A multivariate analysis of the alternation between the adessive case and postposition *peal* ‘on’ in Estonian dialectal data

Jane Klavan <jane.klavan@gmail.com>
University of Tartu, Estonia

Paper presented at the conference "Grammar of non-standard varieties in the East of the Circum-Baltic area" at the University of Tartu, Estonia, 2 February 2018

Estonian adessive case and postposition *peal* ‘on’

(1)	*suured*	*liha*	*kaosid*	*ollid*	***lava***	***pial*** (MUH)
	big-NOM	meat-GEN	bowl-NOM	be.PST	table-GEN	on

‘the big bowls of meat were **on the table**’

(2)	*nied*	*õlid*	*kaa*	***lauwal***	*sis* (LÜG)
	they-NOM	be.PST	also	table-ADE	then

‘these were also **on the table** then’

Estonian adessive case and postposition *peal* ‘on’

(1) *suured* *liha* *kaosid* *ollid* ***lava*** ***pial*** (MUH)

big-NOM meat-GEN bowl-NOM be.PST table-GEN on

‘the big bowls of meat were **on the table**’

(2) *nied* *õlid* *kaa* ***lauwal*** *sis* (LÜG)

they-NOM be.PST also table-ADE then

‘these were also **on the table** then’

TRAJECTOR

LANDMARK

The polysemous nature of ADE construction

(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.'

Overall research question and aims

CVL = corpus-based variationist linguistics

I am interested in:

- morophosyntactic alternation phenomenon, e.g. locative cases vs. postpositions (ade vs. *peal*)
  - what drives speakers' choice between morphosyntactic alternatives
  - what are the statistical methods available to model these choices, e.g. regression, memory-based learning, NDL, analogical modelling, the "tree & forest" method, Bayesian modelling

Overall research question and aims

I subscribe to:

- empirical, usage-based linguistics (using either corpus-based or experimental data)
- probabilistic view of grammar (cf. Bresnan 2007, Bresnan et al. 2007, Bresnan & Ford 2010, Tagliamonte & Baayen 2012, Divjak & Arppe 2013, Szmrecsanyi 2013)

DISCLAIMER: I am not a dialectologist, but...

The Corpus of Estonian Dialects (CED 2015)

http://www.murre.ut.ee/mkweb/

- Data collected in October 2014 -> the corpus consisted of 834,311 morphologically annotated tokens in total from 10 dialect areas
- Recordings mostly from 1960s - 1970s
- Traditional dialect interviews where informants talk about past events, customs, work and their everyday lives
- Long monologous passages
- “Traditional problems” (cf. Tagliamonte & Baayen 2012: 142-143):
  - not highly educated
  - non-mobile
  - elderly people

Overview of the dataset

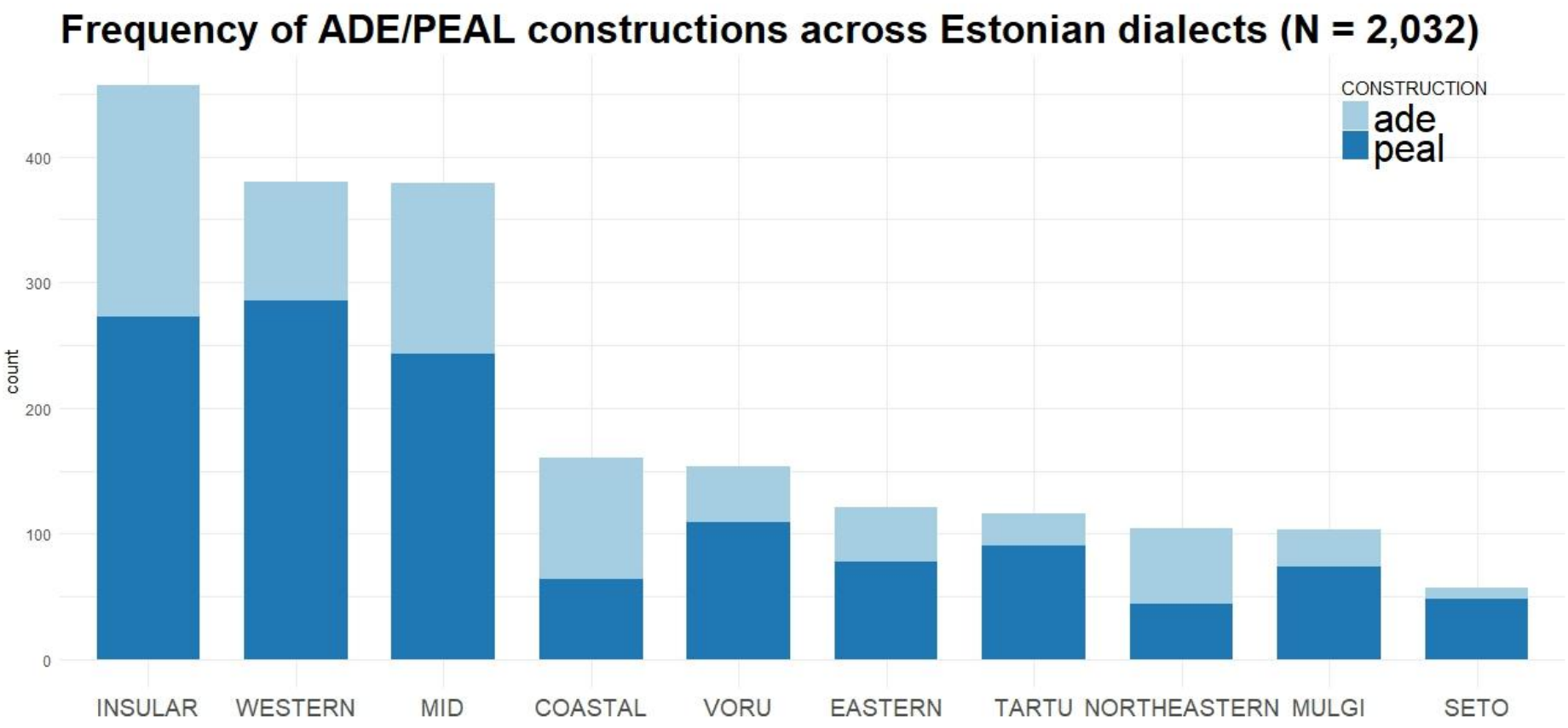
Jane Klavan, Maarja-Liisa Pilvik and Kristel Uiboaed[1]

