# The role of frequency in morpho-syntactic alternations: An experimental study from Estonian

Jane Klavan[1] & Alice Blumenthal-Dramé[2]
[1] University of Tartu
[2] Albert-Ludwigs-Universität Freiburg

# Constructional alternations

- Constructional alternations = alternative linguistic means used to designate the “same” concept or linguistic function
- The  language user can choose among a variety of grammatical and lexical items to construe an experience or a situation
- Even if two linguistic units do express broadly the same function, they do it in different ways: they allow for a different construal of the same situation (*the no-synonymy hypothesis*)

# Modelling native speakers' preferences

- An issue that has received substantial amount of attention: Bresnan (2007), Bresnan et al. (2007), Bresnan & Ford (2010), Arppe & Abdulrahim (2013), Divjak et al. (2016); see Klavan & Divjak (2016) for an overview
- Multivariate analyses = corpus-based + experimental research
- A number of variables that significantly affect subjects' preferences across a range of different paradigms and languages (syntactic, semantic, discourse)
- Overall: subjects' preferred choices reliably pick out the same choices made in the original corpus sample => a high and significant correlation between the proportions of selected constructions and the matching corpus-based probability estimates

## OUR CASE STUDY

Morpho-syntactic alternation from Estonian: adessive case (ex 1) vs *peal* ‘on’ (ex 2)

(1) *Raamat* *on* ***laual.***
book.SG.NOM be-PRS.3SG table.SG.ADE
‘The book is **on the table**.’

(2) *Raamat* *on* ***laua*** ***peal.***
book.SG.NOM be-PRS.3SG table.SG.GEN on
‘The book is **on the table**.’

Klavan, Jane, Maarja-Liisa Pilvik, and 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.

**Table 2**. Coefficients for a mixed-effects logistic regression model for Estonian dialect dataset

| | Estimate | Std. Error | z-value | *p*-value |
|---|---|---|---|---|
| Intercept | -1.894 | 0.643 | -2.947 | 0.0032 |
| LENGTHLOG | 1.390 | 0.467 | 2.976 | 0.0029 |
| COMPLEXITY = simple | 1.765 | 0.442 | 3.995 | 0.0001 |
| TYPE = thing | 1.379 | 0.316 | 4.372 | 0.0000 |
| VERBGROUP = existence | -0.531 | 0.177 | -2.996 | 0.0027 |
| VERBGROUP = motion | -1.287 | 0.234 | -5.496 | 0.0000 |
| VERBGROUP = no verb | -0.142 | 0.276 | -0.515 | 0.6069 |
| VERBGROUP = posture | -0.180 | 0.467 | -0.386 | 0.6998 |
| DIALECT = Eastern | 1.266 | 0.550 | 2.303 | 0.0213 |
| DIALECT = Coastal | 0.270 | 0.548 | 0.492 | 0.6224 |
| DIALECT = Insular | 1.137 | 0.460 | 2.472 | 0.0134 |
| DIALECT = Mid | 1.663 | 0.474 | 3.510 | 0.0004 |
| DIALECT = Mulgi | 1.414 | 0.590 | 2.396 | 0.0166 |
| DIALECT = Seto | 2.567 | 0.754 | 3.404 | 0.0007 |
| DIALECT = Tartu | 1.919 | 0.570 | 3.367 | 0.0008 |
| DIALECT = Võru | 1.665 | 0.531 | 3.137 | 0.0017 |
| DIALECT = Western | 2.265 | 0.477 | 4.751 | 0.0000 |

# Adessive vs *peal*

Predictors that play a role in non-standard, spoken Estonian:

- semantic predictors (e.g. type and mobility of the Landmark, type of verb used in the construction)
- morphosyntactic predictors (e.g. length, complexity)
- dialect
- individual speakers

# Adessive vs *peal* (Klavan, J. aop. Pitting corpus-based classification models against each other: A case study for predicting constructional choice in written Estonian. *Corpus Linguistics and Linguistic Theory*)

Table 3. Model comparison statistics

| | Resid. Dev | Df | Deviance | *p*-value | Reduction in AIC |
|---|---|---|---|---|---|
| Intercept | 1247.7 | | | | |
| LENGTH | 1155.7 | 1 | 91.968 | 0.0000 | 90.0 |
| COMPLEXITY | 1115.7 | 1 | 39.965 | 0.0000 | 38.0 |
| MOBILITY | 1079.9 | 1 | 35.807 | 0.0000 | 33.8 |
| VERBGROUP | 1067.2 | 4 | 12.681 | 0.0000 | 4.7 |
| TRWC | 1051.3 | 2 | 15.967 | 0.0000 | 12.0 |
| POSITION | 1047.2 | 1 | 4.088 | 0.0000 | 2.1 |

# Our research question: what about frequency?

- the extent to which speakers' preferences correlate with **usage frequencies** as attested in corpora
- **which** out of a number of competing **frequency metrics** that have wide currency in psycholinguistics and corpus-based cognitive linguistics **is best suited to predict native speaker behaviour**?
  - in a forced choice task
  - in an acceptability rating task

Ellis, Nick C. 2002. Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition,* 24, 143 - 188.

“Frequency plays a large part in explaining sociolinguistic variation and language change.”

“Learners’ sensitivity to frequency in all these domains has implications for theories of implicit and explicit learning and their interactions.”

