Polytomous logistic regression analysis and modeling of linguistic alternations

Antti Arppe

General Linguistics, Department of Modern Languages

University of Helsinki

Concepts – linguistic alternations

- Alternative linguistic ***forms*** which denote roughly the same ***meaning***
  - ***Structural/constructional*** alternations
    - E.g. Finnish/German word order, English dative (Bresnan 2007) or possessive alternations (Gries 2003)
      - *He gave her the book* vs. *He gave the book to her*
      - *The book's title* vs. *the title of the book*
  - ***Lexical*** alternations
    - E.g. (near-)synonymy, social/dialectal variation
      - *Strong* vs. *powerful* (Church et al. 1991)
      - *Small* vs. *wee*

Theoretical assumptions & methodological prerequisites

- ***Monocausal/univariate*** explanations of linguistic phenomena are insufficient or contradictory (e.g. Gries 2003a)

  ← Lexical or syntactic choices made by speakers are determined, and can thus be explained by a **plurality** of factors, in **interaction**

  → necessity of ***multifactorial*** explanatory models → ***multivariate*** statistical analysis

Theoretical assumptions & methodological prerequisites

- ***Probabilistic*** grammar
  - Bod et al. (2003) and Bresnan (2007) have suggested that the selections of alternative selections on context, i.e. outcomes for combinations of variables, are generally speaking **probabilistic**
  - even though the individual choices in isolation are **discrete**
- In other words, the workings of a linguistic **system**, represented by the range of variables according to some **theory**, and its resultant **usage** are
  - in practice **not categorical**, following from **exception-less** rules,
  - but rather exhibit degrees of **potential variation** which becomes evident over **longer stretches of linguistic usage**
  - Integral characteristic of language – ***not*** a result of ***"interference"*** from language-external cognitive processes

Discrete vs. probabilistic

- …
- X**A**Y
- Y**B**X
- X**A**Y
- X**A**Y
- X**A**Y
- X**A**Y
- Y**B**X
- X**C**Y
- …

- X_Y
  - A:4
  - C:1
- Y_X
  - B:2

- X,Y
  - A:5
  - B:2
  - C:1

Discrete vs. probabilistic – Interpretation of the previous data

- If we assume categorical rules, can we extract them?
  - Y_X -> B
  - X_Y -> A?/C?
  - X,Y -> A?/B?/C
- What do we assume about the nature of these rules and their relationship with the data?
  - Is e.g. ***feature order*** a permissible or truly relevant characteristic?
    - Y_X -> B ~ X_Y -> B?
  - Do we expect that some ***additional variables*** (e.g. extralinguistic or stylistic) – yet unnoticed – might explain away the remaining irregularities?
    - X_Y -> A
    - X_Y**W** -> C
  - Can we explain all cases ***exhaustively*** and ***categorically*** by adding new explanatory variables?

Probabilistic syntax visualized (Bresnan 2007)

- Or do we rather allow *a priori* for variation and proportionate occurrence in the scrutinized contexts
  - X,Y ->
    - A (62.5%) |
    - B (25%) |
    - C (12.5%)