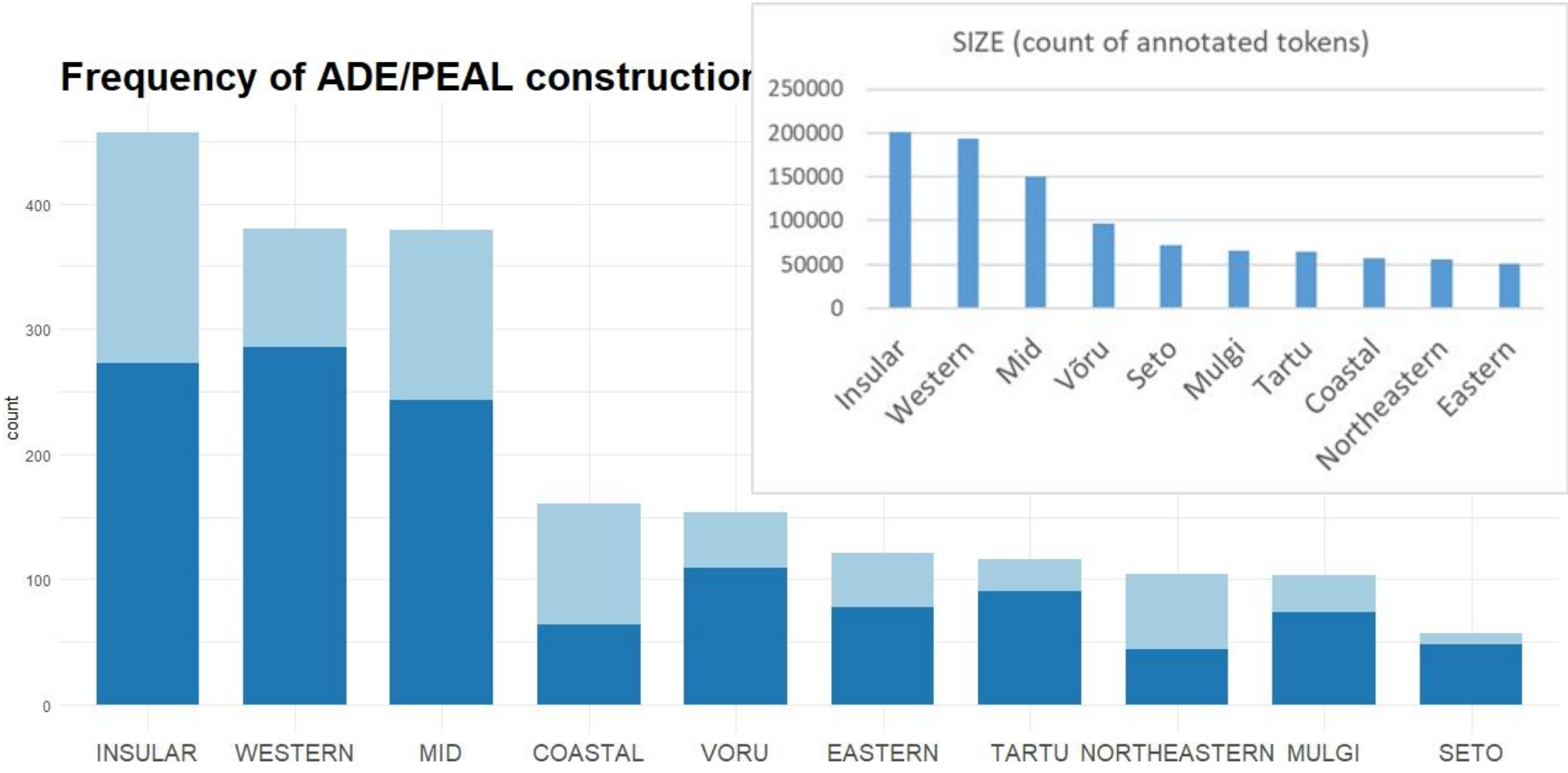
The Use of Multivariate Statistical Classification Models for Predicting Constructional Choice in Spoken, Non-Standard Varieties of Estonian[2]

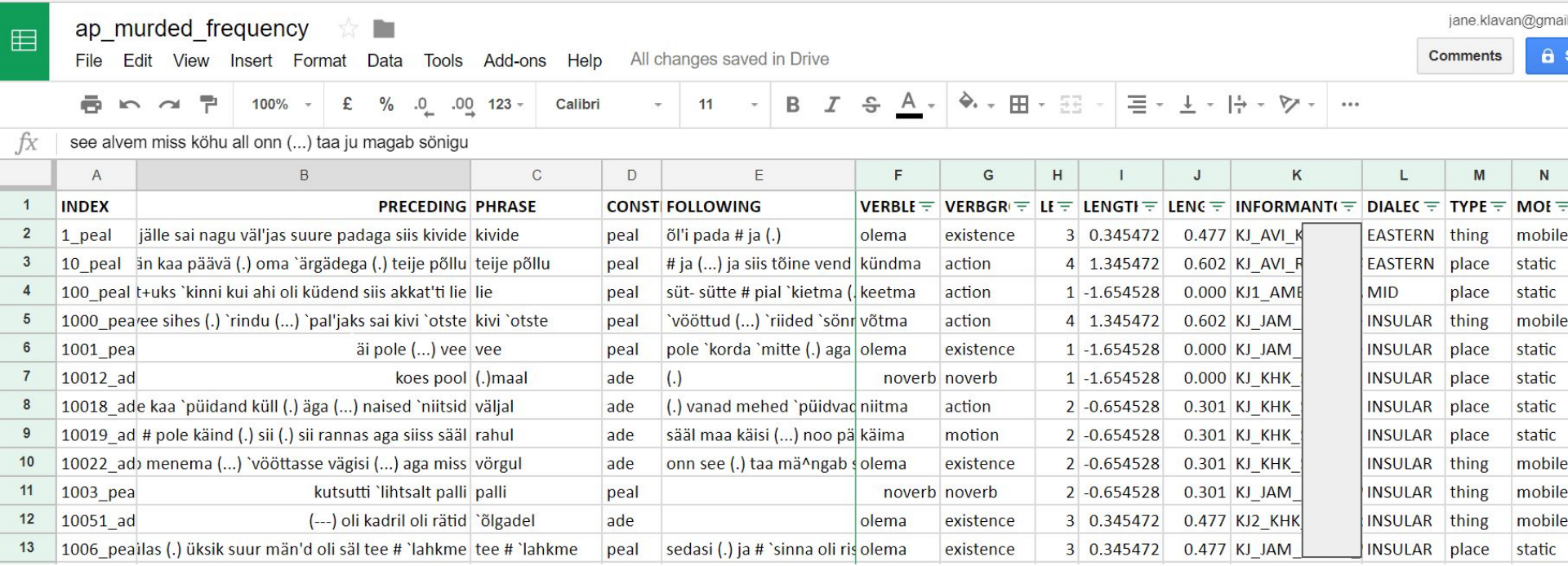
SKY Journal of Linguistics 28 (2015), 187–224

- Data collected and annotated
- in collaboration with **Maarja-Liisa Pilvik** and **Kristel Uiboaed**
- Data collected automatically with an R script
- We extracted all the instances of the adessive case and the adposition *peal* with the symmetrical context span of 10 words

	TOTAL	OUR DATASET
Adessive	14,710	**722**
Peal	1,586	**1,310**
TOTAL	16,296	**2,032**


Frequency of ADE/PEAL constructions across Estonian dialects (N = 2,032)
CONSTRUCTION
ade
peal
count
400
300
200
100
0
INSULAR
WESTERN
MID
COASTAL
VORU
EASTERN
TARTU
NORTHEASTERN
MULGI
SETO