(Ellis 2002: 143)

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

# Exp 1: Forced choices

- **Stimuli:** 30 sentences from annotated corpus sample (Klavan 2012)
- **Participants**: 96 native speakers of Estonian were recruited via the Internet using social media
  - randomly assigned to one of the four versions
  - 47 male participants
  - ranged in age from 18 to 54 (mean 29, SD = 9.5)

# Exp 2: Acceptability ratings

- **Stimuli:** same 30 items; 2 alternative sentences construed
- **Participants**: 98 native speakers of Estonian were recruited via the Internet using social media
- randomly assigned to one of the eight lists (~ 12 participants per list)
- 48 male participants
- ranged in age from 15 to 66 (mean 31, SD = 10.7)

# A. Sample item for the forced choice task

* Malka istus ..... ja luges midagi.

○ suvekohviku valge korvtooli peal ○ suvekohviku valgel korvtoolil

- an alternative paraphrase was constructed for each sentence
- both alternatives were presented together with the original context
- each subject completed the task with the same 30 sentences
- **Instructions:** “Which of the two constructions suits into the blank better?”

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


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


**Instructions:** "Rate the naturalness of the phrase between the square brackets on a 10-point scale ranging from very strange to completely natural"

It was decided not to show both alternatives to one and the same participant

- 60 experimental items were divided into two lists of 30 items each

# General predictions & assumptions

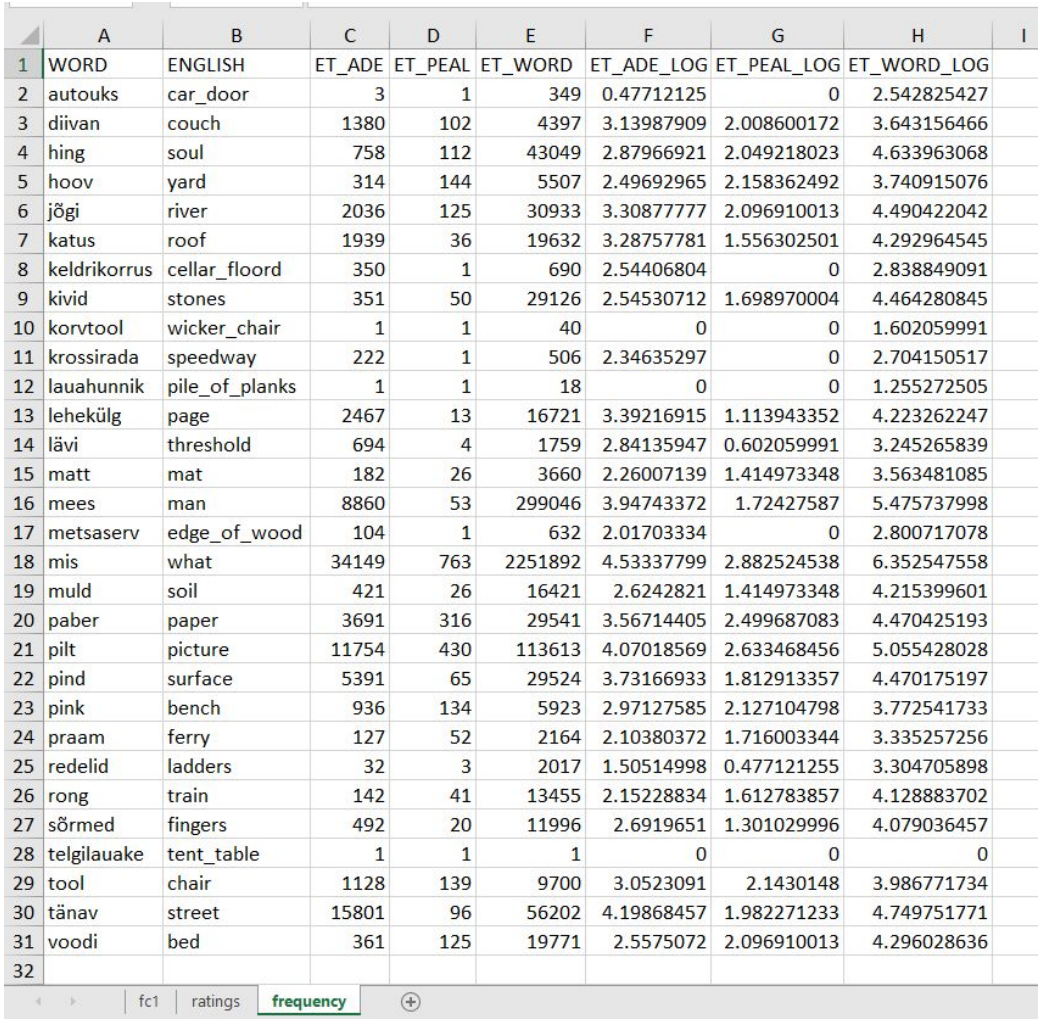
- **Language users' preferences are influenced by the relative frequencies with which certain nouns appear with different locative cases and postpositions**
  - *Frequency is predictive of the speakers' choices and ratings in the experimental studies of adessive vs peal 'on'*
- Assumption: such information is acquired through experience with input that exhibits distributional properties (Ellis 2002: 144)
- "The effects of frequency in input are modulated by the need to simultaneously satisfy the constraints of all other constructions that are represented in the learner's system." (Ellis 2002: 145)

# Which frequency measures?

## A wide variety of measures are available:

- Token frequencies (cf. exemplar-based theories)
- Analysis of contingency (e.g. collocation and collostructional metrics) => widely used in the corpus-linguistic community (Gries & Stefanowitsch 2004, Schmid & Küchenhoff 2013)
- Information-theoretic metrics (e.g. entropy and surprisal) => enjoy increasing popularity in psycho- and neurolinguistics circles (Hale 2016)

All metrics for the 30 experimental items were extracted from etTenTen13 (~ 260 million words from 686,000 webpages in Estonian)