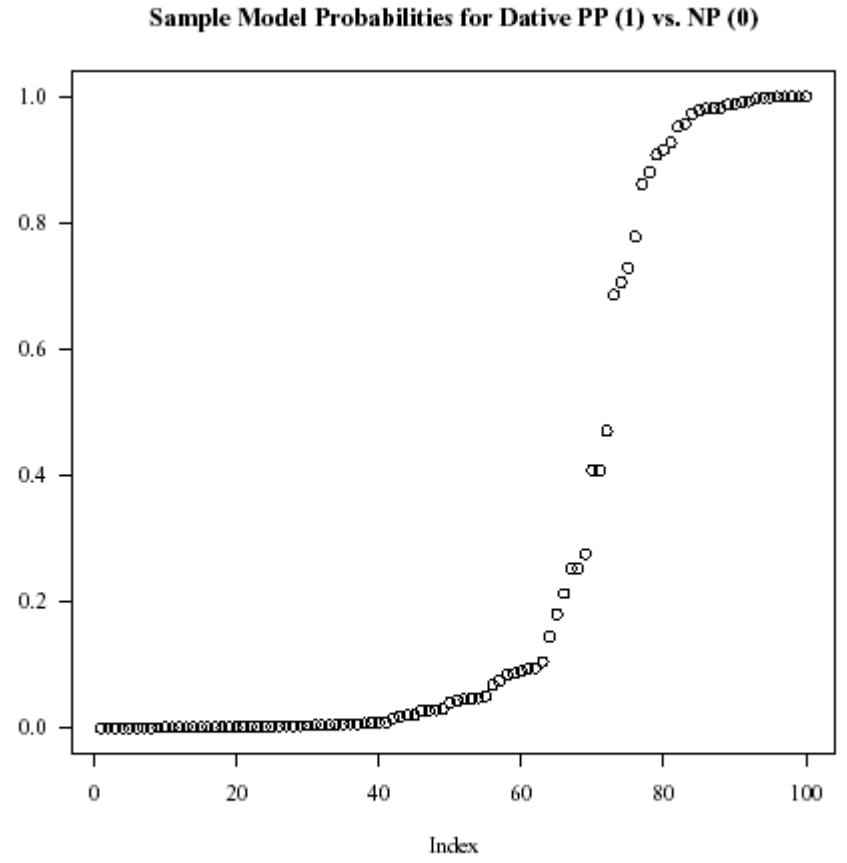


Figure 1: Sample probabilities from the corpus model of Bresnan et al.

Theoretical assumptions & methodological prerequisites

- ***Polytomous*** vs. ***dichotomous*** linguistic alternations: often more than two alternatives (cf. Divjak & Gries 2007; any [synonym] dictionary)
  - Structural alternation: English relative clauses
    - The book ***which*** I read was good.
    - The book ***that*** I read was good.
    - The book **[]** I read was good.
  - Lexical alternations: (English) synonyms
    - Do you ***understand*** what I mean?
    - Do you ***comprehend*** what I mean?
    - Do you ***grasp*** what I mean?
    - Do you ***get*** what I mean?
    - ….

Lexical alternation – practical example case

- Set of the most frequent synonyms denoting THINK in Finnish
  - *ajatella* < *ajaa* ’to drive habitually (in one’s mind)’
  - *miettiä* < *smetit’* Slavic (Baltic?) loan to the Fennic languages (i.e. 2000-3000 years old) cf. Swedish/Germanic *mäta* ’to measure’
  - *pohtia* ~ pohtaa < archaic/agricultural (→1950s) ’to winnow’
  - *harkita* < *harkki* archaic/agricultural ’dragnet’ ~ *haroa/haravoida* ’to rake’
  - [*tuumia/tuumata* < Russian *dumat’* ’to think’ (Slavic loan) cf. Swedish/Scandinavian *dömma* ’to judge, deem’]
- Currently translatable into English as:
  - ’think, reflect, ponder, consider’

Research corpus – two sources

- two months worth (January–February 1995) of written text from Helsingin Sanomat (1995)
  - Finland's major daily newspaper
  - 3,304,512 words of body text
  - excluding headers and captions, as well as punctuation tokens
  - 1,750 representatives of the studied THINK verbs
- six months worth (October 2002 – April 2003) of written discussion in the SFNET (2002-2003) Internet discussion forum, namely regarding
  - (personal) relationships (sfnet.keskustelu.ihmissuhteet)
  - politics (sfnet.keskustelu.politiikka)
  - 1,174,693 words of body text
  - excluding quotes of previous postings as well as punctuation tokens
  - 1,654 representatives of the studied THINK verbs
- the proportion of the THINK lexemes in the Internet newsgroup discussion text is more than twice as high as the corresponding value in the newspaper corpus
- The individual overall frequencies among the studied THINK lexemes in the research corpora were
  - 1492 for *ajatella*
  - 812 for *miettiä*
  - 713 for *pohtia*
  - 387 for *harkita*

Explanatory variables – overview

Selected on the basis of extensive univariate analysis

→ Altogether 48 contextual feature variables:

- Morphological features pertaining to the node-verb or the entire verb-chain they are components of (10)
- semantic characterizations of verb-chains (6)
- syntactic argument types, without any subtypes (10)
- Syntactic arguments combined with their semantic and structural subtypes (20)
- extra-linguistic features (2)