Frequency of ADE/PEAL constructio
count
400
300
200
100
0
INSULAR
WESTERN
MID
COASTAL
VORU
EASTERN
TARTU
NORTHEASTERN
MULGI
SETO
SIZE (count of annotated tokens)
250000
200000
150000
100000
50000
0
Insular
Western
Mid
Võru
Seto
Mulgi
Tartu
Coastal
Northeastern
Eastern



ap_murded_frequency

jane.klavan@gmail

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

Comments

100% £ % .0 .00 123 Calibri 11

see alvem miss köhu all onn (...) taa ju magab sönigu

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	INDEX	PRECEDING	PHRASE	CONST	FOLLOWING	VERBLE	VERBGR	LE	LENGTI	LENG	INFORMANT	DIALEC	TYPE	MOI
2	1_peal	jälle sai nagu väl'jas suure padaga siis kivide	kivide	peal	õl'i pada # ja (.)	olema	existence	3	0.345472	0.477	KJ_AVI_K	EASTERN	thing	mobile
3	10_peal	än kaa päävä (.) oma `ärgädega (.) teije põllu	teije põllu	peal	# ja (...) ja siis tõine vend	kündma	action	4	1.345472	0.602	KJ_AVI_R	EASTERN	place	static
4	100_peal	t+uks `kinni kui ahi oli küdend siis akkat'ti lie	lie	peal	süt- sütte # pial `kietma (.	keetma	action	1	-1.654528	0.000	KJ1_AME	MID	place	static
5	1000_pea	ee sihes (.) `rindu (...) `pal'jaks sai kivi `otste	kivi `otste	peal	`vööttud (...) `riided `sönr	võtma	action	4	1.345472	0.602	KJ_JAM_	INSULAR	thing	mobile
6	1001_pea	äi pole (...) vee	vee	peal	pole `korda `mitte (.) aga	olema	existence	1	-1.654528	0.000	KJ_JAM_	INSULAR	place	static
7	10012_ad	koes pool	(.)maal	ade	(.)	noverb	noverb	1	-1.654528	0.000	KJ_KHK_	INSULAR	place	static
8	10018_ade	kaa `püidand küll (.) äga (...) naised `niitsid	väljal	ade	(.) vanad mehed `püidvac	niitma	action	2	-0.654528	0.301	KJ_KHK_	INSULAR	place	static
9	10019_ad	# pole käind (.) sii (.) sii rannas aga siiss sääl	rahul	ade	sääl maa käisi (...) noo pä	käima	motion	2	-0.654528	0.301	KJ_KHK_	INSULAR	place	static
10	10022_ad	menema (...) `vööttasse vägisi (...) aga miss	võrgul	ade	onn see (.) taa mä^ngab s	olema	existence	2	-0.654528	0.301	KJ_KHK_	INSULAR	thing	mobile
11	1003_pea	kutsutti `lihtsalt palli	palli	peal		noverb	noverb	2	-0.654528	0.301	KJ_JAM_	INSULAR	thing	mobile
12	10051_ad	(---) oli kadril oli rätid	`õlgadel	ade		olema	existence	3	0.345472	0.477	KJ2_KHK	INSULAR	thing	mobile
13	1006_pea	ilas (.) üksik suur män'd oli säl tee # `lahkme	tee # `lahkme	peal	sedasi (.) ja # `sinna oli ris	olema	existence	3	0.345472	0.477	KJ_JAM_	INSULAR	place	static

Table 1. Overview of the dataset (N = 2,032: adessive = 722, *peal* = 1,310) (Klavan, Pilvik & Uiboaed 2015: 198)

Predictor	**Levels**
MOBILITY (mobility of Landmark)	mobile, static
TYPE (type of Landmark)	place, thing
VERBLEMMA (lemma of the verb used together with the adessive or *peal construction*)	191 verb lemmas
VERBGROUP	activity, existence, motion, no verb, posture
LENGTH (length of the Landmark phrase in syllables)	ranging from 1 to 9 syllables
COMPLEXITY (morphological complexity of the word used in the adessive or *peal construction*)	compound, simple
TRWORDCLASS (word class of the Trajector phrase)	NP, pronoun, VP
POSITION (relative position between the Trajector and Landmark phrase)	lm_tr, tr_lm
DIALECT (the dialect area)	Coastal, Eastern, Insular, Mid, Mulgi, Northeastern, Seto, Tartu, Võru, Western
INFORMANT	286 informants
LEMMA (lemma of the word used in the adessive or *peal construction*)	420 lemmas

Klavan, Pilvik & Uiboaed (2015)

- In order to test the relevance of the different predictors, we applied several methods for statistical analysis:
  - Mixed-effects logistic regression (Pinheiro & Bates 2002)
  - Classification trees & random forests (Breiman 2001, Strobl et al. 2009)

Formula for mixed-effects logistic regression:

CONSTRUCTION ~ LENGTHLOG + COMPLEXITY + TYPE + VERBGROUP + DIALECT + (1|LEMMA) + (1|INFORMANT)

[11] adePealinf.cforest = cforest(CX ~ VERBGROUP + LENGTHLOG + DIALECT + TYPE + COMPLEXITY + POSITION + TRWORDCLASS, data=dat, controls=cforest_unbiased(ntree=4000, mtry=3))

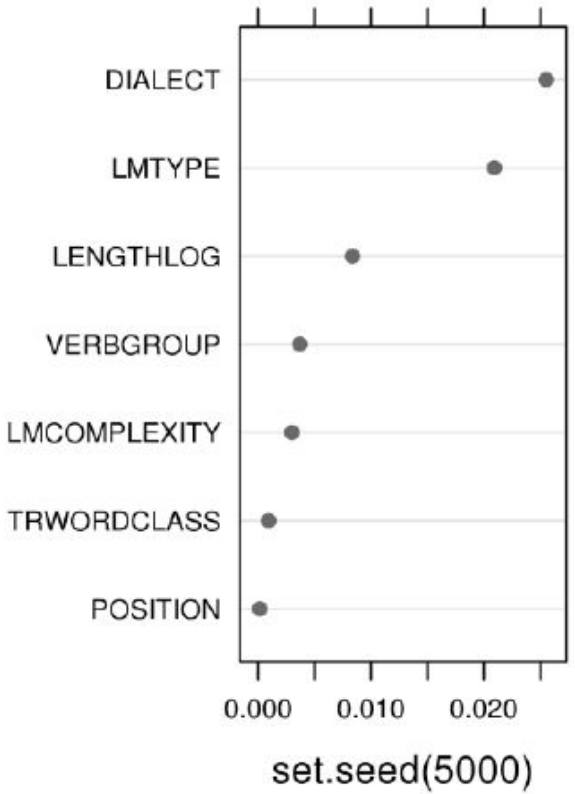
C = 0.94 Accuracy = 87% Baseline = 64%

Table 3. Model comparison statistics for the Estonian dialect dataset

	logLik	Chisq	Chi.Df	*p*-value	Reduction in AIC
INFORMANT	-1277.8				86.6
LEMMA	-1044.2	467.34	1	0.000	465.4
LENGTHLOG	-1041.3	5.7657	1	0.0163	3.7
COMPLEXITY	-1030.2	22.113	1	0.000	20.2
TYPE	-1020.4	19.689	1	0.000	17.6
VERBGROUP	-1001	38.842	4	0.000	30.9
DIALECT	-979.45	43.025	9	0.000	25

(Klavan, Pilvik & Uiboaed 2015: 203)

Figure 5. Variable importance in random forest analysis (Klavan, Pilvik & Uiboaed 2015: 212)



C = 0.83
Accuracy = 76%
Baseline = 64%

To re-cap: Klavan, Pilvik & Uiboaed (2015)

- We showed that the variation between ADE/PEAL in non-standard, spoken language is not free -> the choice depends on:
  - Specific lemmas
  - Length, complexity and type of Landmark
  - Verb
  - Dialect
- But, there are important factors missing from the models, e.g. frequency & persistence
- Going on a fishing trip...

