - Describe the construction in a constructicographic resource
    - Etc.

# Methods for categorising the words that appear in a construction in corpus data

- Categorisation based on semantic inventories like event types (Goldberg 1995:39, Levshina 2016), verb classes (Barðdal 2008:63-68, Dekalo & Hampe 2017), or FrameNet (Fillmore et al. 2002) frames (Sundquist 2020, Bonial 2014)
- Categorisation using semantic similarity measures based on distribution (word embeddings) or semantic relations (Word Net), e.g.:
    - cluster analysis (Perek 2016, 2018, Dekalo and Hampe 2017) or network science (Ellis et al. 2014, 2016, Dekalo & Hampe 2017, Cheng-Hsien Chen 2020) based on similarity measures
    - cluster analysis based on collocates and covarying collexemes (Gries & Stefanowitsch 2010)
- Categorisation based on collostructional strength (Stefanowitsch & Gries 2003)

# Goal & the expected benefits

To test an additional method for describing the semantic profile of a constructional slot, using semantic descriptors gleaned from dictionary definitions.

**Expected benefits of the method:**

- Making use of the rich data in existing lexicographic resources.
- Objectivity: DDs are constructed by different lexicographers at different times, independently.
- There may be aspects of meaning that are relevant to the meaning and productivity of a construction but are not captured by semantic inventories, distributional similarities or semantic relations. Such aspects of meaning could be gleaned from dictionary definitions, which go beyond recording the broad semantic type and semantic relations of a lemma.

# Expected benefits (continued)

- Unlike semantic inventories, the method can be applied to any set of words defined by a construction, not only to words belonging to a particular syntactic category (e.g., verbs) or semantic domain (e.g., events).
- Unlike cluster analysis, the method can provide the categories emerging from the analysis with semantic labels.
- In view of a constructicographic resource, it allows to extract the semantic definition of a construction from the definitions of the words occurring in it.

# Related work

Previous work using dictionary definitions for the purpose of semantic classification:

- Kazeminejad et al. 2022 use dictionary definitions among other sources to refine the VerbNet classification by annotating verb-specific features.
- Recski 2018 describes a module that builds concept graph definitions for words by automatic processing of entries of large explanatory dictionaries. The resulting set of definition graphs in turn has been used in measuring semantic similarity of words.

How we differ: our aim is not to establish a general ontology but to characterise a class of words defined by a particular construction

# A case study: the Estonian Translative Inchoative Construction

**[‘go’-3sg + Noun$_{[event]}$-sg.tra]**

**Characteristics:**

Syntax: subjectless complex predicate
Semantics: inchoative aspect
Pragmatics: expressive; colloquial register

# Examples

| | | | | | | |
|---|---|---|---|---|---|---|
| ***Wisla*** | ***ja*** | ***Levadia*** | ***fännide*** | ***vahel*** | *läks* | *lööma-ks*. |
| Wisla.GEN | and | Levadia.GEN | fans.GEN.PL | between | **go.PST.3SG** | **fight-TRA** |

‘A fight broke out between the fans of Wisla and Levadia.’

| | | | | | |
|---|---|---|---|---|---|
| ***Naiste*** | ***vahel*** | ***läks*** | ***tribüünidel*** | ***hirmsaks*** | ***kismaks.*** |
| woman-pl.gen | between | **go-3** | grandstand-pl.ade | fierce-sg.tra | **affray-sg.tra** |

‘a fierce fight of women broke out on the grandstands’.