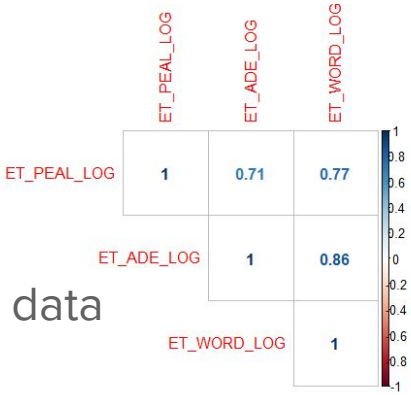
| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | WORD | ENGLISH | ET_ADE | ET_PEAL | ET_WORD | ET_ADE_LOG | ET_PEAL_LOG | ET_WORD_LOG | |
| 2 | autouks | car_door | 3 | 1 | 349 | 0.47712125 | 0 | 2.542825427 | |
| 3 | diivan | couch | 1380 | 102 | 4397 | 3.13987909 | 2.008600172 | 3.643156466 | |
| 4 | hing | soul | 758 | 112 | 43049 | 2.87966921 | 2.049218023 | 4.633963068 | |
| 5 | hoov | yard | 314 | 144 | 5507 | 2.49692965 | 2.158362492 | 3.740915076 | |
| 6 | jõgi | river | 2036 | 125 | 30933 | 3.30877777 | 2.096910013 | 4.490422042 | |
| 7 | katus | roof | 1939 | 36 | 19632 | 3.28757781 | 1.556302501 | 4.292964545 | |
| 8 | keldrikorrus | cellar_floord | 350 | 1 | 690 | 2.54406804 | 0 | 2.838849091 | |
| 9 | kivid | stones | 351 | 50 | 29126 | 2.54530712 | 1.698970004 | 4.464280845 | |
| 10 | korvtool | wicker_chair | 1 | 1 | 40 | 0 | 0 | 1.602059991 | |
| 11 | krossirada | speedway | 222 | 1 | 506 | 2.34635297 | 0 | 2.704150517 | |
| 12 | lauahunnik | pile_of_planks | 1 | 1 | 18 | 0 | 0 | 1.255272505 | |
| 13 | lehekülg | page | 2467 | 13 | 16721 | 3.39216915 | 1.113943352 | 4.223262247 | |
| 14 | lävi | threshold | 694 | 4 | 1759 | 2.84135947 | 0.602059991 | 3.245265839 | |
| 15 | matt | mat | 182 | 26 | 3660 | 2.26007139 | 1.414973348 | 3.563481085 | |
| 16 | mees | man | 8860 | 53 | 299046 | 3.94743372 | 1.72427587 | 5.475737998 | |
| 17 | metsaserv | edge_of_wood | 104 | 1 | 632 | 2.01703334 | 0 | 2.800717078 | |
| 18 | mis | what | 34149 | 763 | 2251892 | 4.53337799 | 2.882524538 | 6.352547558 | |
| 19 | muld | soil | 421 | 26 | 16421 | 2.6242821 | 1.414973348 | 4.215399601 | |
| 20 | paber | paper | 3691 | 316 | 29541 | 3.56714405 | 2.499687083 | 4.470425193 | |
| 21 | pilt | picture | 11754 | 430 | 113613 | 4.07018569 | 2.633468456 | 5.055428028 | |
| 22 | pind | surface | 5391 | 65 | 29524 | 3.73166933 | 1.812913357 | 4.470175197 | |
| 23 | pink | bench | 936 | 134 | 5923 | 2.97127585 | 2.127104798 | 3.772541733 | |
| 24 | praam | ferry | 127 | 52 | 2164 | 2.10380372 | 1.716003344 | 3.335257256 | |
| 25 | redelid | ladders | 32 | 3 | 2017 | 1.50514998 | 0.477121255 | 3.304705898 | |
| 26 | rong | train | 142 | 41 | 13455 | 2.15228834 | 1.612783857 | 4.128883702 | |
| 27 | sõrmed | fingers | 492 | 20 | 11996 | 2.6919651 | 1.301029996 | 4.079036457 | |
| 28 | telgilauake | tent_table | 1 | 1 | 1 | 0 | 0 | 0 | |
| 29 | tool | chair | 1128 | 139 | 9700 | 3.0523091 | 2.1430148 | 3.986771734 | |
| 30 | tänav | street | 15801 | 96 | 56202 | 4.19868457 | 1.982271233 | 4.749751771 | |
| 31 | voodi | bed | 361 | 125 | 19771 | 2.5575072 | 2.096910013 | 4.296028636 | |
| 32 | | | | | | | | | |

fc1 | ratings | frequency

# TOKEN FREQUENCIES

- **ET_ADE_LOG**: log transformed frequency of a word with the adessive construction in EtTenTen
- **ET_PEAL_LOG**: log transformed frequency of a word with peal construction in EtTenTen
- **ET_WORD_LOG**: log transformed frequency of a word in EtTenTen

# Mixed-effects (logistic) regression models




Frequency counts are highly correlated -> cannot be fitted to the data in a single model => 3 different models for the two datasets

1) Model 1a: `CHOICE_CX ~ ET_ADE_LOG + (1|SUBJECT) + (1|WORD)`
2) Model 1b: `RATING ~ ET_ADE_LOG + (1|SUBJECT) + (1|WORD)`
3) Model 2a: `CHOICE_CX ~ ET_PEAL_LOG + (1|SUBJECT) + (1|WORD)`
4) Model 2b: `RATING ~ ET_PEAL_LOG + (1|SUBJECT) + (1|WORD)`
5) Model 3a: `CHOICE_CX ~ ET_WORD_LOG + (1|SUBJECT) + (1|WORD)`
6) Model 3b: `RATING ~ ET_WORD_LOG + (1|SUBJECT) + (1|WORD)`

