Combining corpus-linguistic and experimental methods for the study of constructional alternations

Jane Klavan <jane.klavan@gmail.com>
University of Tartu, Department of English Studies

Combining corpus-linguistic and experimental methods for the study of constructional alternations

In my talk I focus on combining various linguistic methods to find out what we can infer about linguistic variation from the patterns and structures we see in the language data. The case study I present comes from the Estonian language and pertains to the morpho-syntactic alternation between the adessive case suffix and the adposition *peal* 'on'. I will first fit different machine classifiers (e.g. mixed-effects logistic regression, NDL) to the corpus data in order to see which variables contribute significantly to the model fit. The second part of the talk looks at the results of linguistic experiments (a forced choice task and an acceptability rating task) carried out on the same phenomenon. The discussion of the talk highlights some of the pros and cons of combining different methods for the study of constructional alternations. I will take stock with the issue of whether it makes sense to talk about experimental validation of corpus-based studies or whether we are comparing apples and oranges.

Outline

- Introduction:
  - Why linguistic methodology?
  - Why constructional alternations?
- Phenomenon: constructional alternation in Estonian
  - Study 1: Prediction accuracy of men and machines
  - Study 2: Ratings vs acceptability judgements
- Interim conclusions
- Avenues for future research

Why linguistic methodology?

Cognitive Linguistics

Because cognitive linguistics sees language as embedded in the overall cognitive capacities of man, topics of special interest for cognitive linguistics include: the structural characteristics of natural language categorization (such as prototypicality, systematic polysemy, cognitive models, mental imagery and metaphor); the functional principles of linguistic organization (such as iconicity and naturalness); the conceptual interface between syntax and semantics (as explored by cognitive grammar and construction grammar); the experiential and pragmatic background of language-in-use; and the relationship between language and thought, including questions about relativism and conceptual universals.

Geeraerts 1995: 111-112

http://www.cognitivelinguistics.org/en/about-cognitive-linguistics

Cognitive Linguistics

Because cognitive linguistics sees language as embedded in the overall cognitive capacities of man, topics include: the structural characteristics of natural language categorization (such as prototypicality, systemati and metaphor); the functional principles of linguistic organization (such as iconicity and naturalness); the c semantics (as explored by cognitive grammar and construction grammar); the experiential and pragmatic b relationship between language and thought, including questions about relativism and conceptual universa

Jane Klavan

@JaneKlavan

Usage-based linguist with a bent for methodology

Tartu, Eesti

sisu.ut.ee/janeklavan

Joined October 2016

Photos and videos

Geeraerts 1995: 111-112

http://www.cognitivelinguistics.org/en/about-cognitive-linguistics

Doing linguistics during the “quantitative turn”

(2016 Special Issue in Cognitive Linguistics, edited by Divjak, Levshina & Klavan)

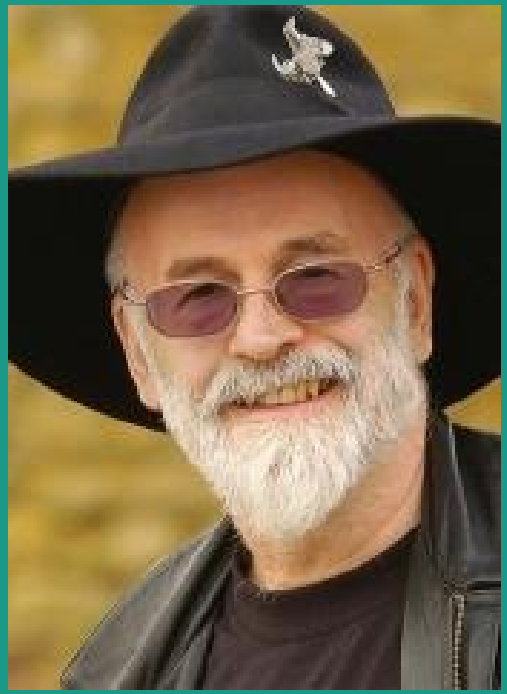
- It is the best of times:
  - Many new methods for data collection and data analysis (e.g. statistical modelling)
  - The age of digitalisation and “big data”
- It is the worst of times:
  - We lack sufficient understanding of what is the nature of the data obtained via the various methods
  - We lack sufficient understanding of what are the underlying assumptions about language made by various algorithms

Why constructional alternations?

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 one and the same function, they do it in different ways: they allow for a different construal of the same situation (*the no-synonymy hypothesis*)
- We may assume that speakers’ choice between alternative forms is influenced by a multitude of factors: semantic, syntactic, morphological, phonological, discourse-related, and lectal features

“There is always a choice”



(Terry Pratchett. 2004. *Going Postal*)

“... 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.

Two studies on a constructional alternation in Estonian

Table 17. The system of adverbial cases (Viitso 2003: 33)

	Directional	Static	Separative
Interior	Illative *ˊkivisse* 'into the stone'	Inessive *ˊkivis* 'in the stone'	Elative *ˊkivist* 'from the stone'
Exterior	Allative *ˊkivile* 'onto the stone'	Adessive *ˊkivil* 'on the stone'	Ablative *ˊkivilt* 'off the stone'
Limited	Terminative *ˊkivini* 'up to the stone'		
Existential	Translative *ˊkiviks* 'into the state of being the stone'	Essive *ˊkivina* 'as the stone'	
Instrumental		Comitative *ˊkiviga* 'with the stone'	Abessive *ˊkivita* 'without the stone'

Adessive case vs *peal* ‘on’ in Estonian

(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**.’