# Methods

1. Data extraction from corpus and collostructional analysis → type and token frequency; collostructional strength → lining up the central vs peripheral members of the category
2. Exploration of the dd's of the nouns; both central and peripheral members of the category → establishing a set of descriptors
3. Relative prominence analysis of the descriptors → establishing the profile of the category

# Step 1: Data extraction

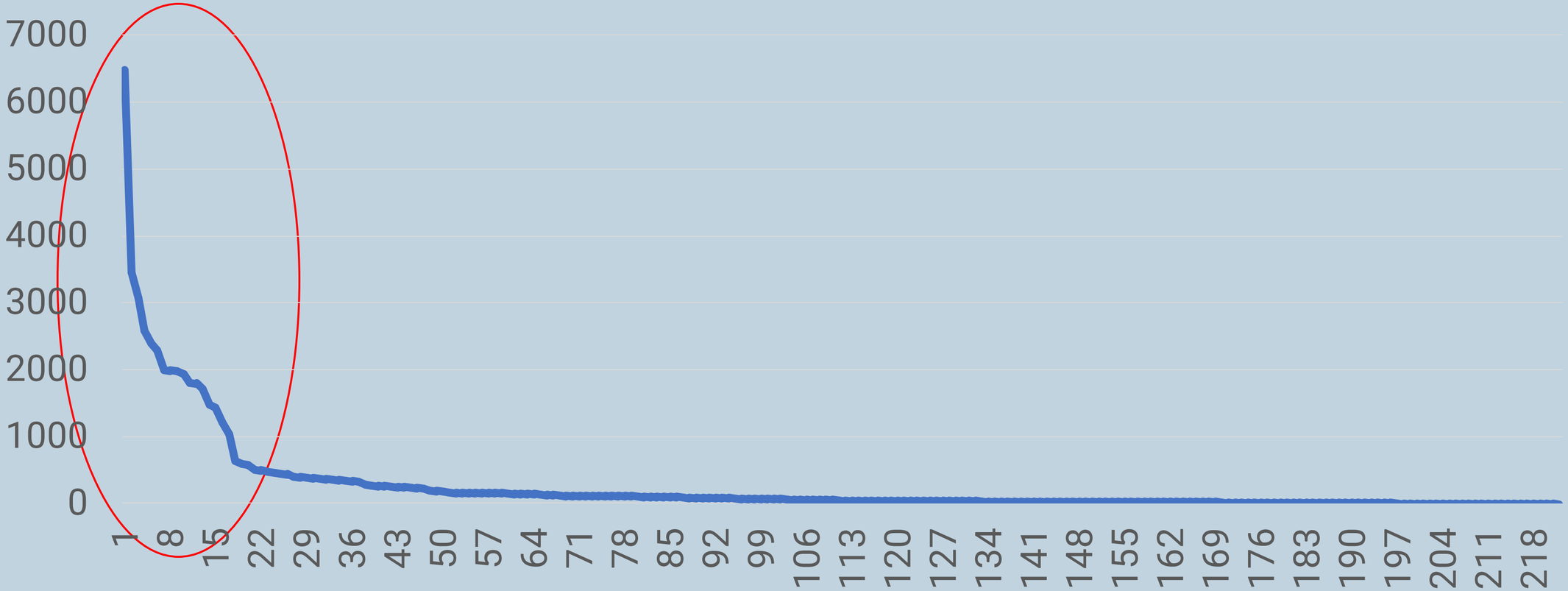
- Estonian National Corpus 2021 (2,945,431,278 tokens), a subcorpus consisting of blogs and forums (839,375,890 tokens)
- Sketchengine.eu corpus query interface

# Results of the data extraction

| Indicator | Value |
|---|---|
| *Token frequency | 20 940 |
| *Type frequency (realised productivity) | 1385 |
| *Hapax legomena | 462 |
| *Hapaxes/tokens (potential productivity) | 0.02 |
| No of the types F≥5 | 552 |
| No of noun lemmas present in EstCombiDic | 232 (42%) |

*Productivity indicators of the construction based on Baayen (2009)