Overall model

- {ajatella|miettiä|pohtia|harkita} ~ Z_ANL_NEG + Z_ANL_IND + Z_ANL_KOND + Z_ANL_PASS + Z_ANL_FIRST + Z_ANL_SECOND + Z_ANL_THIRD + Z_ANL_PLUR + Z_ANL_COVERT + Z_PHR_CLAUSE + SX_AGE.SEM_INDIVIDUAL + SX_AGE.SEM_GROUP + SX_PAT.SEM_INDIVIDUAL_GROUP + SX_PAT.SEM_ABSTRACTION + SX_PAT.SEM_ACTIVITY + SX_PAT.SEM_EVENT + SX_PAT.SEM_COMMUNICATION + SX_PAT.INDIRECT_QUESTION + SX_PAT.DIRECT_QUOTE + SX_PAT. + SX_PAT. + SX_LX_että_CS.SX_PAT + SX_SOU + SX_GOA + SX_MAN.SEM_GENERIC + SX_MAN.SEM_FRAME + SX_MAN.SEM_POSITIVE + SX_MAN.SEM_NEGATIVE + SX_MAN.SEM_AGREEMENT + SX_MAN.SEM_JOINT + SX_QUA + SX_LOC + SX_TMP.SEM_DEFINITE + SX_TMP.SEM_INDEFINITE + SX_DUR + SX_FRQ + SX_META + SX_RSN_PUR + SX_CND + SX_CV + SX_VCH.SEM_POSSIBILITY + SX_VCH.SEM_NECESSITY + SX_VCH.SEM_EXTERNAL + SX_VCH.SEM_VOLITION + SX_VCH.SEM_TEMPORAL + SX_VCH.SEM_ACCIDENTAL + Z_EXTRA_SRC_sfnet + Z_QUOTE

Selection of multivariate statistical method

- ***Logistic regression*** – *WHY?*
  - Looks at outcomes as proportions among all observations with the same context
    - rather than individual *either-or* dichotomies of occurrence vs. non-occurrence
    - Thus estimates ***probabilities of occurrence*** given a particular context
    - Thus, also compatible with the probabilistic view of language
  - Estimates variable parameters which can be interpreted "naturally" as ***odds*** (Harrell 2001)
    - How much does the existence of a variable (i.e. feature) in the context increase (or decrease) the *chances* of a particular outcome (i.e. lexeme) to occur, with all the other explanatory variables being equal?

Logistic regression – formalization of binary (dichotomous) setting

- Model X with M explanatory variables {X} and parameters $\{\alpha_k, \beta_k\}$ for outcome Y=k:

  $X=\{X_1, \ldots, X_M\}$

  $\beta_k X = \beta_{k,1}X_1 + \beta_{k,2}X_2 + \ldots + \beta_{k,M}X_M$

  $P_k(X) = P(Y=k|X)$; $P_{\neg k}(X) = P(Y=\neg k|X) = 1 - P(Y=k|X)$

- **$\text{logit}[P_k(X)] = \log_e\{P_k(X)/[1-P_k(X)]\} = \alpha_k + \beta_k X$**

  $\Leftrightarrow P_k(X)/[1-P_k(X)] = \exp(\alpha_k+\beta_k X)$

  $\Leftrightarrow$ **$P_k(X)/[1-P_k(X)] = \exp(\alpha_k)\cdot\exp(\beta_k X)$**

  $= \exp(\alpha_k)\cdot\exp(\beta_{k,1}X_1)\cdot \ldots \cdot\exp(\beta_{k,M}X_M)$

  $\Leftrightarrow P_k(X) = 1/[1+\exp(-\alpha_k-\beta_k X)]$

Binary logistic regression – a concrete example …

Miten$_{MANNER+GENERIC}$ *ajattelit*$_{INDICATIVE+SECOND,\ COVERT,\ AGENT+INDIVIDUAL}$ *erota*$_{PATIENT+INFINITIVE}$ *… jostain … SAKn kannattajasta?* [sfnet]

'How did you think to differ at all from some dense supporter of class-thinking in SAK?'

Context ⊂ X =

{MANNER:GENERIC,

INDICATIVE, SECOND_PERSON,

COVERT_AGENT, AGENT:INDIVIDUAL,

PATIENT:INFINITIVE,

SFNET}

Binary logistic regression – a concrete example …

Miten$_{MANNER+GENERIC}$ ***ajattelit***$_{INDICATIVE+SECOND,\ COVERT,\ AGENT+INDIVIDUAL}$ *erota*$_{PATIENT+INFINITIVE}$ *… jostain … SAKn kannattajasta?* [sfnet]

'How did you **think** to differ at all from some dense supporter of class-thinking in SAK?'