ADESSIVE (*-l*)

LOCATION (on):

Raamat	on	**laua-l**.
book.NOM	be.3SG	table-ADE

“The book is on the table.”

TEMPORAL:

Professori	loeng	on	**esmaspäeva-l**.
professor.GEN	lecture.NOM	be.3SG	Monday-ADE

“Professor’s lecture is on Monday.”

POSSESSION:

Professori-l	on	uus	raamat.
professor-ADE	be.3SG	new.NOM	book.NOM

“The professor has a new book.”

Data from Klavan (2012)

- extraction of contextual data, i.e. semantic and morpho-syntactic information found within clause boundaries, from the corpus of present-day written Estonian
- random sample of 900 occurrences (450 per construction)
- fiction (108 authors) and newspaper texts from 1980s to 2000s
    - the Morphologically Disambiguated Corpus (MDCE 2015, size 215,000 words)
    - the Balanced Corpus of Estonian (BCE 2015; size 10 million words)

Data from Klavan (2012)

Table 1. Annotation schema (listed alphabetically)

Variable name	Levels
ANIMACY (animacy of LM)	animate, inanimate
CLAUSE	main, subordinate
COMPLEXITY (morphol. complexity of LM)	compound, simple
CONSTRUCTION (response)	adessive, peal
LEMMA (lemma of the word used as the LM)	397 lemmas
LENGTH (length of LM phrase in syllables)	from 1 to 41 syllables (log. transf.)
LMNR (number of LM)	plural, singular
LMTRSIZE (relative size of TR & LM)	conventional, same, unconventional
LMWC (word class of LM)	noun, pronoun
MOBILITY (mobility of LM phrase)	mobile, static
POSITION (relative position btw TR & LM)	lm_tr, tr_lm

Variable name	Levels
RELTYPE (type of relation btw LM & TR)	abstract, spatial
SYNFUN (syntactic function of LM)	adverbial, modifier
TRANIM (animacy of TR)	animate, inanimate
TRCASE (case form of TR)	nom., part., other, not applicable
TRMOBILITY (mobility of TR)	mobile, static
TRNR (number of TR)	plural, singular
TRTYPE (type of TR)	abstract, object
TRWC (word class of TR phrase)	NP, other
TYPE (type of LM)	place, thing
VERBGROUP	action, existence, motion, posture, no verb
WOPOSITION (position of LM phrase)	initial, middle, final

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.’

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	EXAMPLE	CX	VERB	LEMMA	SOURCE	RELTYPE	LENGTH	LENGTHLC	TYPE	ANIMACY	MOBILITY	COMPLEXITY	LMWC
2	Malka istus	ade	istuma	korvtool	fiction	spat	10	1.38629	thing	inanim	mobile	compound	noun
3	**suvekohviku valgel**												
4	**korvtoolil** ja luges	LMNR	SYNFUN	TRWC	TRANIM	TRMOBILI	TRNR	TRCASE	TRMEANI	TRTYPE	POSITION	WOPOSITION	CLAUSE
5	midagi.	sg	adverbial	noun	anim	mobile	sg	nom	object	object	tr_lm	final	main
6													

Two studies:

1. Prediction accuracy of men and machines (Klavan *to appear*)
2. Ratings vs acceptability judgements (Klavan & Veismann 2017)

Background: Klavan (2012)

Klavan (2012) fitted a binary logistic regression model to the corpus data in order to determine which of the variables are more decisive and predictive for the choice between the two constructions

Statistic	Model 1
model L.R.	200.48 (df = 10)
generalised R^2	0.266
C	0.761
Somers' D_{xy}	0.522

```
Call:
glm(formula = CONSTRUCTION ~ LM_LENGTHSYLLOG + LM_COMP
LM_MOBILITY + VERB_GROUP + TR_WC + WO_LM, family = "binomial
data = basic)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.14872  -0.96006   0.03036   0.94679   2.30562

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.01344    0.34283   0.039 0.968731
LM_LENGTHSYLLOG    -0.92625    0.13590  -6.816 9.39e-12 ***
LM_COMPsimple       1.16446    0.22373   5.205 1.94e-07 ***
LM_MOBILITYstatic  -0.89143    0.15347  -5.808 6.30e-09 ***
VERB_GROUPexistence 0.70680    0.20904   3.381 0.000722 ***
VERB_GROUPmotion    0.03419    0.24192   0.141 0.887624
VERB_GROUPnoverb   -0.15370    0.24082  -0.638 0.523338
VERB_GROUPposture   0.06270    0.23049   0.272 0.785581
TR_WCpronoun        0.63322    0.20578   3.077 0.002090 **
TR_WCVP             0.54636    0.20992   2.603 0.009251 **
WO_LMtr_lm          0.32789    0.16261   2.016 0.043759 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1247.7  on 899  degrees of freedom
Residual deviance: 1047.2  on 889  degrees of freedom
AIC: 1069.2

Number of Fisher Scoring iterations: 4
```

Figure 4. Output for binary logistic regression model (multivariate corpus study)

Background: Klavan (2012)

- Given that we are we are predicting a choice between two near-synonyms & both alternatives can be used in all of the studied contexts:
  - ➢ Is it realistic to hope for classification accuracy above $C = 0.76$?
  - ➢ How good is human classification behaviour?