# Distribution of the data by the association strength (LLR) between the construction and the nouns (N=222)



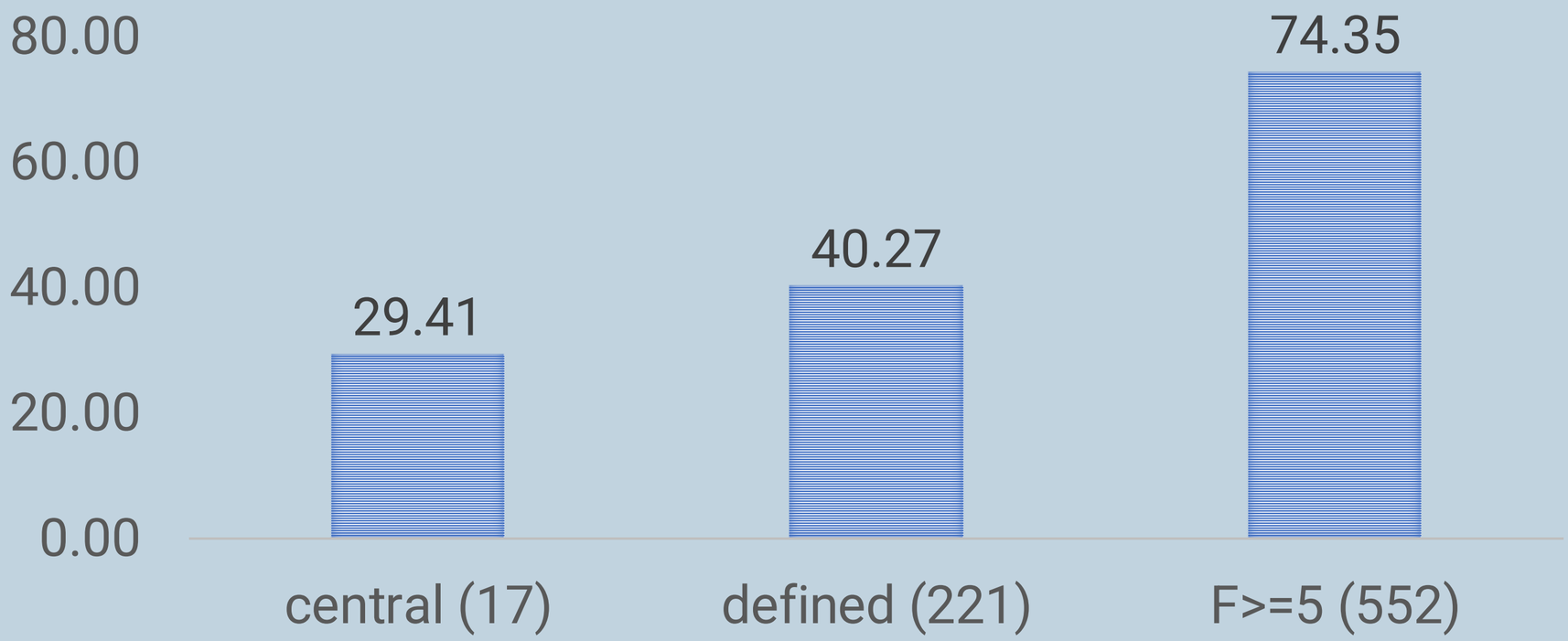

Calculated using Coll.analysis 4.0 (Gries 2022, https://www.stgries.info/teaching/groningen/index.html)

# Results of step 1: Typical instances

- 20 most strongly associated lemmas in terms of LLR

| lemma | N | LLR | lemma | N | LLR |
|---|---|---|---|---|---|
| jama ‘trouble’ | 1121 | 6479.42 | vahetus ‘exchange’ | 422 | 1971.32 |
| ~~andmine ‘giving’~~ | ~~485~~ | ~~3489.24~~ | läbu ‘brawl’ | 206 | 1932.63 |
| tegemine ‘doing, making’ | 969 | 3455.52 | rebimine ‘contest’ | 174 | 1802.71 |
| madin ‘struggle’ | 267 | 3073.24 | möll ‘turmoil’ | 205 | 1795.93 |
| ~~tellimine ‘ordering’~~ | ~~354~~ | ~~2830.31~~ | ~~lammutamine ‘demolition’~~ | 201 | 1754.77 |
| lööm ‘fight’ | 178 | 2581.58 | kisma ‘strife’ | 139 | 1714.08 |
| kaklus ‘fight’ | 250 | 2385.84 | ~~küsimine ‘asking’~~ | 192 | 1534.18 |
| ostmine ‘buying’ | 364 | 2282.92 | vaidlus ‘dispute’ | 222 | 1476.54 |
| sõit ‘drive’ | 464 | 1988.66 | vahetamine ‘exchanging’ | 236 | 1436.32 |
| ost ‘purchase’ | 350 | 1987.44 | ehitamine ‘building’ | 234 | 1352.07 |

