19TH ANNUAL CONFERENCE OF APPLIED LINGUISTICS, JUNE 16 - 17, 2022, TALLINN

The Making and Breaking of Classification Models in Linguistics: A Multimethod Perspective on Alternations

Jane Klavan (University of Tartu, Estonia)
jane.klavan@ut.ee

UNIVERSITY OF TARTU
1632

Usage-based linguistics and Cognitive Linguistics

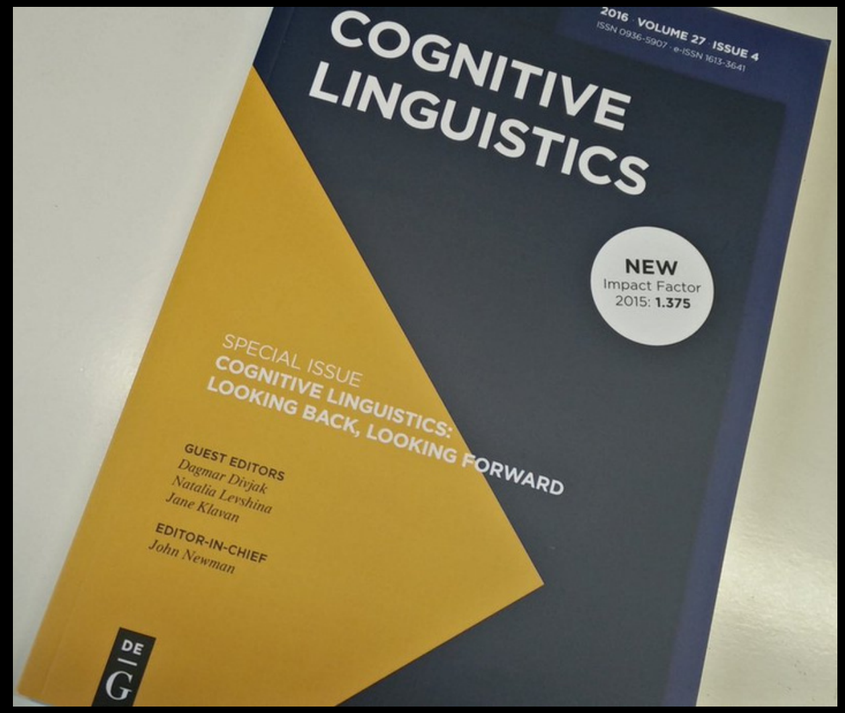
UNIVERSITY OF TARTU
1632

Today's talk

- Introduction
  - Cognitive Linguistics and the Quantitative Turn
  - Alternations - what, why, & how?
- Combining methods:
  - corpus-based study of alternations
  - linguistic experiments with alternations
  - corpora vs. experiments
- Interim conclusions
- Discussion: work in progress