$C = 0.5$ – no discrimination
$0.5 < C < 0.7$ – poor discrimination
$0.7 \leq C < 0.8$ – acceptable discrimination
$0.8 \leq C < 0.9$ – excellent discrimination
$C \geq 0.9$ – outstanding discrimination

(Hosmer et al. 2013: 177)

Study 1: Prediction accuracy of men and machines

Klavan, J. 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.*

Study 1: Previous studies

1. The performance of alternative statistical modelling techniques is compared by pitting them against each other on one and the same dataset (Baayen et al. 2013, Theijssen et al. 2013, Baayen 2011)
   ➢ *classification is similar across the different techniques*
2. The performance of a corpus-based model is explicitly compared to the classification behaviour of native speakers in (psycho)linguistic experiments (Bresnan 2007, Bresnan et al. 2007, Divjak et al. 2016, Arppe & Abdulrahim 2013)
   ➢ *the performance of the corpus-based model, by and large, reflects human behaviour*

Study 1: Prediction accuracy of men and machines

- Two distinct modelling techniques (mixed-effects logistic regression & Naive Discriminative Learning) are applied to predict the choice between two constructional alternatives in written Estonian
- Human performance in the forced choice task
- Two aims:
  1. to compare the classification accuracy of both models in order to assess their usefulness in modelling constructional choice
  2. to set the upper and lower boundaries for human classification behaviour and to compare it to the performance of corpus-based models

Study 1: Research question

- The central idea is to have both the machine classifiers and native speakers perform one and the same task in an equally artificial setting on one and the same set of data

How well do corpus-based models perform compared to each other and compared to native speakers?

Mixed-effects logistic regression
(Pinheiro & Bates 2002)

- uses an algorithm that maximize likelihood using optimization techniques
- do humans exhibit (near-)otpimal behaviour? (Milin et al. 2016: 508)
- one of the best classifiers available

Naive Discriminative Learning (NDL)
(Baayen 2011, Milin et al. 2016)

- a classifier that provides a cognitively grounded framework for classification
- uses an algorithm that makes use of the Rescorla-Wagner rule (Rescorla & Wagner 1972) which defines how a system learns from its own errors and iteratively corrects the erroneous predictions for upcoming events (Milin et al. 2016)

Study 1: Model building

- model formula for NDL:

```
CONSTRUCTION ~ RELTYPE + LMTRSIZE + TYPE + LMANIMACY + LMWC + LMNR +
SYNFUN + TRANIM + TRMOBILITY + TRNR + TRCASE + TRTYPE + CLAUSE +
LENGTH + COMPLEXITY + MOBILITY + VERBGROUP + TRWC + POSITION +
WOPOSITION + LEMMA
```

- model formula for mixed-effects logistic regression:

```
CONSTRUCTION ~ LENGTH + COMPLEXITY + MOBILITY + TRWC + (1|LEMMA)
```

Study 1: Evaluation of model fit

Table 2. Model accuracies (overall accuracy, improvement over baseline, and *C* measure) for the two corpus models

Model	Accuracy	Improvement	*C* value
logistic regression	80%	1.6	0.88
NDL	89%	1.8	0.96

*Overall accuracy = cross-tabulating the two possible outcomes by high and low probabilities based on a cut-off point set at 0.5; the model makes a correct prediction if the estimated probability for *peal* construction is greater than or equal to 0.5 and the *peal* construction was actually observed in the data

**Overfitting is detected by the NDL: under ten-fold cross-validation, *C* value ranges from 0.69 to 0.84 (mean = 0.76), and accuracy from 66% to 78% (mean = 71%).