FREQUENCY

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)

General predictions & assumptions

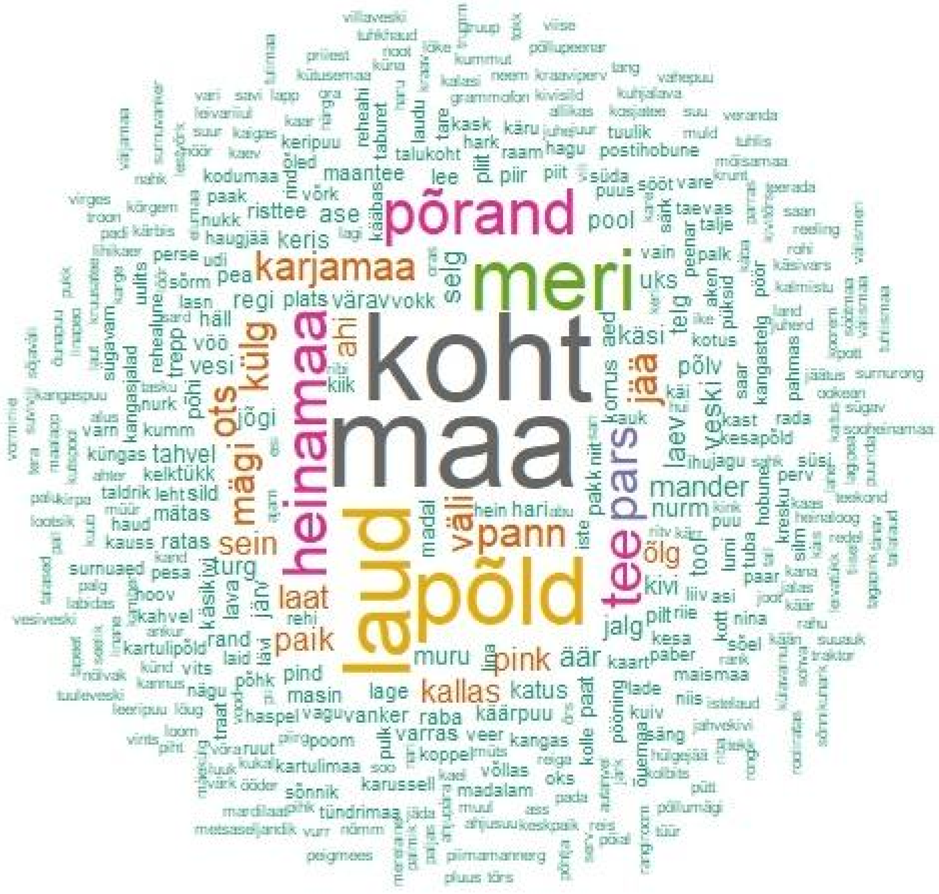
- Language users know the relative frequencies with which certain nouns appear with different locative cases and postpositions
- Assumption: such information is acquired through experience with input that exhibits distributional properties (Ellis 2002: 144)
- “The effects of frequency in input are modulated by the need to simultaneously satisfy the constraints of all other constructions that are represented in the learner’s system.” (Ellis 2002: 145)

C = 0.94 Accuracy = 87% Baseline = 64%

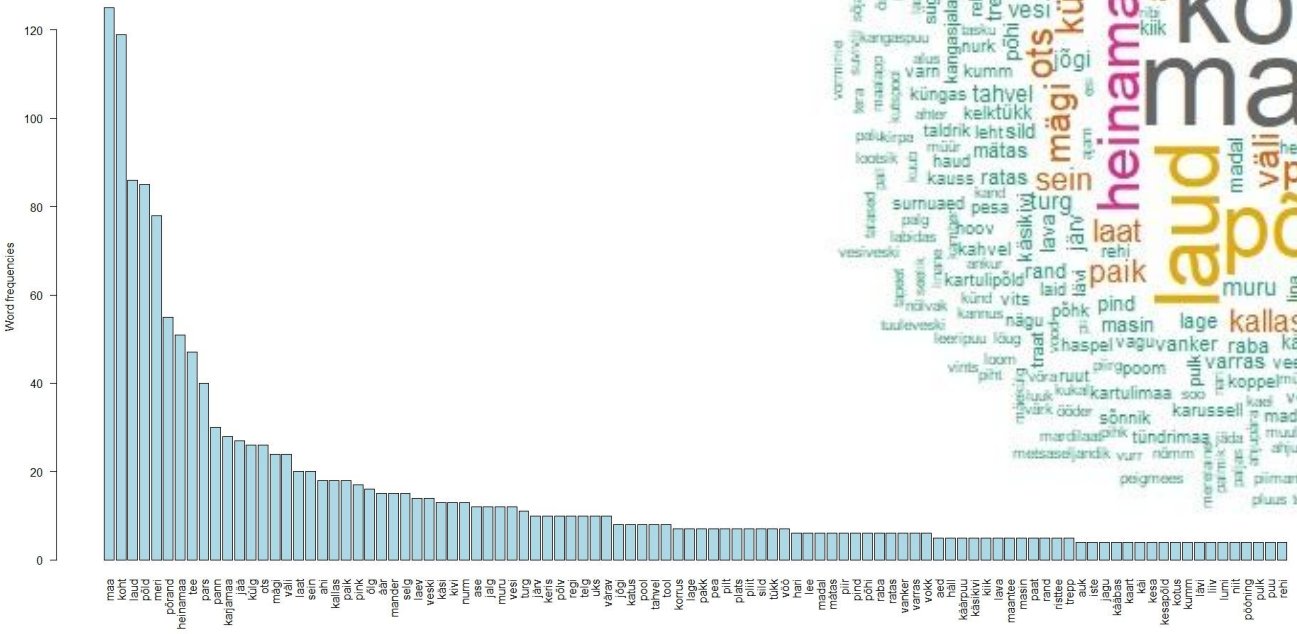
Table 3. Model comparison statistics for the Estonian dialect dataset

	logLik	Chisq	Chi.Df	*p*-value	Reduction in AIC
INFORMANT	-1277.8				86.6
LEMMA	-1044.2	467.34	1	0.000	465.4
LENGTHLOG	-1041.3	5.7657	1	0.0163	3.7
COMPLEXITY	-1030.2	22.113	1	0.000	20.2
TYPE	-1020.4	19.689	1	0.000	17.6
VERBGROUP	-1001	38.842	4	0.000	30.9
DIALECT	-979.45	43.025	9	0.000	25

Wordcloud of the 419 lemmas used together with the adessive or *peal* construction in the dataset



Frequency of the 419 lemmas used together with the adessive or *peal* construction in the dataset



Which frequency measures? 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: 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: Schematic co-occurrence table of token frequencies for association measures

Observed frequencies	Element *y* = laud ‘table’	Other elements	Totals
Element x = ade	*a*	*b*	*a* + *b*
Other elements	*c*	*d*	*c* + *d*
Totals	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d* = *N*

Table: Schematic co-occurrence table of token frequencies for association measures

Observed frequencies	Element *y* = laud ‘table’	Other elements	Totals
Element x = ade	*a* = 16	*b*	*a* + *b*
Other elements	*c*	*d*	*c* + *d*
Totals	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d* = *N*

Table: Schematic co-occurrence table of token frequencies for association measures

Observed frequencies	Element *y* = laud ‘table’	Other elements	Totals
Element x = ade	*a* = 16	*b* = 706	*a* + *b*
Other elements	*c*	*d*	*c* + *d*
Totals	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d* = *N*

Table: Schematic co-occurrence table of token frequencies for association measures

Observed frequencies	Element *y* = laud ‘table’	Other elements	Totals
Element x = ade	*a* = 16	*b* = 706	*a* + *b*
Other elements	*c* = 845	*d*	*c* + *d*
Totals	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d* = *N*

Table: Schematic co-occurrence table of token frequencies for association measures

Observed frequencies	Element *y* = laud ‘table’	Other elements	Totals
Element x = ade	*a* = 16	*b* = 706	*a* + *b*
Other elements	*c* = 845	*d* = 832,744	*c* + *d*
Totals	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d* = *N*

N = corpus size (total number of annotated tokens = 834,311)

Frequency measures included in the study

- -log10$p_{\text{Fisher-Yates exact test}}$
- Delta P

All computations were done with Gries' R script for coll.analysis 3.2

Other measures to be calculated in further work:

- Surprisal
- Entropy

The Fisher-Yates exact test (Ellis & Ferreira-Junior 2009)

All computations were done with Gries' R script for coll.analysis 3.2:

- the script uses an exact binomial test (the Fisher Yates exact test) to quantify the association strength between the noun and the ade/peal construction they occur in
- it provides a p-value for each noun with a construction
- it log transforms the p-value (to the base of 10) such that highly positive and highly negative values indicate a large degree of attraction and repulsion respectively, while 0 indicates random co-occurrence
- an (absolute) p_{log} value that is equal to or higher than 1.3 corresponds to a probability error of 5% or less

The Fisher-Yates exact test (Gries 2007)

Why this measure is good?