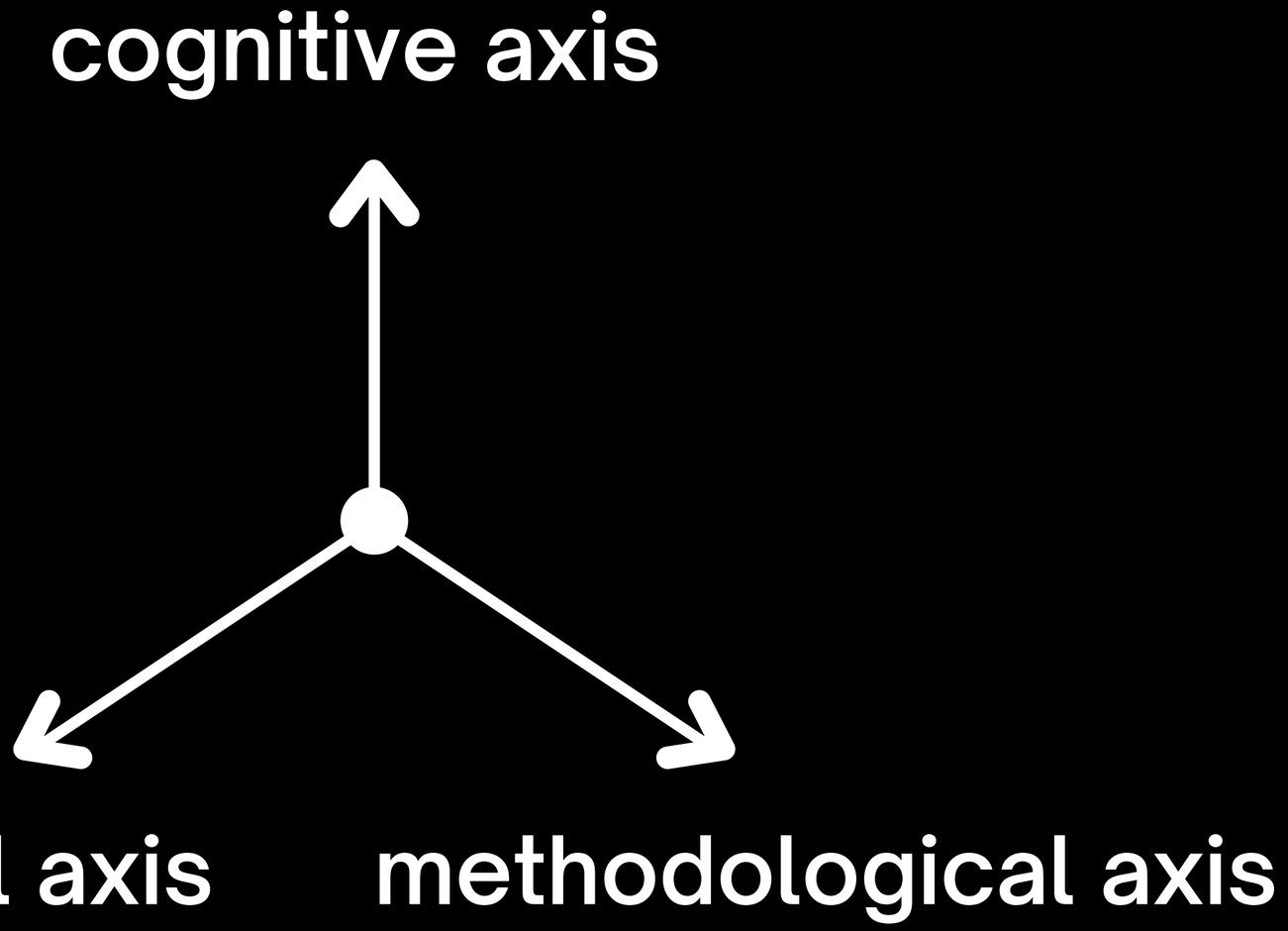
Divjak, Dagmar, Natalia Levshina, and Jane Klavan. 2016. Cognitive Linguistics: Looking back, looking forward. *Cognitive Linguistics* 27 (4): 447-463.

"The historical "prototype" of Cognitive Linguistics may be described as predominantly of mentalist persuasion, based on introspection, specialized in analysing language from a synchronic point of view, focused on West-European data (English in particular), and showing limited interest in the social and multimodal aspects of communication."



UNIVERSITY OF TARTU
1632

Divjak, Dagmar, Natalia Levshina, and Jane Klavan. 2016. Cognitive Linguistics: Looking back, looking forward. *Cognitive Linguistics* 27 (4): 447-463.



Cognitive Linguistics and the Quantitative Turn

- Introspection is deeply embedded in Cognitive Linguistics for both historical as well as theoretical reasons
- The mid-1990s saw a shift in paradigm
  - For the journal *Cognitive Linguistics* the year 2008 "marks the quantitative turn" (Janda 2013: 2)
- It is the discipline's theoretical assumptions, namely its cognitive nature, its usage-based perspective, and its contextualizing approach (Geeraerts 2006: 31) that make Cognitive Linguistics a particularly good candidate for championing the methodological progress of linguistics.