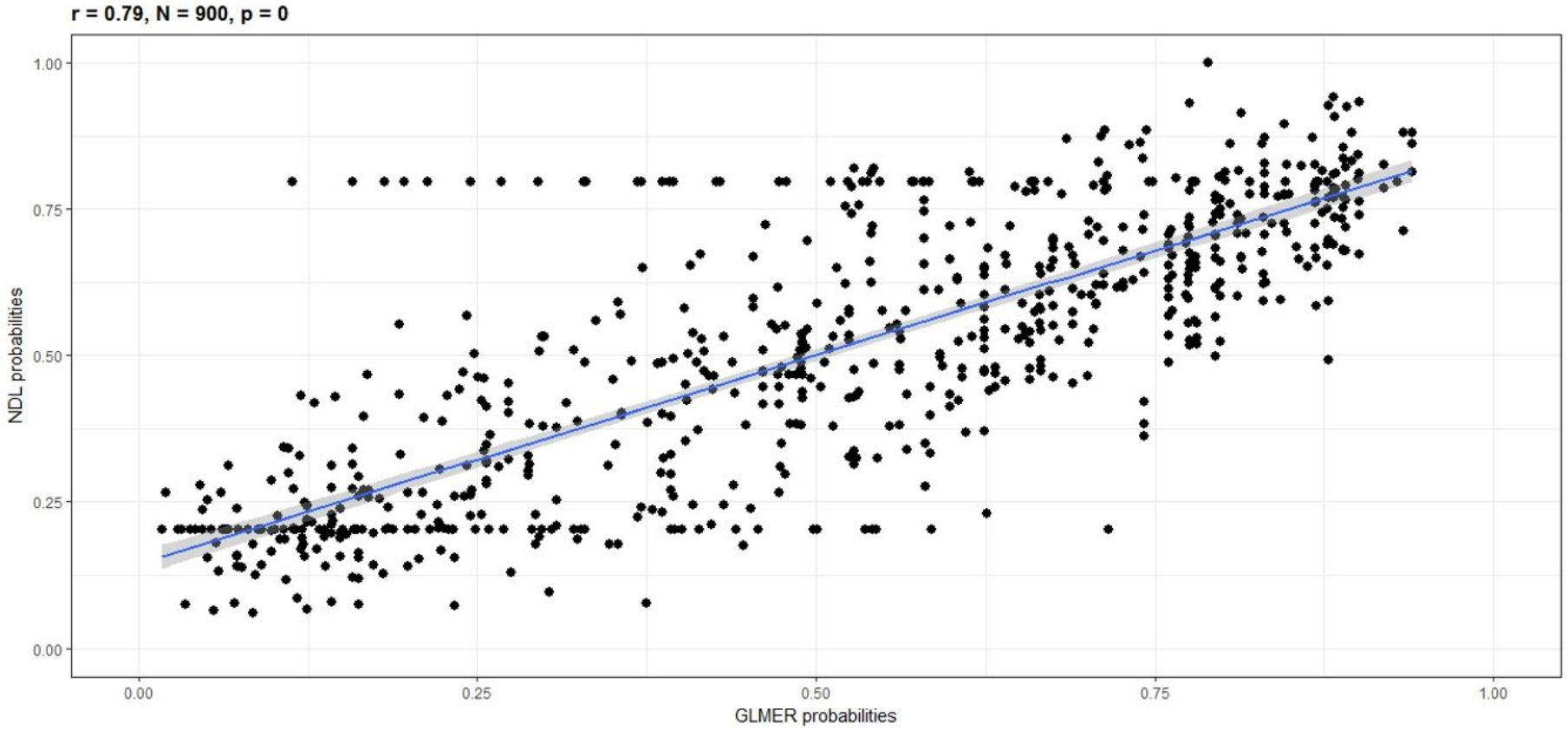


Figure 1. The probabilities estimated by the mixed-effects logistic regression (GLMER) and the NDL-based models for the two constructions attested in the corpus sample

Table 3. Model comparison statistics for the mixed-effects logistic regression corpus-model

	logLik	Chisq	Chi.Df	*p*-value	Reduction in AIC
LEMMA	-589.97				65.7
LENGTH	-553.29	73.356	1	0.000	71.4
COMPLEXITY	-534.22	38.154	1	0.000	36.2
MOBILITY	-524.86	18.717	1	0.000	16.7
TRWC	-517.00	15.716	1	0.000	13.7

Table 3. Model comparison statistics for the mixed-effects logistic regression corpus-model

	logLik	Chisq	Chi.Df	*p*-value	Reduction in AIC
LEMMA	-589.97				65.7
LENGTH	-553.29	73.356	1	0.000	71.4
COMPLEXITY	-534.22	38.154	1	0.000	36.2
MOBILITY	-524.86	18.717	1	0.000	16.7
TRWC	-517.00	15.716	1	0.000	13.7

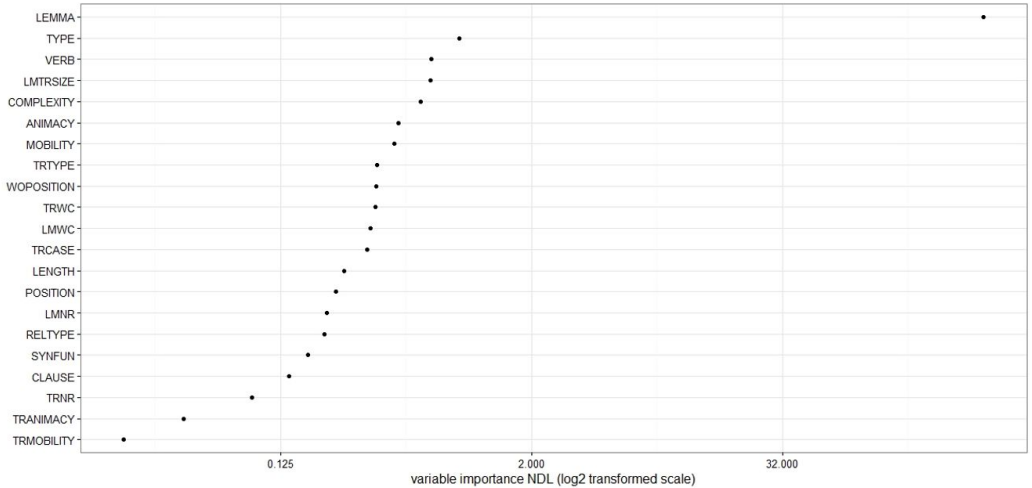


Figure 2. Variable importance for the NDL-based corpus model

Table 4. Coefficients for the mixed-effects logistic regression model of corpus data