# Mixed-effects (logistic) regression models

Frequency counts are highly correlated -> cannot be fitted to the data with a single model => 6 different models:

1) Model 1a: `CHOICE_CX ~ ET_ADE_LOG + (1|SUBJECT) + (1|WORD)`
2) Model 1b: `RATING ~ ET_ADE_LOG + (1|SUBJECT) + (1|WORD)`
3) Model 2a: `CHOICE_CX ~ ET_PEAL_LOG + (1|SUBJECT) + (1|WORD)`
4) Model 2b: `RATING ~ ET_PEAL_LOG + (1|SUBJECT) + (1|WORD)`
5) Model 3a: `CHOICE_CX ~ ET_WORD_LOG + (1|SUBJECT) + (1|WORD)`
6) Model 3b: `RATING ~ ET_WORD_LOG + (1|SUBJECT) + (1|WORD)`

**None of the token frequency counts are significant in predicting speakers' preferences**

# A different frequency measure: **RATIO**: ET_ADE_LOG/ET_PEAL_LOG

Model 4a: `CHOICE_CX ~ RATIO + (1|SUBJECT) + (1|WORD)`

Family: binomial ( logit )

Formula: CHOICE_CX ~ RATIO + (1 | SUBJECT) + (1 | LEMMA)

Fixed effects:

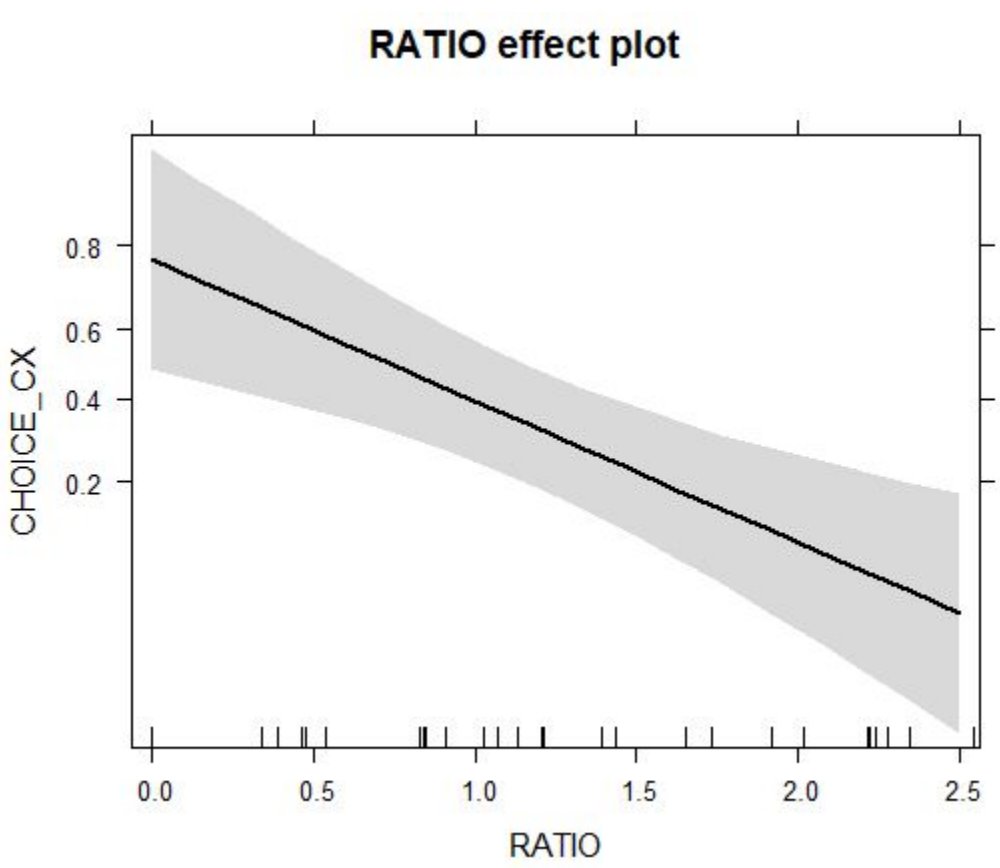
| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | 1.0032 | 0.6826 | 1.470 | 0.141646 | |
| **RATIO** | **-1.6297** | **0.4788** | **-3.404** | **0.000665** | *** |

Model 4b: `RATING ~ RATIO + (1|SUBJECT) + (1|WORD)`

Formula: RATING ~ RATIO + (1 | SUBJECT) + (1 | LEMMA)

Fixed effects:

| | Estimate | Std. Error | df | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|---|
| (Intercept) | 7.6632 | 0.3313 | 35.3498 | 23.131 | <2e-16 | *** |
| **RATIO** | **-0.3059** | **0.2126** | **26.8841** | **-1.438** | **0.162** | |


RATIO effect plot
CHOICE_CX
0.8
0.6
0.4
0.2
0.0
0.5
1.0
1.5
2.0
2.5
RATIO

# Analysis of contingency

- A wide variety of measures are available to determine the degree of association between a cue and an outcome, or, in the case of language, between a linguistic form and its function.