- It is preferred to the more common χ^2 because it does not violate distributional assumptions

Why this measure is not so good?

- It is a measure of the two-way dependency between a pair of events, but associations are not necessarily reciprocal in strength - we need additional measures to assess separately the directional relations

ΔP (Allan 1980; Ellis & Ferreira-Junior 2009)

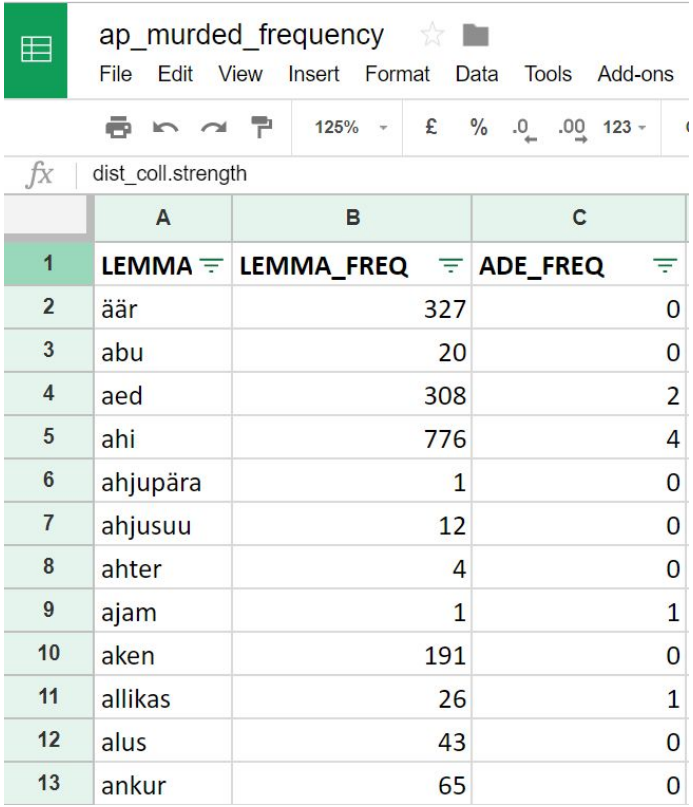
a. $\Delta P_{y|x} = \frac{a}{a+b} - \frac{c}{c+d}$

b. $\Delta P_{x|y} = \frac{a}{a+c} - \frac{b}{b+d}$

- ΔP is the probability of the outcome given the cue P(O|C) minus the probability of the outcome in the absence of the cue P(O|-C)
- when these are the same, when the outcome is just as likely when the cue is present as when it is not, there is no covariation between the two events and $\Delta P = 0$
- ΔP approaches 1.0 as the presence of the cue increases the likelihood of the outcome and approaches -1.0 as the cue decreases the chance of the outcome

=> It can thus be used as a measure of the degree to which a particular noun is distinctive in signaling either ADE or PEAL cx, or, in turn, the degree to which a ADE/PEAL cx selects a particular noun

The input data



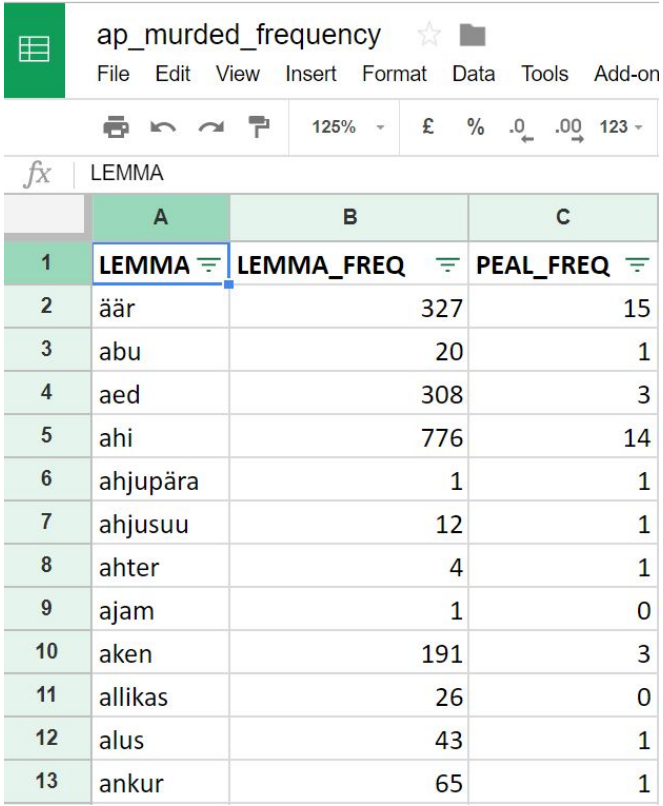
ap_murded_frequency

File Edit View Insert Format Data Tools Add-ons

125% £ % .0 .00 123

fx dist_coll.strength

	A	B	C
1	LEMMA	LEMMA_FREQ	ADE_FREQ
2	äär	327	0
3	abu	20	0
4	aed	308	2
5	ahi	776	4
6	ahjupära	1	0
7	ahjusuu	12	0
8	ahter	4	0
9	ajam	1	1
10	aken	191	0
11	allikas	26	1
12	alus	43	0
13	ankur	65	0