	Estimate	Std. Error	z-value	*p*-value
Intercept	0.244	0.387	0.630	0.5288
LENGTH	-1.075	0.179	-5.991	0.0000
COMPLEXITY = simple	1.517	0.300	5.052	0.0000
MOBILITY = static	-0.958	0.219	-4.363	0.0000
TRWC = other	0.730	0.189	3.858	0.0001

Study 1: Corpus-based results (machine vs machine)

Study 1 reflects the findings of the previous studies comparing alternative modelling techniques on one and the same data (Baayen 2011):

- models perform at an almost comparable level in terms of classification accuracy
- the ranking of the predictors differs

Study 1: Corpus-based results

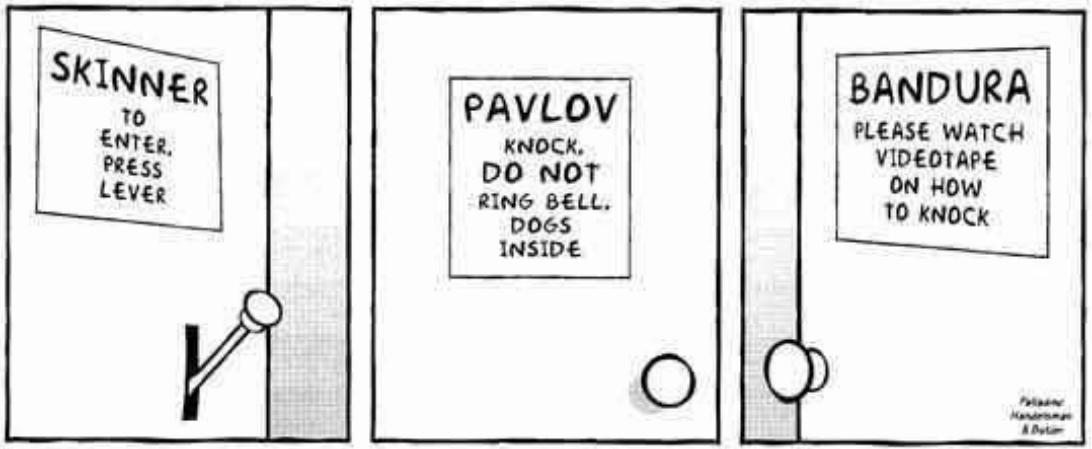
The two models provide a good fit:

- the fit is not outstanding, but we are predicting a choice between two near-synonyms
- relatively similar underlying probabilities are only to be expected since, in principle, both alternatives can be used in all of the studied contexts

Study 1: Corpus-based results

- Is the best performing model necessarily a model that is cognitively plausible?
  - ➢ As linguists we are looking for a model that is sufficiently accurate while at the same time giving useful information about the linguistic phenomenon
- Is native speaker performance as attested in a linguistic experiment comparable to the two corpus-based models?

Study 1: Enter linguistic experiments



Study 1: Research question

- The central idea is to have both the machine classifiers and native speakers perform one and the same task in an equally artificial setting on one and the same set of data

How well do corpus-based models perform compared to each other and compared to native speakers?

Study 1: Classification behaviour of native speakers in a forced choice task (cf. Bresnan 2007, Divjak et al. 2016)

1. should native speaker performance be on a par with that of the model fitted to corpus data, we can add certainty to the conclusion that the model we have selected “has a good fit”;
2. should native speaker performance be inferior to the corpus-based model, it may be suspected that the model is more complex than the actual reality;
3. should native speaker performance be superior to the corpus-based model, there are most likely some important predictors missing from the model formula.

Study 1: Forced choice task

(cf. Bresnan 2007, Divjak et al. 2016)

- 30 corpus sentences with a blank for the original construction followed by the two constructional alternatives
- the sentences represent the full probability scale:
  - ➢ the stimuli ranged from sentences where one construction was very probable (near-categorical preferences) to sentences where both constructions were equally probable (approximately equal probability estimates for both choices) according to the binary logistic regression model fitted by Klavan (2012)

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
- “Which of the two constructions suits into the blank better?”
- items were pseudo-randomized
- four versions of the questionnaire to diminish potential order effects

Study 1: Participants & Procedure

- 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)
- each subject completed the task with the same 30 sentences
  - one sentence at a time, not possible to change their answers
  - ~ 10 minutes to complete it

Study 1: Results of the forced choice task

- Analysis 1: classification accuracy of corpus-based models vs native speakers as a group
  - a direct comparison of how the corpus models performed compared to native speakers as a group
  - Which of the two corpus-based models performs closest to native speakers in terms of overall accuracy?
- Analysis 2: agreement at the level of individual items
  - graphical explorations allow us to assess the agreement between native speakers and the two corpus-based models at the level of individual experimental items

Study 1: Analysis 1 (cf. Divjak et al. 2016)

- 30 sentences used in the task are excluded from the corpus dataset
- corpus-models are trained on the remaining 870 sentences
- probability of the two constructions in the 30 sentences computed based on the re-fitted models
- “correct” response is taken to be the construction that is actually used in the original sentence
- chance performance is 15/30