- The following measures are among the most widely used (Gries & Ellis 2015: 23):

(1) a. pointwise $MI = log_2 \frac{a}{a_{expected}}$

b. $z = \frac{a - a_{expected}}{\sqrt{a_{expected}}}$

c. $t = \frac{a - a_{expected}}{\sqrt{a}}$

d. $G^2 = 2 \cdot \sum_{1}^{4} obs \cdot log \frac{obs}{exp}$

e. $-\log 10\, p_{\text{Fisher-Yates exact test}}$

# Table 1. Schematic co-occurrence table of token frequencies for association measures (Gries & Ellis 2015: 236)

| Observed frequencies | Element *y* | Other elements | Totals |
|---|---|---|---|
| Element x | *a* | *b* | *a* + *b* |
| Other elements | *c* | *d* | *c* + *d* |
| Totals | *a* + *c* | *b* + *d* | *a* + *b* + *c* + *d* = *N* |

# Table 1. Schematic co-occurrence table of token frequencies for association measures (Gries & Ellis 2015: 236)

| Observed frequencies | Element *y* = laud 'table' | Other elements | Totals |
|---|---|---|---|
| Element x = ade | *a* = 4745 | *b* = 7595510 | *a* + *b* = **7600255** |
| Other elements | *c* = 34976 | *d* = 252924598 | *c* + *d* |
| Totals | *a* + *c* | *b* + *d* | *a* + *b* + *c* + *d* = *N* |

*N* = corpus size (260,559,829)

# Collostructional metrics

- Calculating the collostructional strength
  - log-likelihood
- All computations were done with Gries' R script for coll.analysis 3.2 (Gries, Stefan Th. 2007. Coll.analysis 3.2a. A program for R for Windows 2.x.)
- Fitting the mixed-effects models with the collostructional metrics as predictors -> the metrics are strongly correlated => different models

1) Model 1a: `CHOICE_CX ~ ADE_COLL + (1|SUBJECT) + (1|WORD)`
2) Model 1b: `RATING ~ PEAL_COLL + (1|SUBJECT) + (1|WORD)`
3) Model 2a: `CHOICE_CX ~ PEAL_COLL + (1|SUBJECT) + (1|WORD)`
4) Model 2b: `RATING ~ PEAL_COLL + (1|SUBJECT) + (1|WORD)`

# Collostructional metrics

- Calculating the collostructional strength
  - log-likelihood
- All computations were done with Gries' R script for coll.analysis 3.2 (Gries, Stefan Th. 2007. Coll.analysis 3.2a. A program for R for Windows 2.x.)
- Fitting the mixed-effects models with the collostructional metrics as predictors -> the metrics are strongly correlated => different models

1) Model 1a: `CHOICE_CX ~ ADE_COLL + (1|SUBJECT) + (1|WORD)`
2) Model 1b: `RATING ~ PEAL_COLL + (1|SUBJECT) + (1|WORD)`
3) Model 2a: `CHOICE_CX ~ PEAL_COLL + (1|SUBJECT) + (1|WORD)`
4) Model 2b: `RATING ~ PEAL_COLL + (1|SUBJECT) + (1|WORD)`

**None of the collostructional metrics are significant in predicting speakers' preferences**

## **RATIO**: ET_ADE_LOG/ET_PEAL_LOG

### Model 4a: `CHOICE_CX ~ RATIO + (1|SUBJECT) + (1|WORD)`

Family: binomial ( logit )

Formula: CHOICE_CX ~ RATIO + (1 | SUBJECT) + (1 | LEMMA)

Fixed effects:

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | 1.0032 | 0.6826 | 1.470 | 0.141646 | |
| **RATIO** | **-1.6297** | **0.4788** | **-3.404** | **0.000665** | ******* |

### Model 4b: `RATING ~ RATIO + (1|SUBJECT) + (1|WORD)`

Formula: RATING ~ RATIO + (1 | SUBJECT) + (1 | LEMMA)

Fixed effects:

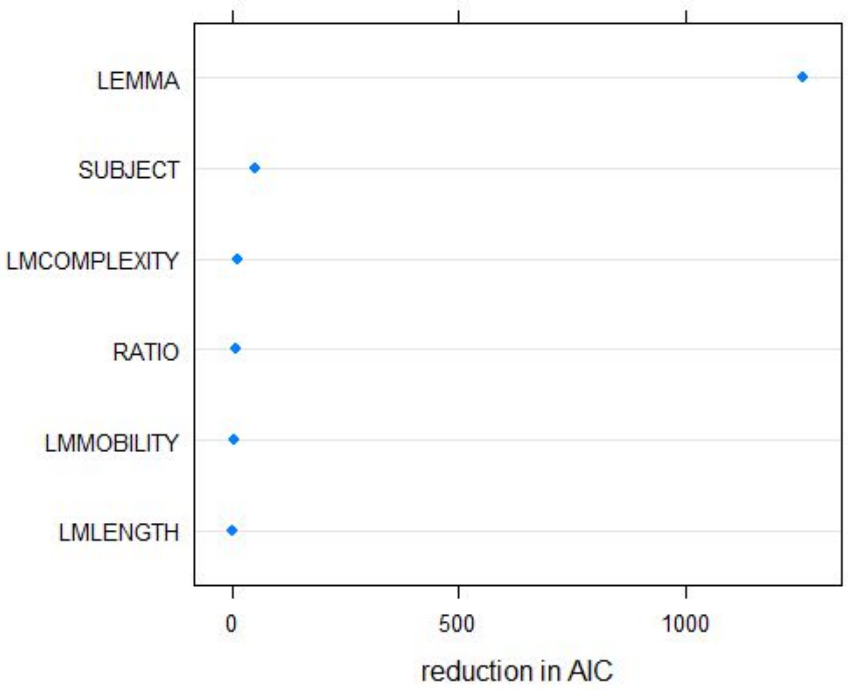
| | Estimate | Std. Error | df | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|---|
| (Intercept) | 7.6632 | 0.3313 | 35.3498 | 23.131 | <2e-16 | *** |
| **RATIO** | **-0.3059** | **0.2126** | **26.8841** | **-1.438** | **0.162** | |