# An aspect of productivity – *ad hoc* noun derivation from verbs




% of lemmas ending with the fully productive nominalizing affix *-mine*

# For comparison: 20 most strongly attracted collocates of the regular inchoative verb *hakkama* ‘start’ in the same subcorpus

| Collocate | Freq | log likelihood | Collocate | Freq | log likelihood |
|---|---|---|---|---|---|
| mõtlema ‘think’ | 6747 | 50577.65 | uurima ‘find out’ | 2158 | 16922.89 |
| naerma ‘laugh’ | 3182 | 33622.62 | valutama ‘ache’ | 1307 | 14164.59 |
| nutma ‘cry’ | 2605 | 28344.77 | minema ‘go’ | 3479 | 14149.14 |
| liikuma ‘move’ | 3011 | 24852.18 | mängima ‘play’ | 1976 | 13316.41 |
| rääkima ‘speak’ | 3519 | 22297.99 | käima ‘go’ | 2645 | 13101.93 |
| sadama ‘rain’ | 1805 | 19354.32 | kehtima ‘be valid’ | 1360 | 12672.36 |
| otsima ‘look for’ | 2480 | 17915.66 | kasutama ‘use’ | 2300 | 12204.97 |
| tulema ‘come’ | 4151 | 17451.83 | meeldima ‘like’ | 1945 | 11641.69 |
| tekkima ‘appear’ | 2522 | 17170.17 | toimuma ‘occur’ | 1772 | 11508.56 |
| tegema ‘do, make’ | 4364 | 17036.76 | tunduma ‘seem’ | 1981 | 11473.97 |

# Step 2: Analysis of the dictionary definitions

1. The lemmas (N=552) were provided with the dictionary definitions (The EKI Combined Dictionary 22)

2. Search for certain keywords in the text of definition fields. The initial set originates in a pilot study; here; verification and finding the new ones:

- Conflict (fight)
- Quarrel
- Collective involvement
- Intensity
- […]

→ Up to 50 descriptors

# Illustration of noun lemmas in the central part of the category, their definitions and defining vocabulary

| lemma | Translation of the definition |
|---|---|
| jama ‘trouble’ | 2. a situation that causes disturbance, strife or problem |
| tegemine ‘doing, making’ | 1.1 about effort, strain, the hassle of working on something |
| madin ‘struggle’ | 1. a brawl, fight, scuffle or other fierce (noisy) action; (about a battle, war); 1.1 a fierce dispute, war of words or quarrel; 2. the sound of footsteps, the noise of scuffling |
| lööm ‘fight’ | hand-to-hand fighting, sharing blows while bickering |
| kaklus ‘fight’ | 1. hand-to-hand fighting, sharing blows while bickering |
| ostmine ‘buing’ | buying something, acquiring something for money |
| sõit ‘ride’ | 1. moving with a vehicle, driving; trip, journey |
| ost ‘buing’ | 1.buying something, acquiring something for money |

# Examples of 2-step abstraction

Step 1. The stem-based cluster-like descriptors (N=50), e.g

**INTENSE** [fierce/intense/active]

**ARM**[arm/shooting/bomb]

Step 2. The higher-order descriptors (N = 17) based on semantic relatedness, e.g:

**VIOLENT**[VIOL[violence] / KILL[kill/destroy/murder] / ARM[arm/shooting/bomb] /WAR[war/battle]]

**EVENT**[ACTION[act/action/doing/making/proceeding] / PROCESS[process] / STATE[situation/state]]

# The relative prominence of the descriptors

232 definitions; 445 occurrences of the descriptors; average 2 per definition; StDev=1.3