$\log_e$[P(*ajatella*|*Context*)/ P(¬*ajatella*|Context)]	⇔	P(*ajatella*|*Context*)/ P(¬*ajatella*|Context)	⇔	P(*ajatella*|*Context*)
=0.5 ≈ $\log_e$[(3404-1492)/3404]		= 3:2		= 319/(1+319)
+3.0 ~ MANNER:GENERIC		· (41:2) ~ MANNER:GENERIC		≈ 1.0
+0.6 ~ INDICATIVE		· (13:7) ~ INDICATIVE		
–(0.5) ~ SECOND_PERSON		· (1:2) ~ SECOND_PERSON		
+(0.0) ~ COVERT_SUBJECT		· (1:1) ~ COVERT_SUBJECT		
–(0.2) ~ AGENT:INDIVIDUAL		· (5:6) ~ AGENT:INDIVIDUAL		
+(1.8) ~ PATIENT:INFINITIVE		· (6:1) ~ PATIENT:INFINITIVE		
+(0.5) ~ [INTERNET-GENRE]		· (3:2) ~ [INTERNET-GENRE]		
≈ +5.8		= 319:1		

Binary logistic regression – another concrete example …

Vilkaise$_{CO\text{-}ORDINATED\_VERB(+MENTAL)}$ *joskus*$_{FREQUENCY(+SOMETIMES)}$ *valtuuston esityslistaa ja* ***mieti***$_{(IMPERATIVE+)SECOND,COVERT,\ AGENT+INDIVIDUAL}$ *monestako*$_{PATIENT+INDIRECT\_QUESTION}$ *asiasta sinulla on jotain tietoa.* [sfnet]

'Glance sometimes at the agenda for the council and **think** on how many issues you have some information.'

$\log_e$[P(*miettiä*|*Context*)/ P(¬*miettiä*|Context)] ⇔ P(*miettiä*|*Context*)/ P(¬*miettiä*|Context) ⇔ P(*miettiä*|*Context*)

=–2.0 ≈ $\log_e$(812/3404)
+ 0.8 ~ CO-ORDINATED_VERB
+ 0.6 ~ FREQUENCY
+ 0.7 ~ SECOND_PERSON
(+ 0.1) ~ COVERT_SUBJECT
(+ 0.0) ~ AGENT:INDIVIDUAL
+ 1.6 ~ PATIENT:INDIRECT_Q…
+ 0.7 ~ [INTERNET-GENRE]
≈ +2.5

=2:15 (Intercept)
· 29:13 ~ CO-ORDINATED_VERB
· 17:9 ~ FREQUENCY
· 2:1 ~ SECOND_PERSON
· (1:1) ~ COVERT_SUBJECT
· (1:1) ~ AGENT:INDIVIDUAL
· 24:5 ~ PATIENT:INDIRECT_Q…
· 2:1 ~ [INTERNET-GENRE]
≈ 12.6:1

= 12.6/(1+12.6)
≈ 0.93 (→ 0.88)

Binary logistic regression – still another concrete example …

Tarkastusviraston mielestä$_{META}$ *tätä ehdotusta*$_{PATIENT+ACTIVITY}$ *olisi*$_{CONDITIONAL+THIRD,\ COVERT}$ *syytä*$_{VERB\_CHAIN+NECESSITY}$ ***pohtia*** *tarkemmin*$_{MANNER+POSITIVE}$. [766/hs95_7542]

'In the opinion of the Revision Office there is reason to **ponder** this proposal more thoroughly.'

P(*pohtia*|*Context*)/P(¬*pohtia*|Context)
= 1:5 ~ Intercept (≈ 719/3404)
· (3:4) ~ META-COMMENT
· (4:3) ~ PATIENT:ACTIVITY
· (4:5) ~ CONDITIONAL (MOOD)
· (8:9) ~ THIRD_PERSON
· (8:9) ~ COVERT_AGENT
· (1:1) ~ VERB-CHAIN:NECESSITY
· (5:6) ~ MANNER:SUFFICIENT
≈ 4:33 ≈ 0.122:1 ≈ 1:8.2

⇔

P(*pohtia*|*Context*)
= 0.12/(1+0.12)
≈ 0.11 (→ 0.125)

Binary logistic regression – still another concrete example …

Tarkastusviraston mielestä$_{META}$ *tätä ehdotusta*$_{PATIENT+ACTIVITY}$ *olisi*$_{CONDITIONAL+THIRD,\ COVERT}$ *syytä*$_{VERB\_CHAIN+NECESSITY}$ ***pohtia*** *tarkemmin*$_{MANNER+POSITIVE}$. [766/hs95_7542]

'In the opinion of the Revision Office there is reason to **ponder** this proposal more thoroughly.'