# Adessive vs *peal* (Klavan, J. aop. Pitting corpus-based classification models against each other: A case study for predicting constructional choice in written Estonian. *Corpus Linguistics and Linguistic Theory*)

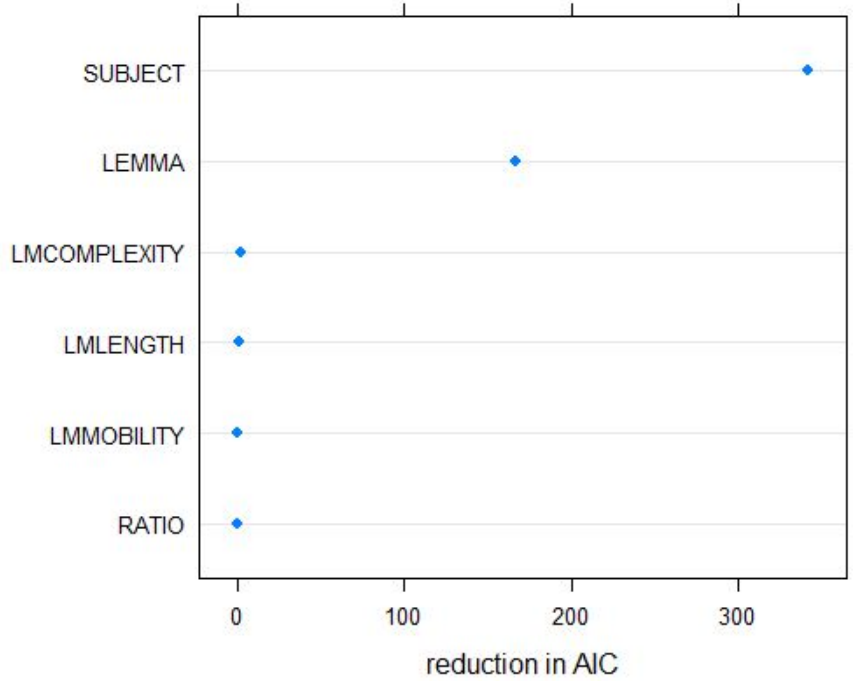
Table 3. Model comparison statistics

| | Resid. Dev | Df | Deviance | *p*-value | Reduction in AIC |
|---|---|---|---|---|---|
| Intercept | 1247.7 | | | | |
| LENGTH | 1155.7 | 1 | 91.968 | 0.0000 | 90.0 |
| COMPLEXITY | 1115.7 | 1 | 39.965 | 0.0000 | 38.0 |
| MOBILITY | 1079.9 | 1 | 35.807 | 0.0000 | 33.8 |
| VERBGROUP | 1067.2 | 4 | 12.681 | 0.0000 | 4.7 |
| TRWC | 1051.3 | 2 | 15.967 | 0.0000 | 12.0 |
| POSITION | 1047.2 | 1 | 4.088 | 0.0000 | 2.1 |

Formula: CHOICE_CX ~ LMLENGTH + LMMOBILITY + LMCOMPLEXITY + RATIO + (1 | SUBJECT) + (1 | LEMMA)




Formula: RATING ~ LMLENGTH + LMMOBILITY + LMCOMPLEXITY + RATIO + (1 | SUBJECT) + (1 | LEMMA)