ap_murded_frequency

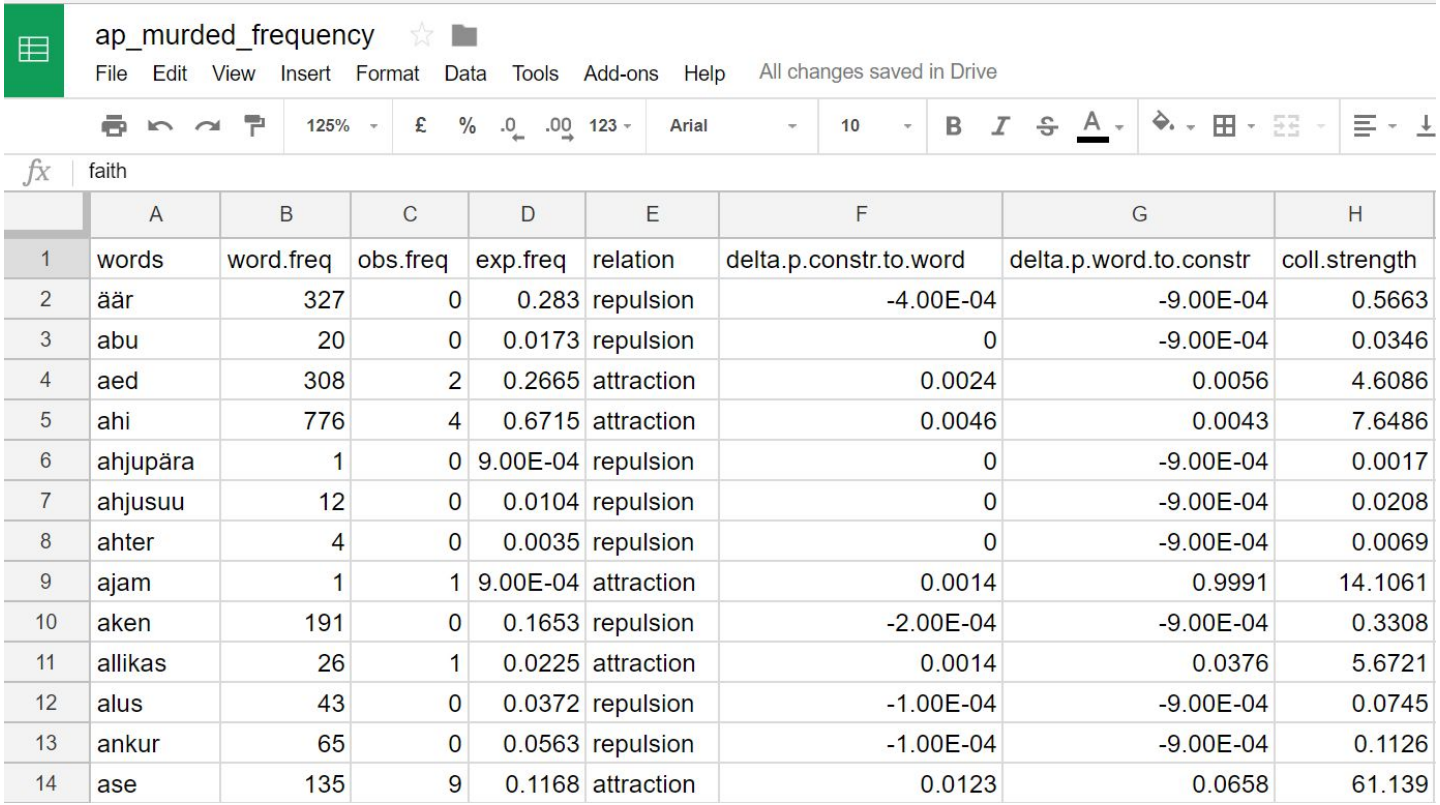
File Edit View Insert Format Data Tools Add-on

125% £ % .0 .00 123

fx LEMMA

	A	B	C
1	LEMMA	LEMMA_FREQ	PEAL_FREQ
2	äär	327	15
3	abu	20	1
4	aed	308	3
5	ahi	776	14
6	ahjupära	1	1
7	ahjusuu	12	1
8	ahter	4	1
9	ajam	1	0
10	aken	191	3
11	allikas	26	0
12	alus	43	1
13	ankur	65	1

The output data: adessive


ap_murded_frequency

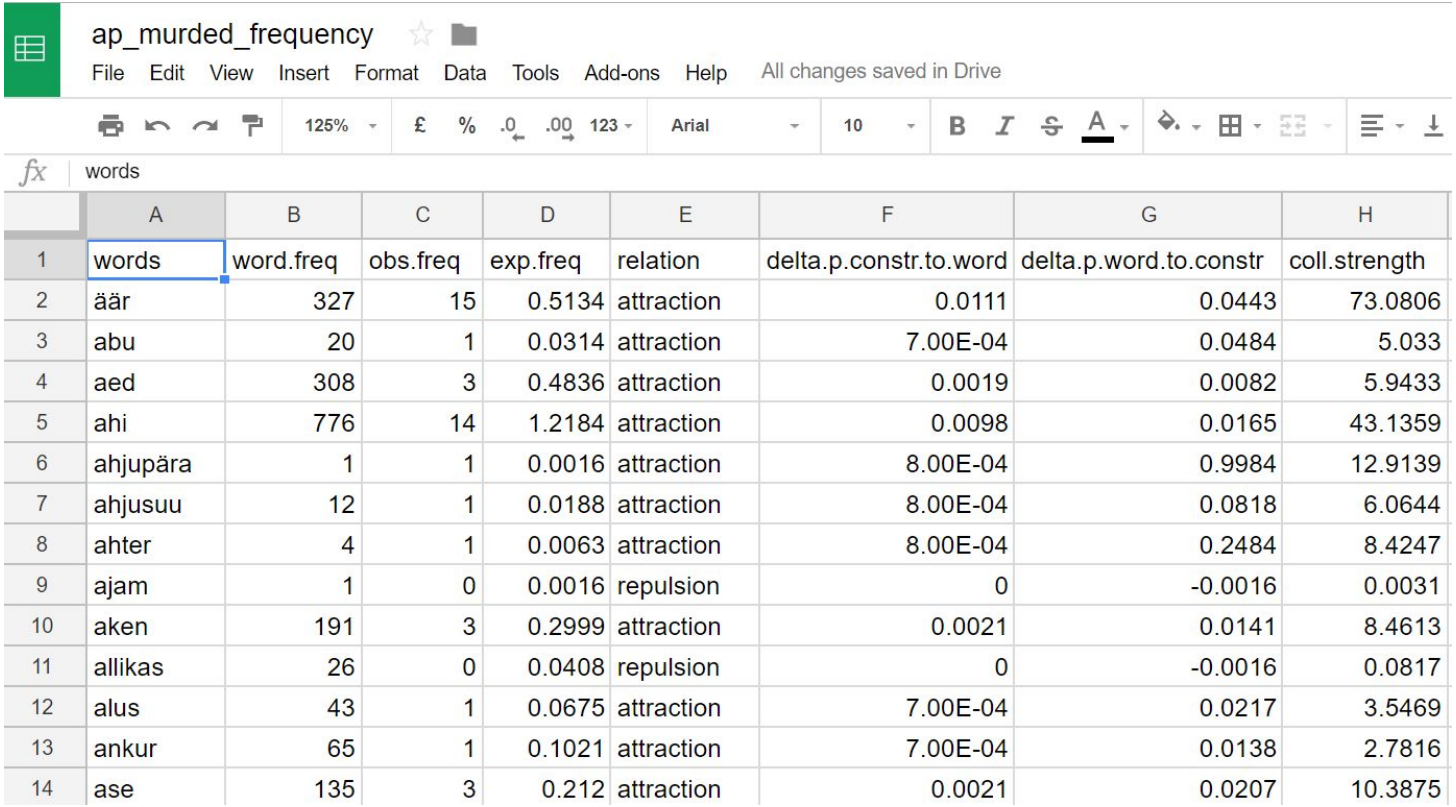
File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

125% £ % .0 .00 123 Arial 10

faith

	A	B	C	D	E	F	G	H
1	words	word.freq	obs.freq	exp.freq	relation	delta.p.constr.to.word	delta.p.word.to.constr	coll.strength
2	äär	327	0	0.283	repulsion	-4.00E-04	-9.00E-04	0.5663
3	abu	20	0	0.0173	repulsion	0	-9.00E-04	0.0346
4	aed	308	2	0.2665	attraction	0.0024	0.0056	4.6086
5	ahi	776	4	0.6715	attraction	0.0046	0.0043	7.6486
6	ahjupära	1	0	9.00E-04	repulsion	0	-9.00E-04	0.0017
7	ahjusuu	12	0	0.0104	repulsion	0	-9.00E-04	0.0208
8	ahter	4	0	0.0035	repulsion	0	-9.00E-04	0.0069
9	ajam	1	1	9.00E-04	attraction	0.0014	0.9991	14.1061
10	aken	191	0	0.1653	repulsion	-2.00E-04	-9.00E-04	0.3308
11	allikas	26	1	0.0225	attraction	0.0014	0.0376	5.6721
12	alus	43	0	0.0372	repulsion	-1.00E-04	-9.00E-04	0.0745
13	ankur	65	0	0.0563	repulsion	-1.00E-04	-9.00E-04	0.1126
14	ase	135	9	0.1168	attraction	0.0123	0.0658	61.139

