Combining Abstractness and Language-specific Theoretical Indicators for Detecting Non-Literal Usage of Estonian Particle Verbs

Eleri Aedmaa[1], Maximilian Köper[2], Sabine Schulte im Walde[2]

[1] Institute of Estonian and General Linguistics, University of Tartu, Estonia
[2] Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany
eleri.aedmaa@ut.ee,{maximilian.koeper,schulte}@ims.uni-stuttgart.de

Student Research Workshop Paper — 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics • Human Language Technologies •

GOALS

- Create two **datasets** for a low-resource Estonian.
- Build a random-forest **classifier** to automatically predict literal vs. non-literal language usage of particle verbs.
- Ascertain **the importance of language-specific features** when combined with language-independent features of abstractness.

DATASETS

- Dataset of literal and non-literal language usage for Estonian PVs: 210 PVs across 34 particles, 1490 sentences.
- Automatically created abstractness ratings for 243,675 Estonian lemmas.
- Available at `http://github.com/eleriaedmaa/`

CLASSIFICATION RESULTS

	feature type	acc	F_1 n-lit	F_1 lit
	majority baseline	74.0%	85.0	0.00
1	particle (p)	73.6%	84.4	13.6
2	base verb (v)	81.2%	87.9	58.0
3	unigrams, f>5 (uni)	82.3%	89.0	54.6
4	average rating of words (abs)	68.1%	79.8	24.5
5	average rating of nouns (abs)	68.5%	79.7	30.1
6	rating of the PV subject (abs)	72.3%	83.1	23.7
7	rating of the PV object (abs)	73.0%	83.5	25.2
8	subject case (case)	74.0%	85.0	0.00
9	object case (case)	74.0%	85.0	0.00
10	subject animacy (animacy)	74.0%	85.0	0.00
11	object animacy (animacy)	74.0%	85.0	0.00
12	case government (govern)	73.8%	84.6	10.1
	p+v, 1–2	85.2%	90.3	68.7
	v+uni, 2–3	84.2%	89.6	66.4
	p+v+uni, 1–3	85.0%	90.1	68.5
	p+v+abs, 1–2, 4–6	86.3%	90.9	72.3
	p+v+abs, 1–2, 4–7	86.0%	90.7	71.3
	p+v+abs, 1–2, 5–6	86.0%	90.7	71.9
	p+v+case, 1–2, 8	85.3%	90.4	68.9
	p+v+case, 1–2, 8–9	84.6%	89.7	69.3
	p+v+animacy, 1–2, 10–11	86.2%	90.8	72.3
	p+v+govern, 1–2, 12	86.2%	90.9	71.6
	p+v+abs+lang, 1–2, 4–6, 10–12	87.3%	91.6	73.8
	p+v+abs+lang, 1–2, 4–12	87.5%	91.8	73.8
	p+v+abs+lang, 1–2, 5–6, 8, 10, 12	**87.9%**	**92.0**	**75.0**

CONCLUSION

Language-specific features **subject case**, **subject animacy** and **case government** combined with **abstractness ratings** as well as **verb** and **particle** information classify literal vs. non-literal usage of PVs with accuracy 87.9%.

ACKNOWLEDGEMENTS

This research was supported by the University of Tartu ASTRA Project PER ASPERA, financed by the European Regional Development Fund (Eleri Aedmaa), and by the Collaborative Research Center SFB 732, financed by the German Research Foundation DFG (Maximilian Köper, Sabine Schulte im Walde).

ESTONIAN PARTICLE VERB

Estonian particle verb = an adverbial particle + base verb (e.g., *alla andma* 'to give up')

Challenging because:

- their components do not always appear adjacent to each other
- the particles are homonymous with adpositions
- the same PV can be used in literal (1) vs. non-literal (2) language

(1) Ta **astu-s** kaks sammu **tagasi**.
he step-PST.3SG two step.PRT back
'He took two steps back.'

(2) Ta **astu-s** ameti-st **tagasi**.
he step-PST.3SG job-ELA back
'He resigned from his job.'

FEATURES

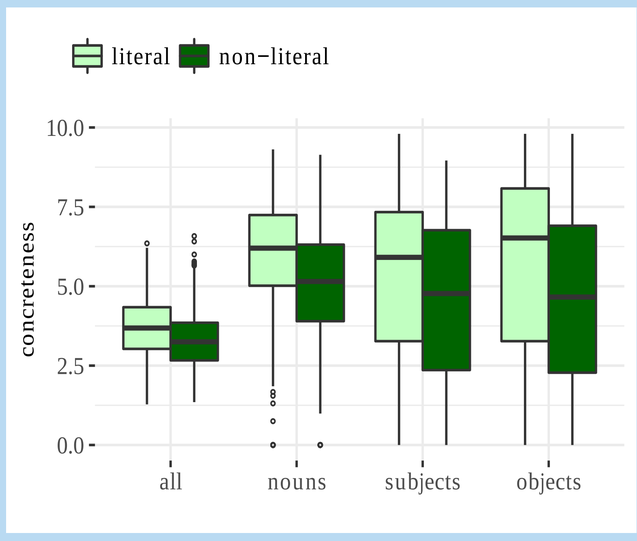
1 **particle** – particle of the particle verb

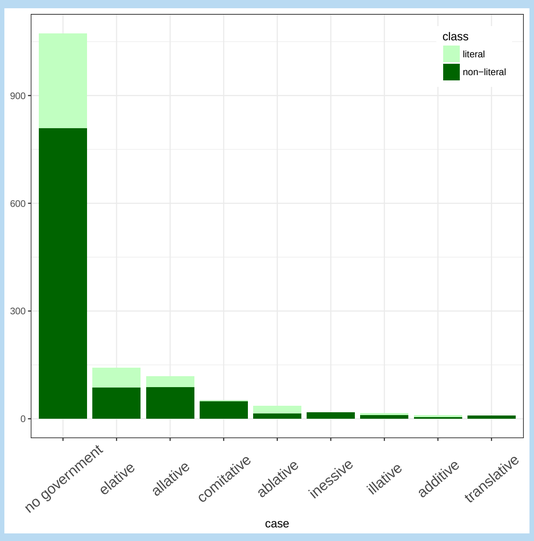
2 **verb** – verb of the particle verb

3 **unigrams** – lemmas of content words that occur in the same sentences with target PVs

4–7 **abstractness features** – average rating of all words in a sentence, average rating of all nouns in a sentence, rating of the PV subject, rating of the PV object

12 **case government** – case of the argument of the particle verb (excluding subject and object); value = one of the 14 cases





8–9 **case features** – case of the PV subject (nominative or partitive) and PV object (nominative, genitive or partitive)

10–11 **animacy features** – whether the subject and object are alive or not