Cognitive Linguistics and the Quantitative Turn

- Exponential growth in studies that use statistical analysis of corpus data or experimental findings
- Publication of edited volumes and monographs on linguistic methodology (e.g. Gonzalez-Marquez et al. 2007, Glynn and Fischer 2010, Newman and Rice 2010, Janda 2013, Glynn and Robinson 2014)
- Textbooks introducing linguists to statistics (e.g. Baayen 2008, Johnson 2008, Gries 2009, Levshina 2015, Winter 2020)

the existential question of a cognitive linguist:

"to be empirical or to be introspective" (Zlatev 2016)?

both approaches are crucial for the development of cognitive linguistics

"qualitative descriptions provide the basis for quantitative methods such as experiment, neural imaging, and computer modeling - they suggest what to look for and allow the interpretation of results" (Langacker 2016)

Constructional alternations - what, why, & how?

UNIVERSITY OF TARTU
1632

Why study constructional alternations?

Linguistic variation in all its guises

What are constructional alternations?

What are constructional alternations?

(1) a. John sent Mary the book.
b. John sent the book to Mary.

(2) a. Picasso painted this picture.
b. This picture was painted by Picasso.

(3) a. John picked up the book.
b. John picked the book up.

(4) a. the university's budget
b. the budget of the university

(5) a. John will send Mary a book.
b. John is going to send Mary a book.

Labovian sociolinguistics

‘alternate ways of saying the same thing’
(Labov 1972: 188)

Redefining alternations

- practical research setup created by the researcher to test more general hypotheses (Arppe et al. 2010: 13–15);
- two or more forms that compete for the same function in a community of language users (Van de Velde 2014, 2017);
- a choice point of the individual language user (Bresnan et al. 2007);
- various constructions that have a special relation to one another in the constructicon, e.g. as allostructions (Cappelle 2006; Perek 2012).

Source:
https://www.uantwerpen.be/en/conferences/construction-grammars/scientific-program/workshops/alternations/

My approach to constructional alternations

"... an expression imposes a particular construal, reflecting just one of the countless ways of conceiving and portraying the situation in question."

"The term construal refers to our manifest ability to conceive and portray the same situation in alternate ways."

Langacker, R. 2008. Cognitive Grammar: A Basic Introduction. Oxford: OUP.

UNIVERSITY OF TARTU
1632

Exterior locative constructions in Estonian

(1) LATIVE

Paneb	raamatu	{lauale	/	laua	peale. }
Put-PRS.3SG	book.SG.GEN	table.SG.ALL		table.SG.GEN	onto

"He/She puts the book on(to) the table."

(2) LOCATIVE

Raamat	on	{laual	/	laua	peal. }
book.sg.nom	be-prs.3sg	table.sg.ade		table.sg.gen	on

"The book is on the table."

(3) SEPARATIVE

Võtab	raamatu	{laualt	/	laua pealt. }	
take-PRS.3SG	book.SG.GEN	table.SG.ABL		table.SG.GEN	from on

"He/She takes the book from the table."

UNIVERSITY OF TARTU
1632

How to study constructional alternations?

My bent for methodology ...

Combining different methods for the study of alternations - corpora and experiments

Taking a leap of faith

behavioural data proxy for cognition

corpus data proxy for language production

experiments proxy for language comprehension

Aims and predictions

to measure the extent and nature of variation as reflected in language production and comprehension

It is expected that the alternations exhibit different constraints on their use as seen in language production and comprehension

01 Factors across varieties

the influence of certain factors across different varieties of the language should be relatively stable in terms of the direction of those factors

02 Factors across constructions

the strength of different factors on speakers' choices will vary by the types and frequencies of constructions

03 Factors across speakers

the variation in the use of alternations may be driven by stylistic preferences, situational forces or by cognitive pressures related to language processing

UNIVERSITY OF TARTU
1632

Corpus-based study of alternations

Estonian National Corpus (1.1 billion words, mainly web-based)

3,000 usages of exterior locative constructions

Corpus data: Exterior Locative Constructions in Estonian

Construction	*F*	*f*	*s*	*Pr* = *s* / *f*
Lative				
Allative	19,187,296	3,017	500	0.166
peale	959,515	2,142	500	0.233
Locative				
Adessive	30,661,120	5,148	500	0.097
peal	241,263	1,210	500	0.413
Separative				
Ablative	2,675,044	1,745	500	0.287
pealt	138,049	872	500	0.573