Table 5. Performance of the corpus-based models and native speakers

Model	**Accuracy**	**Improvement**
logistic regression	28/30 = 93%	1.8
NDL	27/30 = 90%	1.7
native speakers	as a group	23/30 = 77%
	highest performance	28/30 = 93%
	lowest performance	14/30 = 47%

Study 1: Results of Analysis 1

- considerable individual variation (the scores range from 14 to 28 out of 30)
  - ➢ Divjak et al. 2016: different participants rely on different features and collectively they have access to more information than any one individual alone
- corpus-based models are doing an exceptionally good job:
  - ➢ chance: a different set of stimuli ->model performance lower?
  - ➢ models more complex than the representations native speakers operate with?
- Both models are over-optimistic since their prediction accuracy is higher than that of native speakers as a group

Study 1: Analysis 2

To compare the agreement between corpus predictions and native speaker choices for the set of 30 experimental items we can look at:

- the log of the ratio of adessive and *peal* choices for the forced choice task pitted against the probabilities of the two corpus models

* The log odds are calculated by hand; 1 is added to all counts before taking the log in order to avoid dividing by zero:

logit = log((number of adessive constructions + 1)/(number of peal constructions + 1))

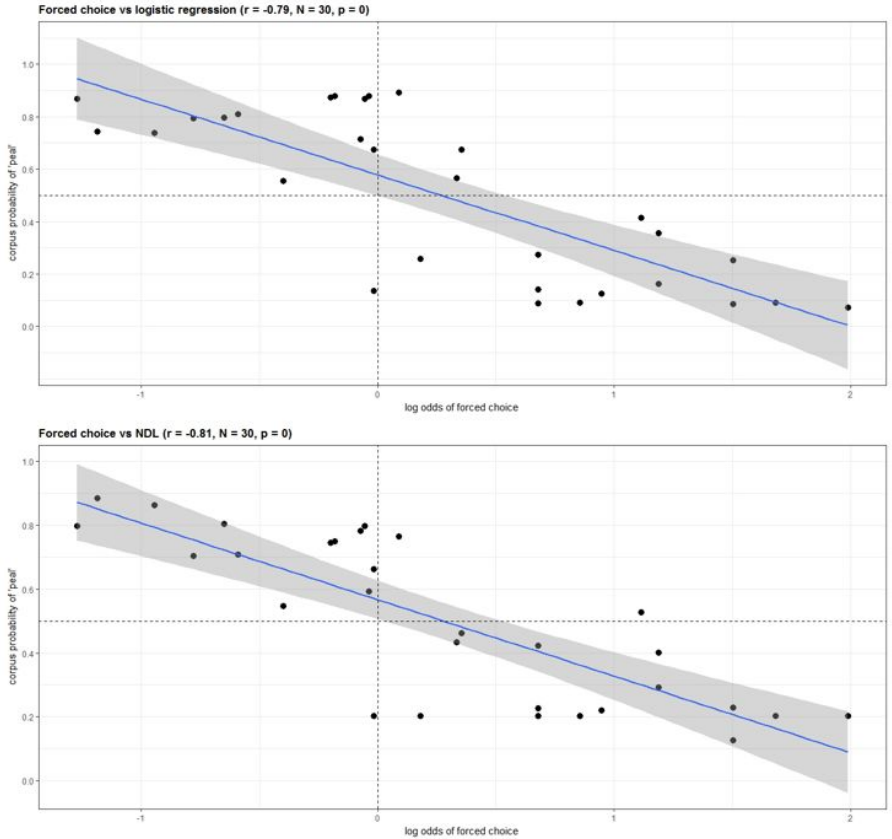


Figure 5. The log odds (of adessive vs *peal*) for each of the 30 experimental items plotted against the respective corpus probabilities of the *peal* construction estimated by the mixed-effects regression model (upper panel) and the NDL-based model (lower panel)

Study 1: Results of Analysis 2

- the default choice for native speakers is the adessive construction

➢ participants frequently chose the adessive construction for items where both the original as well as the predicted construction was the *peal* construction

Possible explanation:

- the adessive construction is 10 times more frequent than the *peal* construction (Klavan 2012: 182-183)
- native speakers are attuned to such global frequency information
- frequency information is not included the two corpus-based models

Study 1: Results of the forced choice task

The two analyses reveal that:

- the native speakers as a group perform worse than the two corpus-based models with the performance of the NDL-based model closer to native speaker performance than the performance of the logistic regression model
- there is a high and significant correlation between the proportions of selected constructions and the matching corpus based probability estimates - as the probability of the construction rises, so does the proportion of selections of that construction

Summary of Study 1

- A probabilistic model based on richly annotated corpus data is superior to an average native speaker.
- Human performance is susceptible to variation, while machine performance (at least regression) aims for mathematical precision.
- For the field of linguistics to move forward we need multivariate corpus research coupled with experimentation and the implementation of more intricate modelling techniques that are cognitively more plausible (e.g. NDL; cf. Milin et al. 2016)

What is the “gold” standard of human performance?

- In Study 1, the evaluation of native speaker performance is limited to a forced choice task.
- Other types of experimental data should ideally complement the results of the forced choice task.

Study 2: ratings vs acceptability judgements

Klavan, J. & A. 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.