P(*harkita*|*Context*)/P(¬*harkita*|Context)
= 4:41 ~ Intercept (≈ 387/3404)
· 3:2 ~ META-COMMENT
· 23:3 ~ PATIENT:ACTIVITY
· 14:5 ~ CONDITIONAL (MOOD)
· (22:15) ~ THIRD_PERSON
· (7:8) ~ COVERT_AGENT
· (10:7) ~ VERB-CHAIN:NECESSITY
· (2:1) ~ MANNER:SUFFICIENT
≈ 12:1

⇔

P(*harkita*|*Context*)
= 12/(1+12)
≈ 0.92 (→ 0.725)

Model fit – observed proportions vs. estimated probabilities

- Most frequent feature combination in data:
  - *n*{Z_ANL_IND, Z_ANL_THIRD, SX_AGE.SEM_INDIVIDUAL, SX_PAT.DIRECT_QUOTE}=88
- Observed frequencies

ajatella	miettiä	pohtia	harkita
0	31	57	0

- Observed proportions

ajatella	miettiä	pohtia	harkita
0.0	0.35	0.65	0.0

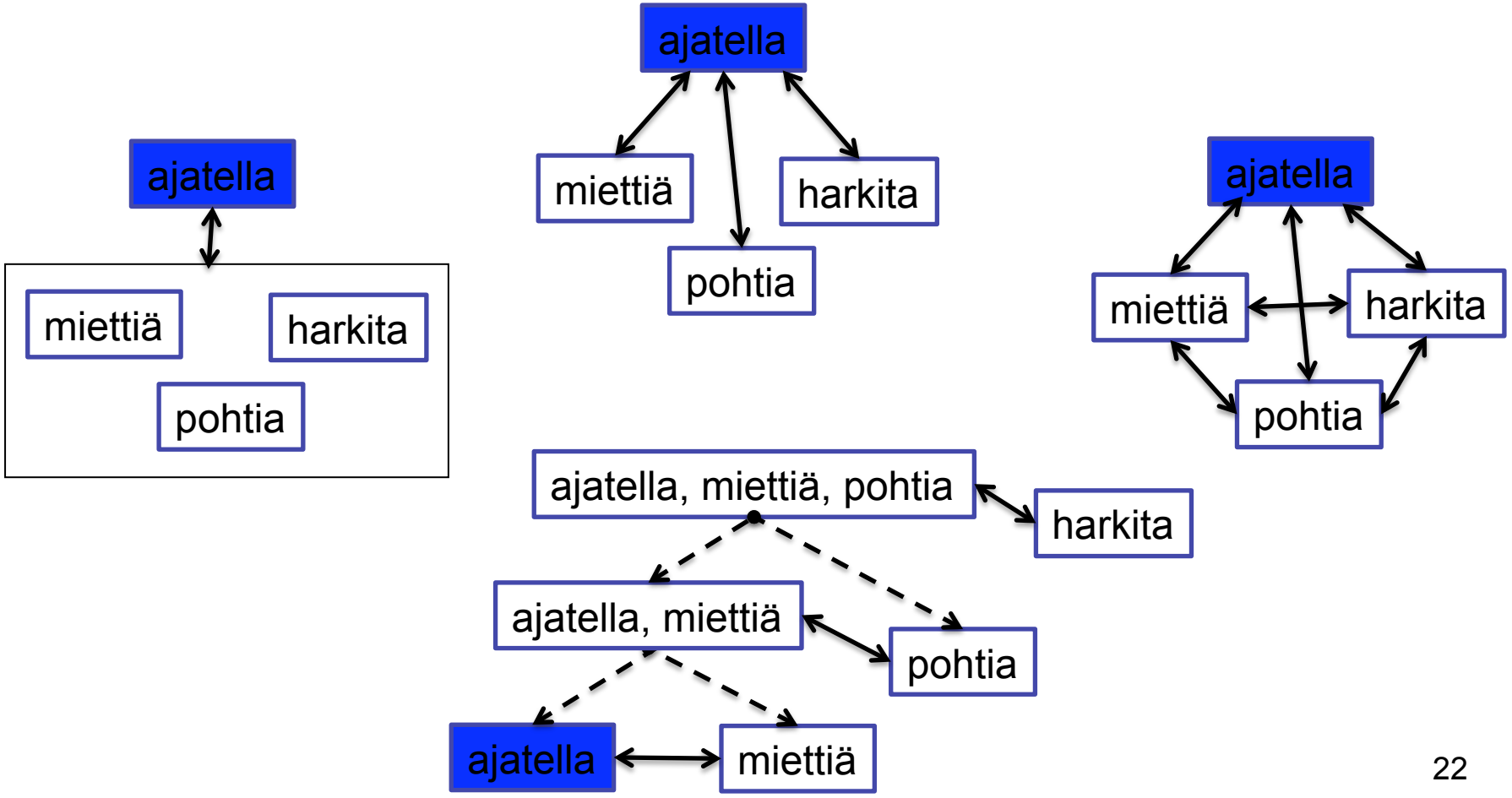
- Estimated probabilities

ajatella	miettiä	pohtia	harkita
0.03	0.37	0.60	0.00

Dichotomous → Polytomous setting

- Example case: four outcomes (i.e. synonyms)
  - {*ajatella, miettiä, pohtia, harkita*}
- How could the selection of these be broken down into a set of binary models?
  - N.B. *nnet:multinom* consists of binary models!

Polytomous outcome setting – *binarization* techniques



Dichotomous → Polytomous setting

- Several heuristic techniques for ***binarizing*** (dichotomizing) polytomous outcome settings
  - Baseline-category multinomial
    - simultaneously/separately fit
  - One-vs-rest (*one-against-all*)
  - Pairwise contrast (*all-against-all*, *round-robin*)
  - Nested dichotomy
  - Ensemble of nested dichotomies (ENDs)

Characteristic dimensions of polytomous logistic regression heuristics

- Number of constituent binary logistic regression models (→ complexity)
- Interpretation of explanatory variables in model(s) as well as the associated odds
  - Outcome-specific odds?
- Direct probability estimates for outcomes?
  - Necessity of normalization?
- Selection algorithm in prediction

Baseline-category multinomial

- **Reasoning**: one outcome is (manually/automatically) selected as a baseline category (most frequent, prototypical, or general), against which the other outcomes are contrasted each individually (Cox 1958)