Polysemy of constructions: allative

ALLATIVE	EXAMPLE SENTENCE	POSTPOSITIONAL ALTERNATIVE	ENGLISH TRANSLATION
Direction of location	Mari pani vaasi **lauale**	Mari pani vaasi **laua peale.**	'Mari put the vase on(to) the table.'
Time	Koosolek viidi üle **neljapäevale.**	Koosolek viidi üle **neljapäeva peale**.	The meeting has been moved to Thursday.'
State	Tüdruku nägu läks **naerule**.	not attested	'The girl started to laugh.'
Addressee	Mari rääkis **Jürile** kõik ära.	not attested	'Mari told Jüri everything.'
Experiencer	**Mulle** meeldib siin elada.	not attested	'I like living here.'
Object of action	Ta lootis **sõpradele**.	Ta lootis sõprade peale.	'He counted on friends.'
Object of emotions	Mihkel on **sõbrale** kade.	Mihkel on **sõbra peale** kade.	'Mihkel is jealous of his friend.'
Without clear meaning	Järgenege mulle.	not attested	'Follow me.'

UNIVERSITY OF TARTU
1632

Polysemy of constructions: adessive

ADESSIVE	EXAMPLE SENTENCE	POSTPOSITIONAL ALTERNATIVE	ENGLISH TRANSLATION
Location	Vaas on **laual**.	Vaas on **laua peal**.	'The vase is on the table.'
Time	Nad sõidavad **neljapäeval** maale.	not attested	'They are driving to the country on Thursday.'
State	Jüri vaatas meid **naerul** näoga.	not attested	'Jüri looked at us with a laughing face.'
Possessor	**Maril** on kaks last.	not attested	'Mari has two children.' (lit. 'On Mary are two children.')
Agent with finite verb forms	See asi ununes **mul** kiiresti.	not attested	'I quickly forgot about that thing.'
Instrument	Mari mängib **klaveril** mõnd lugu.	Mari mängib **klaveri peal** mõnd lugu.	'Mari is playing some tunes on the piano.'
Manner	Mari kuulas kikkis **kõrvul**.	not attested	'Mari listened with her ears pricked up.'

Polysemy of constructions: ablative

ABLATIVE	EXAMPLE SENTENCE	POSTPOSITIONAL ALTERNATIVE	ENGLISH TRANSLATION
Source of location	Mari võttis vaasi **laualt**.	Mari võttis vaasi **laual pealt**.	'Mari took the vase off the table.'
Source	Mari kuulis seda **Jürilt**.	not attested	'Mari heard it from Jüri.'
Modifier of a noun	**Elukutselt** on ta insener.	not attested	'He is an engineer by profession.'

Corpus data: Exterior Locative Constructions in Estonian

Construction	*F*	*f*	*s*	*Pr* = *s* / *f*
Lative				
Allative	19,187,296	3,017	500	0.166
peale	959,515	2,142	500	0.233
Locative				
Adessive	30,661,120	5,148	500	0.097
peal	241,263	1,210	500	0.413
Separative				
Ablative	2,675,044	1,745	500	0.287
pealt	138,049	872	500	0.573

Annotation of the corpus data

Table: Definition of variables

Variable	Category	Scale/levels (reference level stated first for categorical variables)
POSTPOS	outcome	*CASE* *POSTPOSITION*
POSITION	fixed	*post* *pre*
CONCRETENESS	fixed	*CONC_01* *CONC_02* *CONC_03*
MOBILITY	fixed	*MOBILE* *STATIC*
COMPLEXITY	fixed	*SIMPLE* *COMPOUND*
LENGTH	fixed	$\log_2$-length (in syllables) of landmark phrase
RATIO	fixed	$\log_2$-frequency (raw) of landmark lemma used with the case affix relative to the frequency of the lemma used with the postposition
FUNCTION	fixed	*adverbial* *modifier*
LM_LEMMA	random	592 levels (lative) 438 levels (locative) 528 levels (separative)

UNIVERSITY OF TARTU
1632

Annotation: example

Malka	istus	**suvekohviku**	**valgel**
Malka.SG.NOM	sit-PST.3SG	summer café.SG.GEN	white.SG.ADE

korvtoolil	ja	luges	midagi.
wicker chair.SG.ADE	and	read-PST.3SG	something.SG.PRT

‘Malka was sitting on the white wicker chair of the summer café and was reading something.’