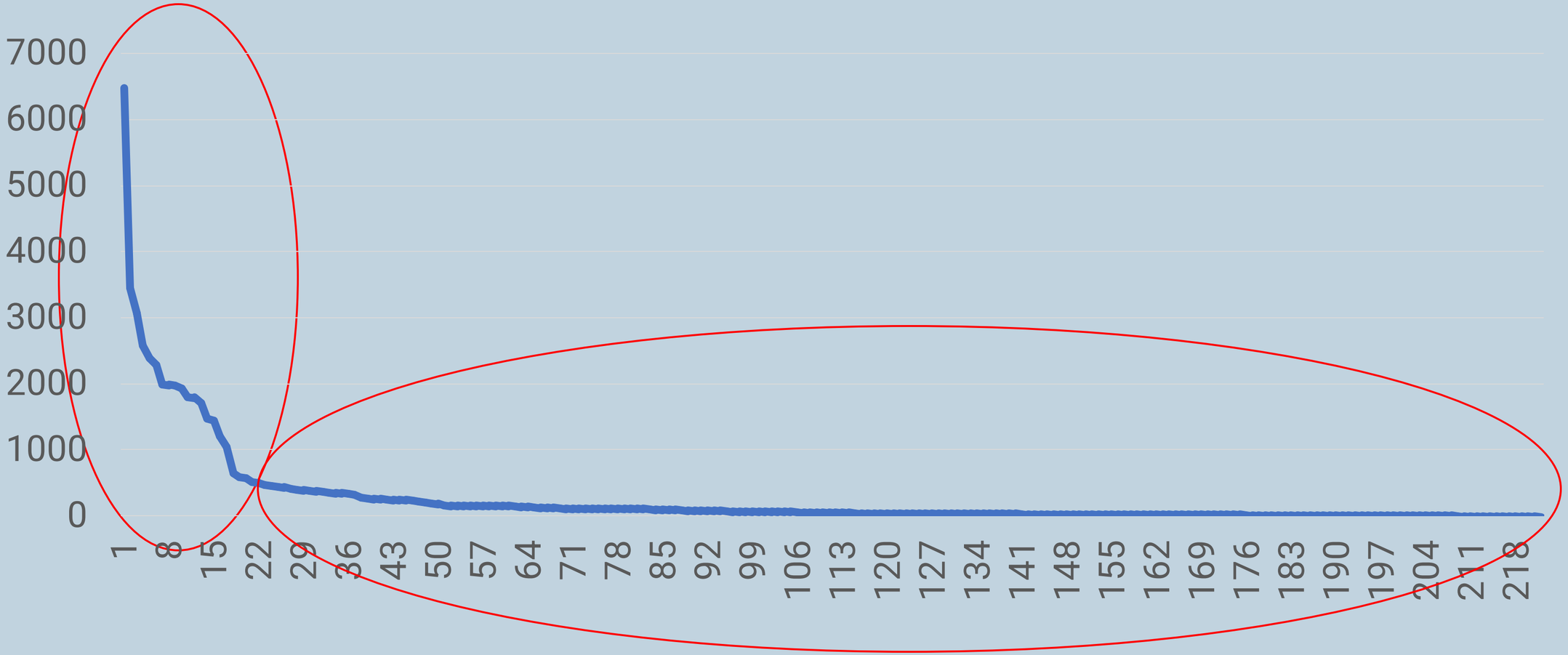
**Two groups in comparison**

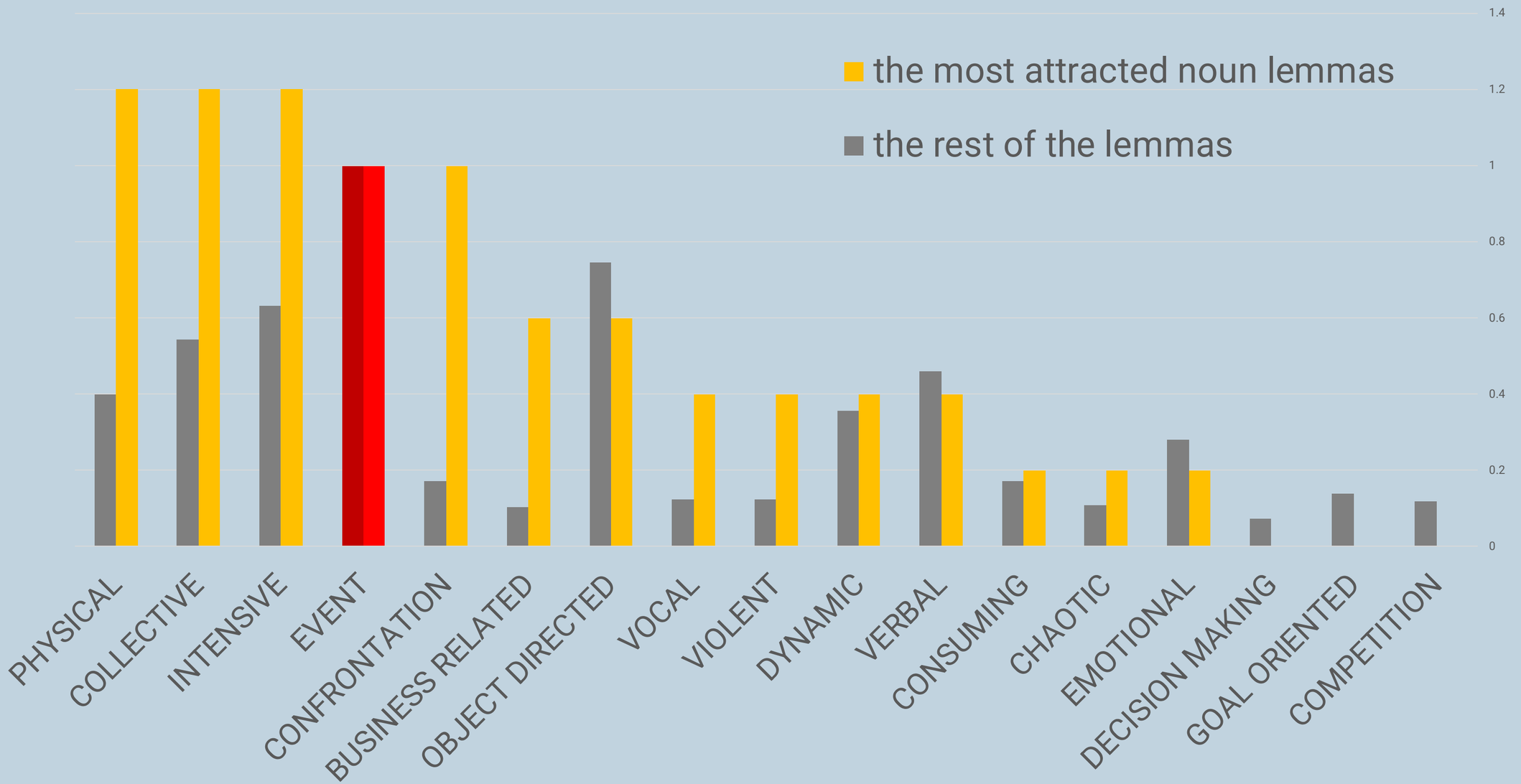
1) The noun lemmas most strongly attracted to the Translative Inceptive Construction (N=17)

2) The rest of noun lemmas occurring in the construction and provided with definitions in the CombiDic (N = 215)

In both groups the prominence of semantic descriptors was compared to the number of the descriptor EVENT

# Distribution of the data by the association strength (LLR) between the construction and the nouns (N=222)




the most attracted noun lemmas
the rest of the lemmas
1.4
1.2
1
0.8
0.6
0.4
0.2
0
PHYSICAL
COLLECTIVE
INTENSIVE
EVENT
CONFRONTATION
BUSINESS RELATED
OBJECT DIRECTED
VOCAL
VIOLENT
DYNAMIC
VERBAL
CONSUMING
CHAOTIC
EMOTIONAL
DECISION MAKING
GOAL ORIENTED
COMPETITION

# Conclusion

The collostructional analysis provided us with list of lemmas occurring in the open slot of a construction.