The output data: *peal*



ap_murded_frequency

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

125% £ % .0 .00 123 Arial 10

fx words

	A	B	C	D	E	F	G	H
1	words	word.freq	obs.freq	exp.freq	relation	delta.p.constr.to.word	delta.p.word.to.constr	coll.strength
2	äär	327	15	0.5134	attraction	0.0111	0.0443	73.0806
3	abu	20	1	0.0314	attraction	7.00E-04	0.0484	5.033
4	aed	308	3	0.4836	attraction	0.0019	0.0082	5.9433
5	ahi	776	14	1.2184	attraction	0.0098	0.0165	43.1359
6	ahjupära	1	1	0.0016	attraction	8.00E-04	0.9984	12.9139
7	ahjusuu	12	1	0.0188	attraction	8.00E-04	0.0818	6.0644
8	ahter	4	1	0.0063	attraction	8.00E-04	0.2484	8.4247
9	ajam	1	0	0.0016	repulsion	0	-0.0016	0.0031
10	aken	191	3	0.2999	attraction	0.0021	0.0141	8.4613
11	allikas	26	0	0.0408	repulsion	0	-0.0016	0.0817
12	alus	43	1	0.0675	attraction	7.00E-04	0.0217	3.5469
13	ankur	65	1	0.1021	attraction	7.00E-04	0.0138	2.7816
14	ase	135	3	0.212	attraction	0.0021	0.0207	10.3875

Top 10 lemma types for the adessive construction

	Frequency in ADE - CX
meri 'sea'	60
maa 'land'	59
heinamaa 'hayfield'	43
põld 'field'	42
põrand 'floor'	41
koht 'place'	36
pars 'bar'	28
karjamaa 'pasture'	27
laat 'fair'	18
paik 'place'	16

Top 10 lemma types for the adessive construction

	Frequency in ADE - CX		Collocation Strength Fisher Yates Exact p log
meri 'sea'	60	meri 'sea'	93.2
maa 'land'	59	põrand 'floor'	82.55
heinamaa 'hayfield'	43	heinamaa	81.95
põld 'field'	42	maa 'land'	70.94
põrand 'floor'	41	põld 'field'	65.26
koht 'place'	36	karjamaa	54.49
pars 'bar'	28	pars 'bar'	48.89
karjamaa 'pasture'	27	mander 'mainland'	41.82
laat 'fair'	18	koht 'place'	40.57
paik 'place'	16	laat 'fair'	33.18

- Gries' script uses an exact binomial test to quantify the association between the words and the adessive/peal construction they occur in.
- It provides a p-value for each word with each ade/peal construction and log transforms it.
- An (absolute) plog value that is equal to or higher than 1.3 corresponds to a probability of error of 5% or less.

Top 10 lemma types for the adessive construction

	Frequency in ADE - CX		Collocation Strength Fisher Yates Exact p log		Delta P Constructio n -> Word
meri 'sea'	60	meri 'sea'	93.2	meri 'sea'	0.08
maa 'land'	59	põrand 'floor'	82.55	maa 'land'	0.08
heinamaa 'hayfield'	43	heinamaa	81.95	põrand 'floor'	0.06
põld 'field'	42	maa 'land'	70.94	heinamaa	0.06
põrand 'floor'	41	põld 'field'	65.26	põld 'field'	0.06
koht 'place'	36	karjamaa	54.49	koht 'place'	0.05
pars 'bar'	28	pars 'bar'	48.89	karjamaa	0.04
karjamaa 'pasture'	27	mander 'mainland'	41.82	pars 'bar'	0.04
laat 'fair'	18	koht 'place'	40.57	mander	0.02
paik 'place'	16	laat 'fair'	33.18	laat 'fair'	0.02

Delta P is the outcome given the cue minus the probability of the outcome in the absence of the cue.

It signals the degree to which a construction selects a particular type in that slot.

Top 10 lemma types for the adessive construction

	Frequency in ADE - CX		Collocation Strength Fisher Yates Exact p log		Delta P Constructio n -> Word		Delta P Word -> Construc tion
meri 'sea'	60	meri 'sea'	93.2	meri 'sea'	0.08	kresku	1.000
maa 'land'	59	põrand 'floor'	82.55	maa 'land'	0.08	laudu	1.000
heinamaa 'hayfield'	43	heinamaa	81.95	põrand 'floor'	0.06	ajam	1.000
põld 'field'	42	maa 'land'	70.94	heinamaa	0.06	grammofon	1.000
põrand 'floor'	41	põld 'field'	65.26	põld 'field'	0.06	heinaloog	1.000
koht 'place'	36	karjamaa	54.49	koht 'place'	0.05	hülgejää	1.000
pars 'bar'	28	pars 'bar'	48.89	karjamaa	0.04	kõrgem	1.000
karjamaa 'pasture'	27	mander 'mainland'	41.82	pars 'bar'	0.04	kosjatee	1.000
laat 'fair'	18	koht 'place'	40.57	mander	0.02	külavainu	1.000
paik 'place'	16	laat 'fair'	33.18	laat 'fair'	0.02	leeripuu	1.000

Top 10 lemma types for the *peal* construction

	Frequency in PEAL construction		Collocation Strength Fisher Yates Exact p log		Delta P Construction -> Word		Delta P Word -> Constructio n
koht 'place'	83	koht 'place'	101.48	koht 'place'	0.06	korrus	1.00
laud 'table'	70	laud 'table'	93.48	laud 'table'	0.05	lage	1.00
maa 'land'	66	maa 'land'	65.4	maa 'land'	0.05	kangastelg	1.00
põld 'field'	43	põld 'field'	56.03	põld 'field'	0.03	madalam	1.00
tee 'road'	42	tee 'road'	51.08	tee 'road'	0.03	õlg_õled	1.00
pann 'pan'	27	pann 'pan'	47.44	pann 'pan'	0.02	õuemaa	1.00
ots 'tip'	24	külg 'side'	32.9	külg 'side'	0.02	sügavam	1.00
külg 'side'	23	sein 'wall'	25.69	ots 'tip'	0.02	ahjupära	1.00
sein 'wall'	19	muru 'grass'	24.59	sein 'wall'	0.01	elumaa	1.00
meri 'sea'	18	pink 'bench'	22.23	muru 'grass'	0.01	kangaspuu	1.00

Distinctive collexeme analysis

Words distinctive fore ADE	freq._ade	freq._peal	coll.strength
meri 'sea'	60	18	13.49
heinamaa 'heyfield'	43	8	12.29
karjamaa 'pasture'	27	1	11.00
põrand 'floor'	41	14	8.58
mander 'mainland'	15	0	6.78
laat 'fair'	18	2	6.21
paik 'place'	16	2	5.39
pars 'bar'	28	12	5.10
vöö 'belt'	7	0	3.15
mägi 'mountain'	16	8	2.75

Words distinctive for PEAL	freq._ade	freq._peal	coll.strength
tee 'road'	5	42	4.03
laud 'table'	16	70	3.44
ots 'tip'	2	24	2.92
äär 'edge'	0	15	2.87
pann 'pan'	3	27	2.82
sein 'wall'	1	19	2.75
õu 'yard'	0	14	2.68
nurm 'meadow'	0	13	2.48
muru 'grass'	0	12	2.29
vesi 'water'	0	12	2.29

Results: 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)
- adessie = 722 tokens (163 types)
- *peal* = 1310 tokens (339 types)
- cf. present-day written language (Klavan 2012):
  - adessive = 450 tokens (255 types)
  - *peal* = 450 tokens (209 types)

Problematic issues when counting frequencies

- How to count the type/token frequency of compounds?
  - *postihobusel* ‘on a post-horse’ vs *hobusel* ‘on a horse’
  - *sooheinamaal* ‘on a swamp hayfield’ = *soo* + *heinamaal* = *soo* + *heina* + *maa*
- What to count as the adessive construction? Adessive in the locative function vs other functions

Observed frequencies	Element *y* = laud ‘table’	Other elements	Totals
Element x = ade	*a* = 16	*b* = 706	*a* + *b* = 722
Other elements	*c* = 845	*d* = 832,744	*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.'