Aims and predictions

to measure the extent and nature of variation as reflected in language production and comprehension

It is expected that the alternations exhibit different constraints on their use as seen in language production and comprehension

Statistical modelling

mixed-effects logistic regression

Harrell 2001, Pinheiro and Bates 2002, Hosmer et al. 2013, Winter 2019

software R

version 3.6.1, R development core team 2019

lme4 package

Bates 2014, Bates et al. 2015

Model formula fitted to the data:
construction ~ log(ratio) + log(length) + concreteness + mobility + complexity + synfun + position + (1|lemma)

Alternation	Model evaluation	
	Model accuracy	C-value
Lative: allative ~ peale	78%	0.87
Locative: adessive ~ peal	86%	0.93
Separative: ablative ~ pealt	79%	0.88

Corpus-based results

Prediction 1:
the influence of certain factors across different varieties of the language should be relatively stable in terms of the direction of those factors

The grammatical knowledge of Estonian exterior locative cases and the corresponding postpositions is probabilistic and regulated by the different factors (the length, complexity, mobility, position, and function of Landmark phrase) in a relatively unifrom way:

Landmark phrases that are simple, shorter, mobile and function as adverbials (rather than modifiers) favour the use of postpositions

Corpus-based results

Prediction 1:
the influence of certain factors across different varieties of the language should be relatively stable in terms of the direction of those factors

written language

Klavan, Jane. 2012. *Evidence in linguistics: corpus-linguistic and experimental methods for studying grammatical synonymy.* (Dissertationes Linguisticae Universitatis Tartuensis). Tartu: University of Tartu Press.

Klavan, Jane. 2020. Pitting corpus-based classification models against each other: a case study for predicting constructional choice in written Estonian. *Corpus Linguistics and Linguistic Theory*, 16 (2), 363–391.

spoken language

Klavan, Jane, Maarja-Liisa Pilvik & Kristel Uiboaed. 2015. The Use of Multivariate Statistical Classification Models for Predicting Constructional Choice in Spoken, Non-Standard Varieties of Estonian. *SKY Journal of Linguistics*, 28, 187–224.

web texts

Klavan, Jane. 2021. The alternation between exterior locative cases and postpositions in Estonian web texts. *ESUKA-JEFUL*, 12 (1), 153–188.

Corpus-based results

Prediction 2:
the strength of different factors on speakers' choices will vary by the types and frequencies of constructions

Ranking of predictors for the three alternations:

allative ~ peale (LATIVE):
RATIO > LEMMA > LENGTH > SYNFUN > MOBILITY

adessive ~ peal (LOCATIVE):
RATIO > LEMMA > LENGTH > MOBILITY > SYNFUN

ablative ~ pealt (SEPARATIVE):
RATIO > LEMMA > COMPLEXITY > MOBILITY > CONC

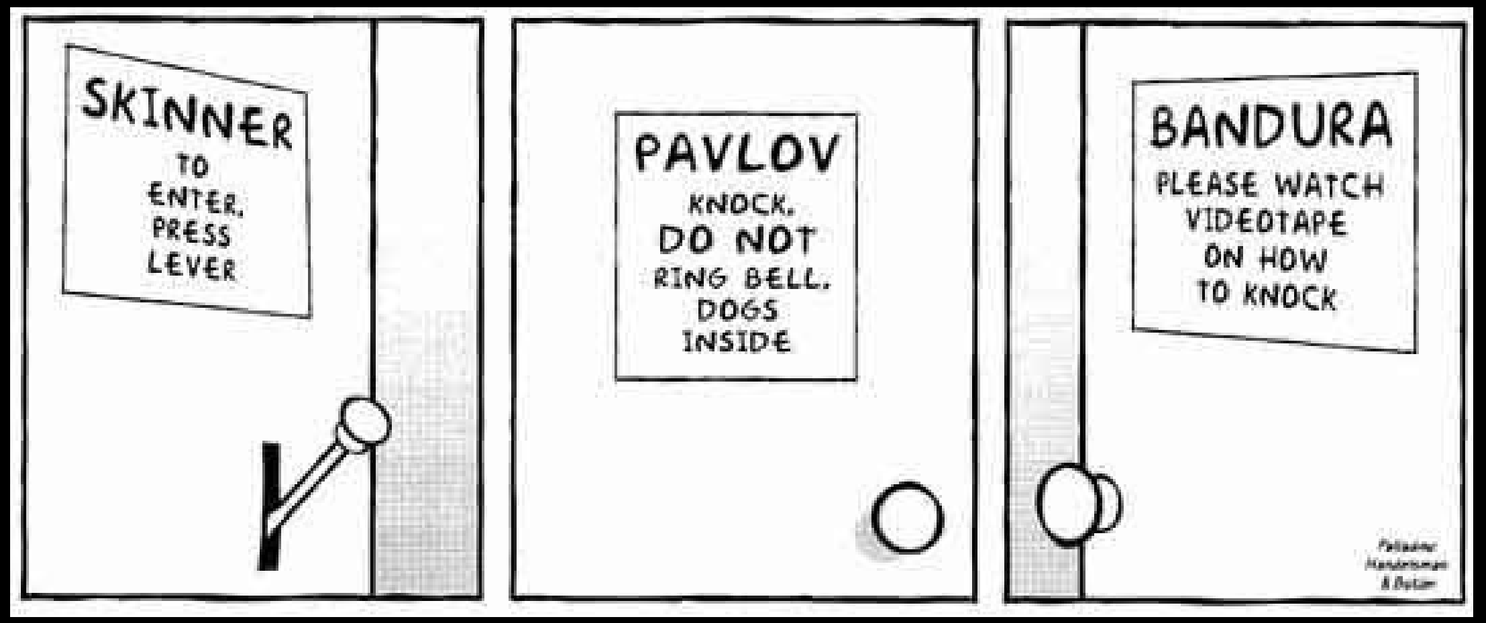
Corpus-based study of alternations: interim summary

Corpora allow me to detect patterns in the data and determine what is typical in the language

Corpus-based study of alternations: interim summary

Corpora don't tell me what is possible in the language and they don't allow me to test specific hypotheses

Enter linguistic experiments



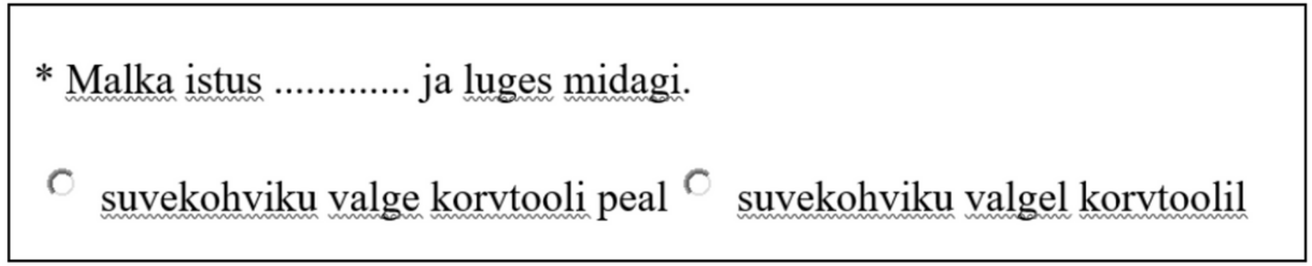
Acceptability rating task

B. Sample item for the rating task (adessive construction)

C. Sample item for the rating task (*peal* construction)

Forced choice task

A. Sample item for the forced choice task



	Forced choice task	Acceptability rating task
Number of participants	75 (60 female, 14 male, 1 preferred not to say)	105 (85 female, 18 male, 2 preferred not to say)
Age of participants	Mean 37, SD = 14.9 (range 19 – 76 years)	Mean 34, SD = 12.6 (range 18 – 66 years)

UNIVERSITY OF TARTU
1632

Results of the forced choice task

Table. Number and proportion of choices for case construction vs postposition construction across the three alternations