Study 2: Original aims & research questions

- To evaluate the performance of a corpus-based mixed-effects logistic regression model by comparing the corpus-based predictions against the preferential choices and acceptability ratings of native speakers
- It is assumed that the predictions made by the corpus-based model are mirrored in the behaviour of native speakers
- RQ: *Are corpus-based predictions reflected equally well in native speakers' preferential choices and their ratings?*
  - ➢ *If not, where and why do they diverge?*

Study 2: Modified aim & research question

- To evaluate the performance of a corpus-based mixed-effects logistic regression model by comparing the corpus-based predictions against the preferential choices and acceptability ratings of native speakers
- RQ:
  - ➢ *What are the upper and lower boundaries of human classification behaviour?*
  - ➢ *Are the boundaries the same across the two types of off-line experiments?*

Study 2: Acceptability ratings

- The experimental items used in the acceptability rating task were the same as the 30 items used in the forced choice task
- For each of the original experimental item an alternative paraphrase was constructed
- The adessive and *peal* constructions were separated from the rest of the sentence by square brackets
- It was explicitly stated in the instructions that the focus of the study is on the alternation between the adessive and *peal*

Study 2: Procedure

- “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
- four versions of the two questionnaires to diminish potential order effects => 8 lists alltogether
- one sentence at a time, not possible to change their answers
- ~ 10 minutes to complete it

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

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

Study 2: Participants & Procedure

- 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)

Study 2: Results of the acceptability task

- Following Divjak et al. (2016) the raw acceptability ratings were residualised against participant and position of the experimental items in the experiment:
  - Ratings were regressed on participant and position
  - The residuals from this regression were then used in subsequent data analysis
  - The residualised ratings were rescaled so that each participant used the entire scale (1 - 10).

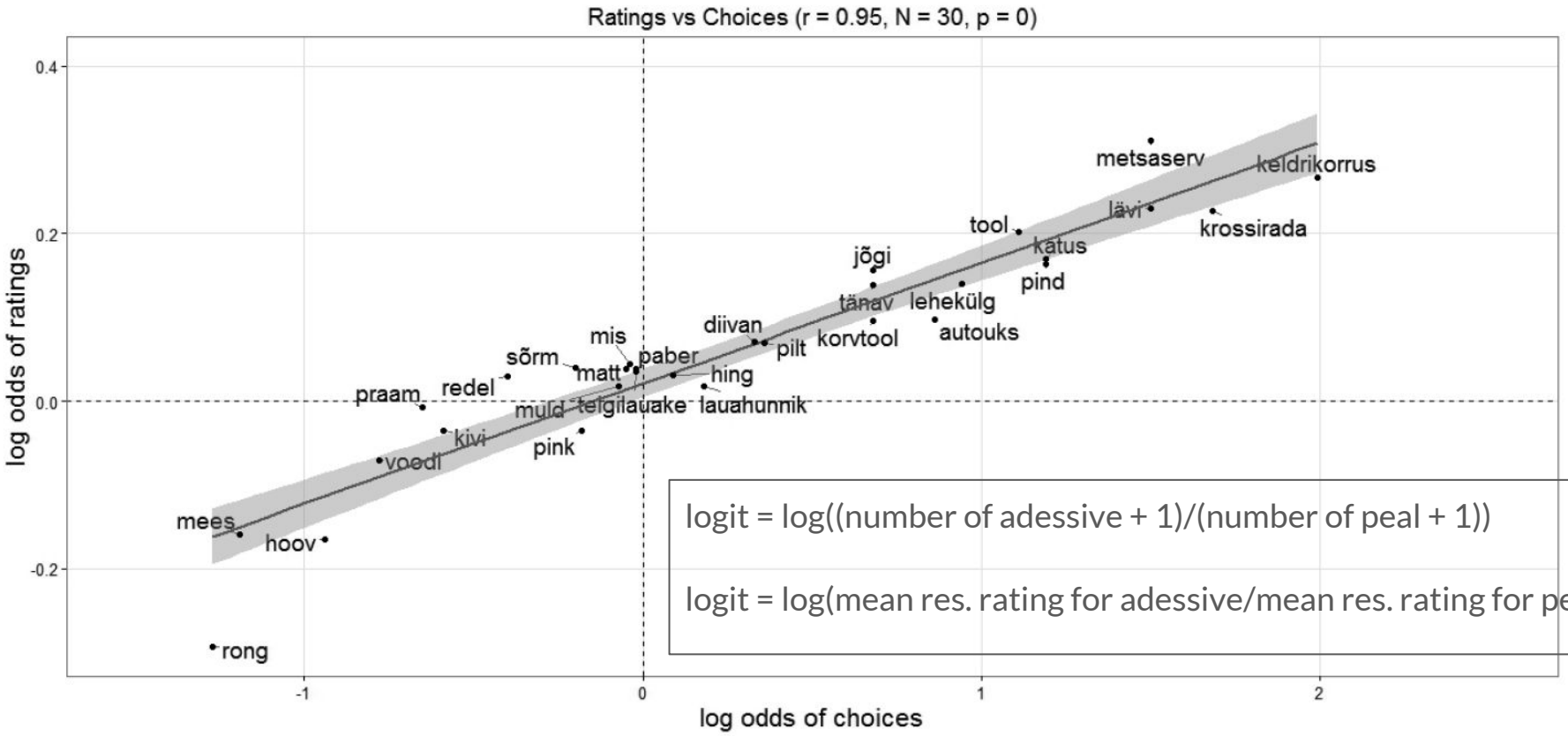
Study 2: Results of the acceptability task