Problematic issues when counting frequencies

- How to count the type/token frequency of compounds?
  - *postihobusel* ‘on a post-horse’ vs *hobusel* ‘on a horse’
  - *sooheinamaal* ‘on a swamp hayfield’ = *soo* + *heinamaal* = *soo* + *heina* + *maa*
- What to count as the adessive construction? Adessive in the locative function vs other functions

14,710 = the total number of all of the adessive tokens in the corpus

Observed frequencies	Element *y* = laud ‘table’	Other element	
Element x = ade	*a* = 16	*b* = 706	*a* + *b* = **722**
Other elements	*c* = 845	*d* = 832,744	*c* + *d*
Totals	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d* = *N*

Conclusions

- Predictors that play a role in the alternation between adessive and *peal* construction 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
- Strong associations between specific words and constructions

But...

- What is the psychological plausibility of the fitted models and the different association measures?
  - in need of controlled experimental data to zoom in on the different variables and the role they play
- What is the range of alternating constructions?
  - accounting for the polysemy of alternating constructions, e.g. adessive in the locative function vs adessive in the other functions
  - the choice is not necessarily binary
- Are the same predictors “responsible” for the alternation between other locative cases and adpositions?
  - *allative~peale; ablative~pealt;* interior cases ~ *sisse/sees/seest*

THANK	YOU!
thank-PRS.1SG	you-2SG

aitäh!

References

Allan, L. G. 1980. A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, 15, 147-149.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston and Wolfgang Sternefeld (eds). Roots: Linguistics in Search of Its Evidential Base, 77–96. Berlin: Mouton de Gruyter.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina and R. Harald Baayen. 2007. Predicting the Dative Alternation. In Gerlof Bouma, Irene Krämer, and Joost Zwarts (eds). Cognitive Foundations of Interpretation, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan and Marilyn Ford. 2010. Predicting syntax: processing dative constructions in American and Australian varieties of English. Language, 86: 1, 186–213.

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.

Ellis, Nick C., and Fernando Ferreira-Junior. "Constructions and their acquisition: Islands and the distinctiveness of their occupancy." *Annual Review of Cognitive Linguistics* 7.1 (2009): 188-221.

Erelt, Mati, Tiiu Erelt and Kristiina Ross. 2007. Eesti keele käsiraamat [Handbook of Estonian]. Tallinn: Eesti Keele Sihtasutus.

Gries, Stefan Th. 2007. Coll.analysis 3.2a. A program for R for Windows 2.x.

Gries, Stefan Th, and Nick C. Ellis. "Statistical Measures for Usage‐Based Linguistics." *Language Learning* 65.S1 (2015): 228-255.

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, 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.

Szmrecsanyi, Benedikt. 2013. Diachronic Probabilistic Grammar. English Language and Linguistics 1 (3): 41–68.

Tagliamonte, Sali A. & Baayen, Harald R. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. Language Variation and Change 24 (02): 135–178.

“Of course, frequency and dimensional central tendency are not the only factors that determine activation of candidate schemata;

there are moderating effects of recency of use and of context”

(Ellis 2002: 147)

PERSISTENCE

Definition of “persistence” (Szmrecsanyi 2006: 2)

the tendency that if speaker A faces a variable Z where he or she has the choice between two or more semantically equivalent variants (regardless of whether they are lexical, morphological, or syntactic in nature), speaker A’s choice will be affected by

> (*α*) previous exposure to the variable Z, such that use of a specific variant (either by speaker A or by another speaker B, to whose output speaker A has been exposed) in previous discourse will make it more likely, all other things being equal, that the same variant will be used again by speaker A (henceforth: *α*-persistence); or by
>
> (*β*) previous exposure to a linguistic pattern Z*, which is not variable in the same way as variable Z but parallel to one of variable Z’s variants, such that the use of the linguistic pattern Z* (either by speaker A herself or by another speaker B, to whose output speaker A has been exposed) in previous discourse will make it more likely, all other things being equal, that the variant of variable Z which is parallel to the linguistic pattern Z* will be used by speaker A (henceforth: *β*-persistence).

Example: MKT0020 (MID dialect), May 1965

LN: *kui ma àkkasin òrjama | ja piàle seľle pel̨è mul̀ ühte kèrɢet̀ tuňᴅiɢi ol̀ᴅ‖ aʙi˛elù aèɢ sie ol'i kā mièsol'i äᴅàlen̨e ja ài̇ɢe‿ja* ‖ ((köhib)) *vi̇̂s las̀t olèn kā ülès̀se kaš̀vat̆tɒnᴅ ja meĭ̮sa pel̀ᴅuᴅel | kaŭwaks meìl sīn | seᴅa oňni |* ((rögisedes)) *kaheks̆ateis̀t kümne ā̀stast sāᴅik̀ on̄ sē ol̀ᴅ* ((köhib)) ‖ *käĭsin meì̮sɒs tüȍl lei̮kkasin lei̮kkust | karttul̨ıᴅ veì̮tmas | poi- po̮èɢ ol'i ŭheksa ā̀stane sie ol'i vaò vaht̀is‿ja veì̮t'sin kaht̀e vaɢù | seᴅàsi oǹ minùl pā̈vaᴅ ol̀ᴅ sīn meì̮sas sin | tüȍᴅ tehà oh̀tu meì̮sas |* ((köhib)) *| lei̮kkasin lei̮kkust‿jɒ ‖ mièsol'i kūwe pā̈va teò̮mes ‖ jā | seľlest saim̆e mōna‿ja ‖ seľlest me elàsime | (--)*

Example: MKT0020 (MID dialect), May 1965

LN: *kui ma akkasin oṙjama | ja piàle seľle pelè mul̀ ühte keṙgeṫ tuńᴅɪɢɪ ol̀ᴅ‖ aʙɪ˛elù aèɢ sie ol'i kā mièѕ ol'i äᴅàlene ja aìɢe‿ja ‖* ((köhib)) *vîs laṡt olèn kā ülesse kašvaťtɒɴᴅ ja meĭsa pel̀ᴅuᴅel | kaŭwaks meil sīn | seᴅa oṅni |* ((rögisedes)) *kaheǩsateiṡt küm̀ne ắṡtast sāᴅik ōn sē ol̀ᴅ* ((köhib)) *‖ käĭsin meìsɒs tüöl leikkasin leikkust | karttuḷɪᴅ vetmas | poi- poèɢ ol'i üheksa ắṡtane sie ol'i vaò vahtis‿ja veťsin kahte vaɢù | seᴅàsi on̄ minùl pä̃vaᴅ ol̀ᴅ sīn meìsas sin | tüöᴅ tehà ohtu meìsas |* ((köhib)) *| leikkasin leikkust‿jɒ ‖ mièѕ ol'i kūwe pä̃va teomes ‖ jā | seľlest saìme mōna‿ja ‖ seľlest me elàsime | (--)*