Type of alternation	Case constructions	Postpositional constructions	Total
Lative: allative ~ peale	310 (69%)	140 (31%)	450 (100%)
Locative: adessive ~ peal	284 (63%)	166 (37%)	450 (100%)
Separative: ablative ~ pealt	294 (65%)	156 (35%)	450 (100%)
Total	888 (66%)	462 (34%)	1350 (100%)

Results of the acceptability rating study

Table. Residualised mean ratings for case construction vs postposition construction across the three alternations

Type of alternation	Case constructions	Postpositional constructions	Overall
Lative: allative ~ peale	6.7	6.4	6.6
Locative: adessive ~ peal	6.8	6.3	6.5
Separative: ablative ~ pealt	6.8	6.6	6.7
Overall	6.8	6.4	6.6

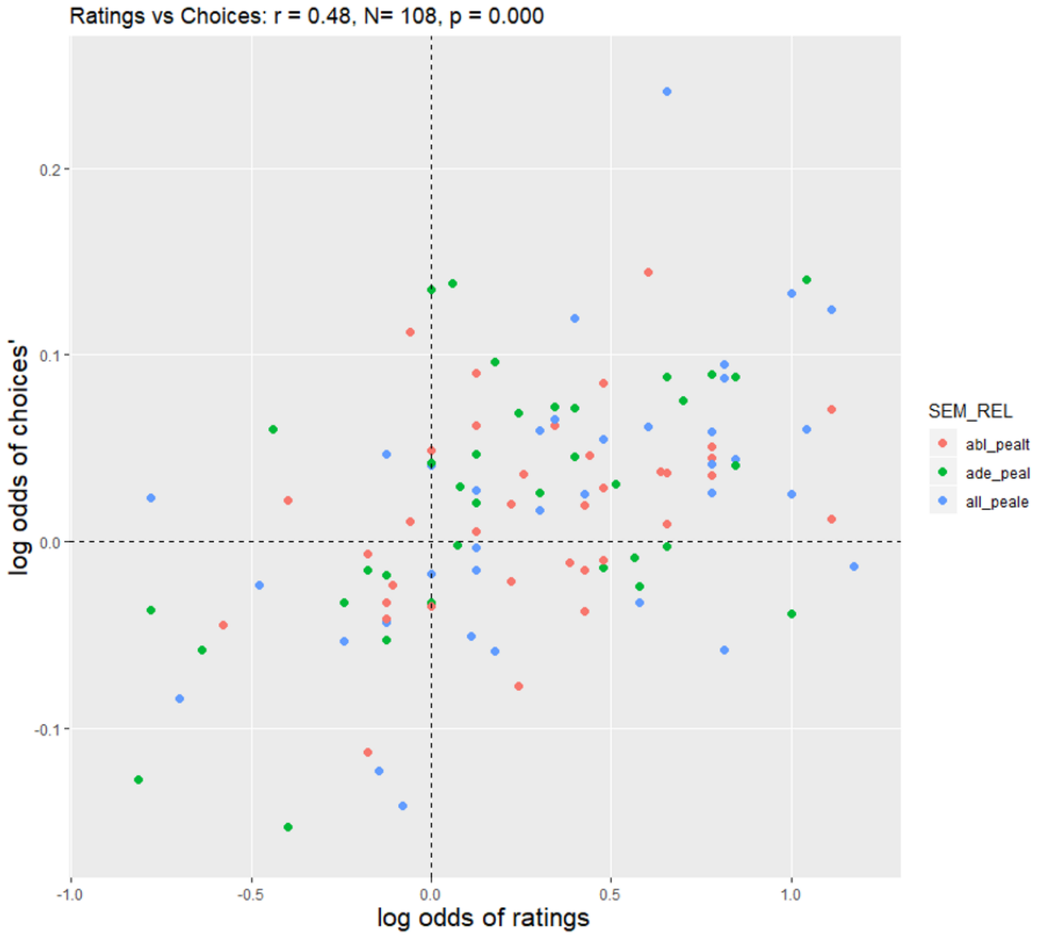


Figure X.1 The log odds (of case vs postposition) for each of the 108 experimental items and the pairwise Pearson correlation between residualised ratings and choices. The cut-off point for both the horizontal and vertical dimension is zero: a dot that falls to the right of or above zero indicates the predominance of the adessive construction, whereas a dot to the left of or below zero indicates the predominance of the peal construction. Positive scores indicate a preference for the case construction, negative scores a preference for the postpositional construction.