- For the purposes of this presentation, the residualised mean ratings are taken to reflect the preferred “choices” of native speakers
- i.e. if the residualised mean rating of the adessive construction for itemi is higher than the residualised mean rating of the *peal* construction for the same item, the preferred “choice” is taken to be the adessive construction
- This crude approach allows us to compare native speaker “performance” across the two tasks and against corpus-based models

Table. Performance of the corpus-based models and native speakers in a forced choice task and in an acceptability rating task

Model	**Accuracy**
logistic regression	28/30 = 93%
NDL	27/30 = 90%
native speakers in a forced choice task	23/30 = 77%
native speakers in an acceptability task	21/30 = 70%

Agreement between choices and ratings



logit = log((number of adessive + 1)/(number of peal + 1))

logit = log(mean res. rating for adessive/mean res. rating for peal)

Summary of Study 2

- There is a strong positive correlation between the corpus-based probability estimates and the experimental data - as the probability of the *peal* construction rises, so does the mean residualised rating for the *peal* construction
- There is a strong correlation between choices and ratings
  - ➢ however, there are also clear instances where the two diverge
- The results of both forced choice task and the acceptability rating task suggest that the default choice for native speakers is the adessive construction; something that is not included in the corpus-based models at the moment

Interim conclusions: why combine corpus-linguistic and experimental methods?

Why do corpus-based studies?

- 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

Why do linguistic experiments?

(cf. Klavan & Divjak 2016, Divjak et al. 2016)

- without behavioural (or experimental) data it would be very difficult if not impossible to provide an adequate assessment of a corpus-based model
  - ➢ linguistic experiments are necessary to calibrate our corpus-based models - sometimes models are very accurate, and sometimes they appear to be accurate
- different types of experimental data give us access to different types of “behaviour” (important complementary information as to the nature of the linguistic phenomenon)

Avenues for further research

Doing linguistics during the “quantitative turn”

(2016 Special Issue in Cognitive Linguistics, edited by Divjak, Levshina & Klavan)

- It is the best of times:
  - Many new methods for data collection and data analysis (e.g. statistical modelling)
  - The age of digitalisation and “big data”
- It is the worst of times:
  - We lack sufficient understanding of what is the nature of the data obtained via the various methods
  - We lack sufficient understanding of what are the underlying assumptions about language made by various algorithms

Item	PUT1358 "The Making and Breaking of Models: Experimentally Validating Classification Models in Linguistics (1.01.2017–31.12.2020)", Jane Klavan, University of Tartu, Faculty of Arts and Humanities, College of Foreign Languages and Cultures.
Project number	PUT1358
Annotation in English	Recent years have witnessed an exponential growth in the use of statistical modelling techniques to analyse linguistic data. There are only few studies that pay close attention to the aspects of the language system captured by these models. The proposed project will use methodological pluralism to enhance our understanding of the mathematical properties underlying the statistical modelling techniques now commonly used in linguistics, impacting directly on how empirical data feeds into constructing cognitively realistic linguistic theories. The project focuses on the following questions: How well do different modelling techniques perform on the same linguistic data? How well do humans perform in comparison to machines? Which (linguistic) features are picked up by both machines and humans? The project proceeds from the assumption that in order to make statistical models, we need to break them by pitting them against each other and against human behaviour in experimental settings.

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.
- Baayen, R. Harald. 2011. Corpus linguistics and naive discriminative learning. *Revista Brasileira de Linguística Aplicada* 11(2). 295-328.
- Baayen, R. Harald, Anna Endresen, Laura A. Janda, Anastasia Makarova & Tore Nesset. 2013. Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics* 37. 253-291.
- 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.
- Divjak, Dagmar, Natalia Levshina, Jane Klavan. 2016. Cognitive Linguistics: Looking back, looking forward. Cognitive Linguistics, 27 (4), 1−17.
- Geeraerts, D. 1995. Cognitive Linguistics. In *The Handbook of Pragmatics.* John Benjamins Publishing Company.

- Hosmer Jr, David W., Stanley Lemeshow & Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. John Wiley & Sons.
- 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 & 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).
- Langacker, Ronald W. 2008. *Cognitive Grammar. A Basic Introduction*. Oxford: Oxford University Press.
- Milin, Petar, Dagmar Divjak, Strahinja Dimitrijević & R. Harald Baayen. 2016. Towards cognitively plausible data science in language research. *Cognitive Linguistics* 27(4). 507-526.
- Pinheiro, José C. & Bates, Douglas M. 2002. *Mixed-Effects Models in S and S-PLUS*. New York: Springer.
- Rescorla, Robert A. & Allan W. Wagner. 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In Abraham H. Black & William F. Prokasy (eds.), *Classical Conditioning II: Current Research and Theory*, 64–99. New York: Appleton Century Crofts.
- Theijssen, Daphne, Louis ten Bosch, Lou Boves, Bert Cranen & Hans van Halteren. 2013. Choosing alternatives: Using Bayesian Networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory* 9(2). 227–262.