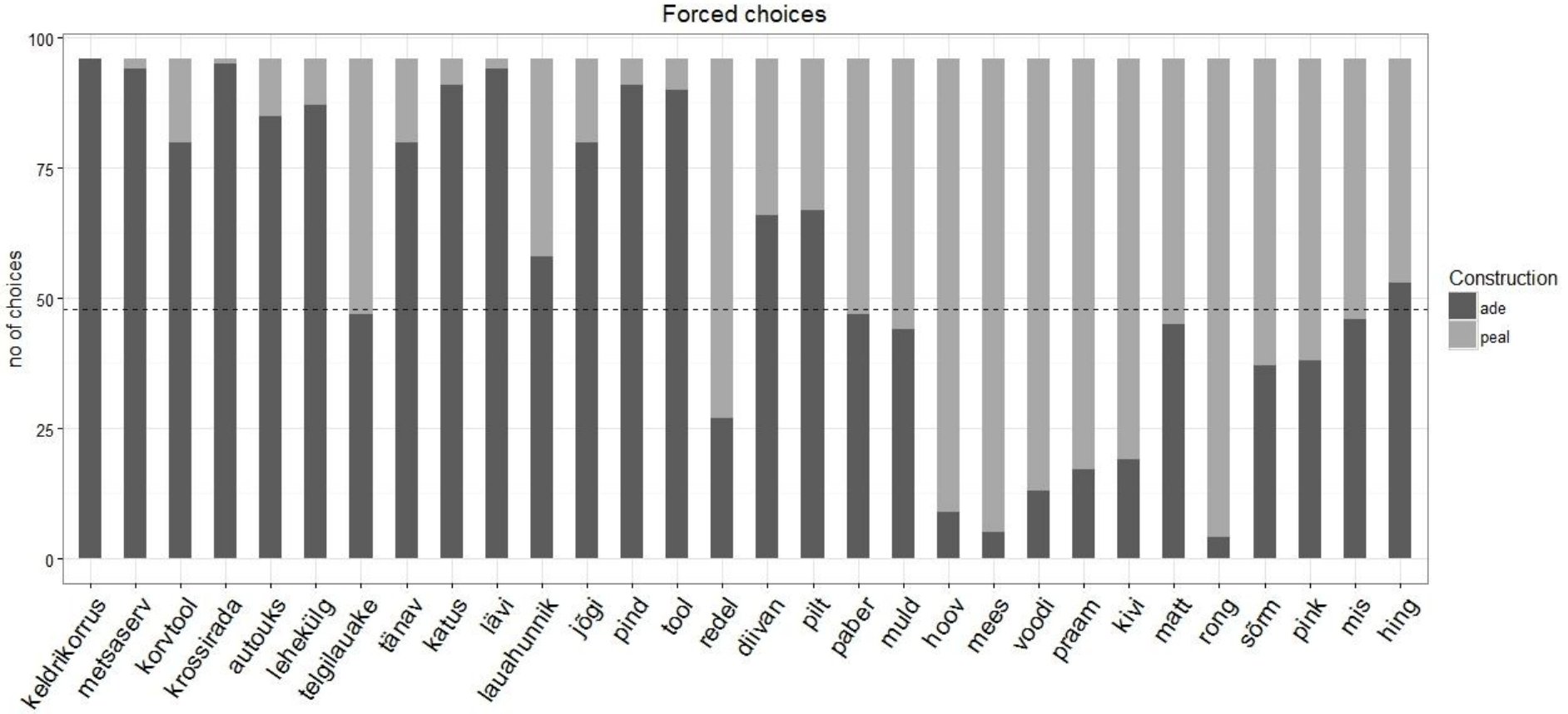


# Discussion: type vs token frequencies

- “Recent work shows that in syntax, as in phonology, the productivity of pattern depends on type frequency of the construction”. (Ellis 2002: 145)
  - adessive = 7,600,255 tokens in etTenTen13 (274,688 types)
  - *peal* = 59,873 tokens in etTenTen13 (9,759 types)
- cf. present-day written language (Klavan 2012):
  - adessive = 450 tokens (255 types)
  - *peal* = 450 tokens (209 types)
- Speakers seem to be attuned to the global frequencies of the two constructions

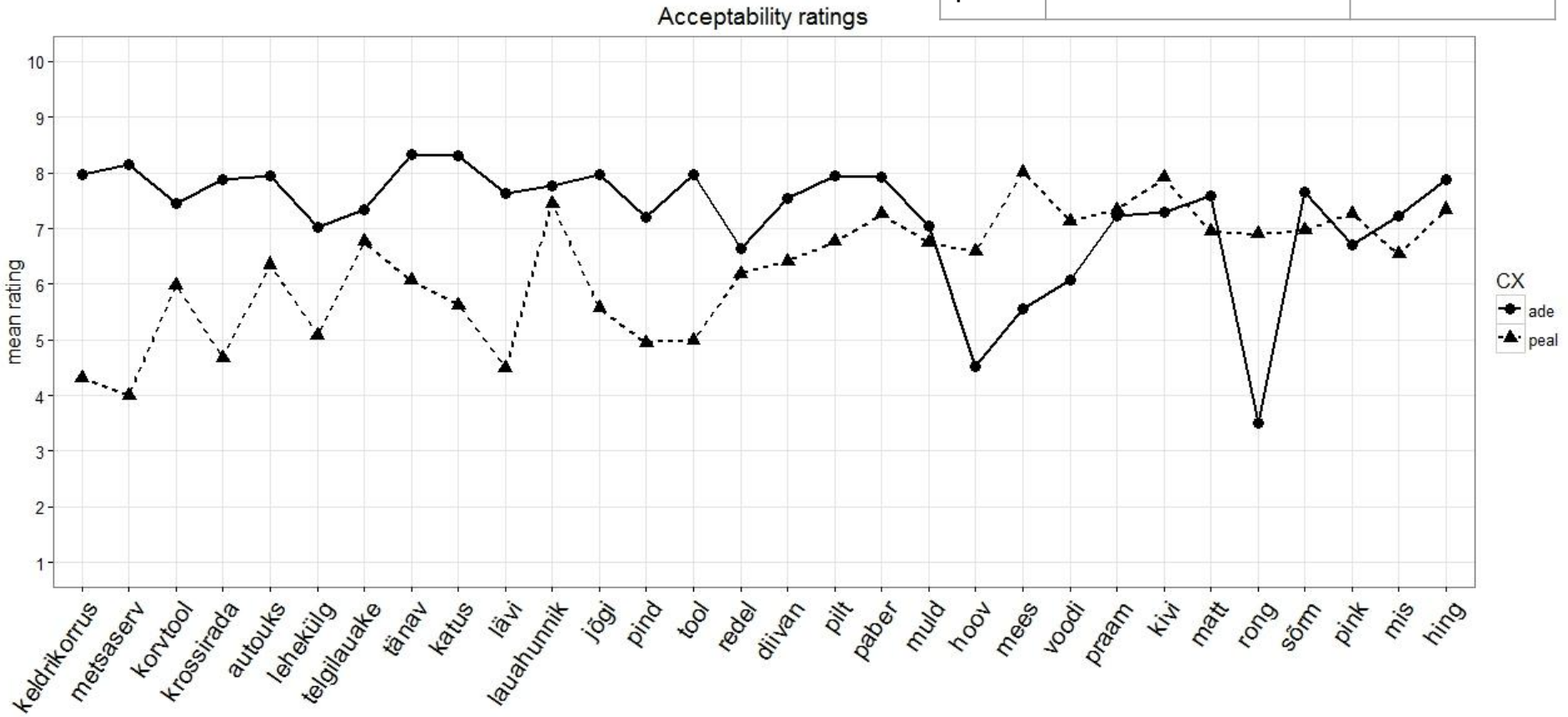
# Results: forced choice data

```
 ade peal  Sum
1705 1175 2880
```