Experiment results

When we produce language, we prefer one construction

When we comprehend language, we judge both constructions as ok

There is a strong correlation between choices and ratings.

There are also some clear instances where the two diverge:

clear preference in the forced choice data, but no difference in the acceptability ratings

corpora vs. experiments

Klavan, Jane. 2020. Pitting corpus-based classification models against each other: a case study for predicting constructional choice in written Estonian. *Corpus Linguistics and Linguistic Theory*, 16 (2), 363–391.

Klavan, Jane & Ann Veismann. 2017. Are corpus-based predictions mirrored in the preferential choices and ratings of native speakers? Predicting the alternation between the Estonian adessive case and the adposition peal 'on'. *ESUKA – JEFUL*, 8 (2), 59–91.

Klavan, Jane & Dagmar Divjak, Dagmar. 2016. The Cognitive Plausibility of Statistical Classification Models: Comparing Textual and Behavioral Evidence. *Folia Linguistica*, 50 (2), 355–384.

Discussion

- How does the polysemy of constructions factor into the (grammatical) knowledge / representation of morphosyntactic alternations?
- Is there a (qualitative) change in the knowledge representations of different alternations speakers draw on in language production and language comprehension?
- Do speakers' choices and ratings in a forced choice task and acceptability rating task vary according to the types and frequencies of constructions?

Conclusion:

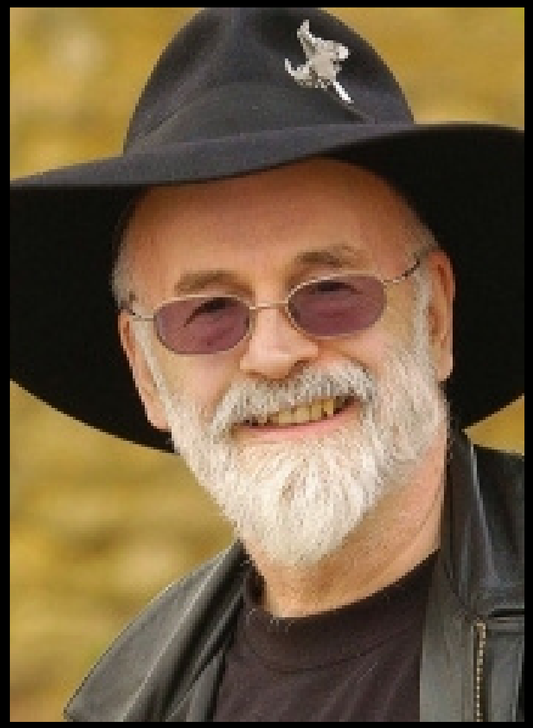
Alternation between exterior locative constructions in Estonian

- the grammatical knowledge of exterior locative alternations in Estonian is probabilistic and regulated by various factors
- the influence of certain factors across different varieties of the language is relatively stable in terms of the direction of those factors
- the Estonian data shows that morphosyntax and semantics both play a role, differently from the syntactic alternations in English, where the main constraining factors have been discourse-related factors (e.g. animacy, givenness, weight)
- the relative importance of factors differs across the different constructions: the separative relation (*ablative ~ pealt*) responds most strongly to the different factors

Conclusion:

The Making and Breaking of Classification Models in Linguistics

- corpus-based studies are necessary because they provide ecologically valid data
- using advanced statistical modelling for a richly annotated corpus sample allows us to capture the speakers' multivariate and probabilistic knowledge quantitatively
- without experimental data it would be very difficult if not impossible to provide an adequate assessment of corpus-based models - linguistic experiments are necessary to calibrate our corpus-based models
- different types of (experimental) data give us access to different types of behaviour which we use as proxy for cognition



“There is always a choice.”

Terry Pratchett. 2004. *Going Postal*.

thank you!