- The central members of the category as the most strongly associated
- The long “tail” of lemmas revealing the productivity of the construction

The semantic profile of the words occurring in an open slot of a construction can, indeed, be gleaned from their dictionary definitions

# Discussion

- The results suggest that the Translative Inceptive Construction primarily denotes the inception of intense collective and confrontational physical activities.
- This semantic profile is in accordance with the expressive and colloquial nature of the construction.
- The semantic profile of the less attracted collexemes suggests the productivity of the construction and reveals the features inherited from the central group: OBJECT-DIRECTED, INTENSIVE, COLLECTIVE, VERBAL.

# Discussion

- The DD-based approach gave rise to heterogenic descriptors. Some of the descriptors are similar to categories like event types or verb classes (e.g., CONSUMING) or frames (e.g., BUSINESS-RELATED), while others are orthogonal to these (e.g., INTENSIVE). This supports the added value of the method with respect to the use of existing ontologies and classifications.
- As a drawback, the method gave rise to a large number of descriptors, introducing the need for a subjective process of abstracting more general descriptors.

# Some limitations

- Applicable only to instances where the definitions are available (15 % of the types used in our corpus)
- Lots of manual work in order to establish the descriptors; polysemy needs manual inspection; automatization of the method is needed
- Outdated? much more effective tools like ChatGPT are available

# Thank you!

# ChatGPT: the central members

1. **Conflict/Aggression**: Words like "jama" (trouble), "madin" (scuffle), "lööm" (hit), "kaklus" (fight), "rebimine" (tearing), and "sõda" (war)
2. **Action/Activity**: Words like "tegemine" (doing), "ostmine" (buying), "sõit" (drive/trip), "möll" (commotion), "vahetus" (exchange), "jagamine" (sharing), and "vahetamine" (changing)
3. **Transaction/Exchange**: "Ostmine" (buying), "ost" (purchase), "vahetus" (exchange), and "vahetamine" (changing)
4. **Disagreement/Argument**: Words like "läbu" (mess), "kisma" (quarrel), "vaidlus" (argument), and "kisma" (quarrel)
5. **Intensity/Chaos**: Words like "möll" (commotion), "läbu" (mess), and "sõda" (war)

# ChatGPT: The profile of the “tail” (not present or not defined in CombiDic)

**Action Verbs**: Words like "andmine" (giving), "tellimine" (ordering), "küsimine" (asking), "paugutamine" (exploding), "lahmimine" (blabbering), "kihutamine" (speeding), etc.,

**Social Interaction**: Verbs like "kiitmine" (praising), "vaidlemine" (arguing), "rabamine" (scrambling), "jutustamine" (narrating), "äriamine" (negotiating), etc., indicate various forms of social interaction.

**Emotional States and Responses**: Words like "nutmine" (crying), "naermine" (laughing), "ohkimine" (sighing), "sünnipäevajutt" (birthday talk),

**Physical Actions**: Verbs like "ronimine" (climbing), "tantsimine" (dancing), "mängimine" (playing), "pesemine" (washing), "kraaklemine" (squabbling),

**Communication**: Words like "lobisemine" (chatting), "kirumine" (cursing), "kommenteerimine" (commenting), "jämmimine" (jamming), "arutamine" (discussing)

**Cognitive Processes**: Verbs like "mõistatamine" (puzzling), "analüüsimine" (analyzing), "teoretiseerimine" (theorizing), "proovimine" (trying)

**Conflict and Competition**: Words like "kaklemine" (fighting), "vandumine" (cursing), "konkureerimine" (competing), "rüselemine" (tussling)

**Creation and Crafting**: Verbs like "meisterdamine" (crafting), "kokkamine" (cooking), "joonistamine" (drawing), "punumine" (weaving)

**Daily Activities**: Words like "koristamine" (cleaning), "ostlemine" (shopping), "söögitegemine" (cooking), "pesemine" (washing)

**Personal Improvement**: Words like "treenimine" (training), "arendamine" (developing)