# Results: acceptability ratings

| | | |
|---|---|---|
| ade | Mean = 8.05 | SD = 2.8 |
| peal | Mean = 6.64 | SD = 3.2 |



# Discussion: issues when counting frequencies

- What to count as the adessive construction? Adessive in the locative function vs other functions

**7,600,255** = the total number of all of the adessive tokens in the corpus

| Observed frequencies | Element *y* = laud 'table' | Other elements | Totals |
|---|---|---|---|
| Element x = ade | *a* = 4745 | *b* = 7595510 | *a* + *b* = **7600255** |
| Other elements | *c* = 34976 | *d* = 252924598 | *c* + *d* |
| Totals | *a* + *c* | *b* + *d* | *a* + *b* + *c* + *d* = *N* |

# The adessive construction = frequency of what?

(6) Functions of the Estonian adessive case (Erelt et al. 2007: 250):

a. Location:

| *Vaas* | *on* | ***laual.*** |
|---|---|---|
| vase.SG.NOM | be-PRS.3SG | table.SG.ADE |

'The vase is **on the table**.'

b. Time:

| *Nad* | *sõidavad* | ***neljapäeval*** | *maale.* |
|---|---|---|---|
| they.NOM | drive-PRS.3PL | Thursday.SG.ADE | country.SG.ALL |

'They are driving to the country **on Thursday**.'

c. State: *Jüri* *vaatas* *meid* ***naerul*** *näoga.*
Jüri.NOM look-PST.3SG us laugh.SG.ADE face.SG.COM
'Jüri looked at us **with a laughing** face.'

d. Possessor: ***Maril*** *on* *kaks* *last.*
Mari.ADE be-PRS.3PL two child.SG.PRT
'**Mari** has two children.' (lit. '**On Mari** are two children.')

e. Agent with finite verb forms:
*See* *asi* *ununes* ***mul*** *kiiresti.*
this.SG.NOM thing.SG.NOM forget-PRS.3SG me.SG.ADE quickly
'**I** quickly forgot about that thing.'

f. Instrument: *Mari* *mängib* ***klaveril*** *mõnd lugu.*
Mari.NOM play-PRS.3SG piano.SG.ADE some tune.SG.PART
'Mari is playing some tunes **on the piano**.'

g. Manner: *Mari* *kuulas* *kikkis* ***kõrvul.***
Mari.NOM listen-PST.3SG pricked.up ear.PL.ADE
'Mari listened **with her ears** pricked up.'

# Conclusions

- **Language users' preferences are influenced by the relative frequencies with which certain nouns appear with different locative cases and postpositions (ratio of $freq_{ade+word}$ / $freq_{peal+word}$)**
  - => artefact of the experiment?
- Different linguistic tasks might be more or less susceptible to quantitative effects (McConnell & Blumenthal-Dramé, in revision 2018)
  - For example, tasks that encourage subjects to attend closely to individual linguistic units are likely to disrupt statistical links between units
  - Tasks that highlight the competition between two alternants might be particularly sensitive to metrics comparing the likelihood of these alternants (cf. the effect of ratio)

# We need follow-up experiments (work in progress)

- Different metrics (comparing backwards and forwards looking metrics)
  - Prediction: forward looking metrics are more relevant in online processing tasks than in off-line tasks
- Using fillers => making the competition of two constructions less salient
  - Prediction: metrics taking into account the whole probability distribution between words and constructions (not only the competing ones) become significant (e.g. entropy)
- Quantifying not only associations between words and constructions, but words and the constructional functions
  - Cf. the polysemy of the adessive case (manual coding)

# References


- Arppe, Antti & Dana Abdulrahim. 2013. Converging linguistic evidence on two flavors of production: The synonymy of Arabic COME verbs. Second Workshop on Arabic Corpus Linguistics, University of Lancaster, UK, 22-26 July, 2013.
- Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the Dative Alternation. In Gerlof Bouma, Irene Krämer & Joost Zwarts (eds.), *Cognitive Foundations of Interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.
- Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in Search of Its Evidential Base*, 77–96. Berlin: Mouton de Gruyter.
- Divjak, Dagmar, Antti Arppe & Ewa Dąbrowska. 2016. Machine meets man: Evaluating the psychological reality of corpus-based probabilistic models. *Cognitive Linguistics* 27(1). 1-33.
- Gries, Stefan Th. 2007. Coll.analysis 3.2a. A program for R for Windows 2.x.
- Gries, Stefan Th. and Nick C. Ellis. 2015. Statistical Measures for Usage-Based Linguistics. *Language Learning* 65 (1), 228–255.
- Gries, Stefan Th. and Anatol Stefanowitsch. 2004. Extending collostructional analysis: a corpus-based perspective on alternations. *International Journal of Corpus Linguistics* 9 (1), 97–129.
- Hale, John. 2016. Information-theoretical Complexity Metrics. *Language and Linguistics Compass* 10 (9), 397–412.

# References


- 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. To appear. Pitting corpus-based classification models against each other: A case study for predicting constructional choice in written Estonian. *Corpus Linguistics and LInguistic Theory.*
- Klavan, Jane, Maarja-Liisa Pilvik, and 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.
- 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. 2016. The Cognitive Plausibility of Statistical Classification Models: Comparing Textual and Behavioral Evidence. *Folia Linguistica* 50(2), 355−384.
- McConnell, Kyla & Alice Blumenthal-Dramé. In revision 2018. Effects of task and corpus-derived association scores on the online processing of collocations.
- Schmid, Hans-Jörg and Helmut Küchenhoff. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24 (3), 531-577.