  - Binary models may be fitted separately or dependently
- {*ajatella* vs. *miettiä*}, {*ajatella* vs. *pohtia*}, {*ajatella* vs. *harkita*}
- Variables and associated odds contrast other outcomes only with baseline (and not with each other)
- Number of binary models: *n(outcomes)–1*
- Direct probability estimates:
  - *P(baseline outcome) = 1-ΣP(non-baseline outcomes)*
  - Normalization of probabilities required, so that *ΣP(all outcomes)=1*

Baseline-category multinomial

(3.17) $P_k(X) = P(Y=k|X)$, with $\sum_{k=1\ldots K} P_k(X)=1$ and $k=\{1, \ldots, K\}$, and $P_K(X)=P(Y=K|X)=1-\sum_{k=1\ldots K-1} P_k(X)$ as the baseline case.

(3.18) $\log_e[P_k(X)/P_K(X)] = \alpha_k+\beta_k X \Leftrightarrow P_k(X) = \exp(\alpha_k+\beta_k X)/[1+\sum_{k=1\ldots K-1}\exp(\alpha_k+\beta_k X)]$ for $k=1\ldots K-1$ and $P_K(X) = 1-\sum_{k=1\ldots K-1} P_k(X)$ (the baseline thus assigned the "left-over" probability)

(3.19) $\beta_k X = \beta_{k,1}X_1 + \beta_{k,2}X_2 + \ldots + \beta_{k,M}X_M$

with classes $k=\{1, \ldots, K-1\}$, and M explanatory variables $X=\{X_1, \ldots, X_M\}$, parameters $\beta = \{(\beta_{1,1}, \ldots, \beta_{1,M}), (\beta_{2,1}, \ldots, \beta_{2,M}), \ldots, (\beta_{k-1,1}\ldots, \beta_{k-1,M})\}$, and constants $\alpha=\{\alpha_1, \ldots, \alpha_{k-1}\}$.

One-vs-rest

- **Reasoning**: Each outcome is contrasted with the undifferentiated bulk of the rest
  - In principle could be simultaneously fitted!
- {*ajatella* vs. ¬*ajatella*}

  ~ {*ajatella* vs. {*miettiä, pohtia, harkita*}, …
- Number of binary models: *n(outcomes)*
- Variables (and odds) distinguish individual outcomes against all the rest lumped together → highlight outcome-specific distinctive features
- Direct probability estimates:
  - *P(outcome)* generated directly, BUT
  - Normalization of probabilities required, so that *ΣP(all outcomes)=1*

One-vs-rest

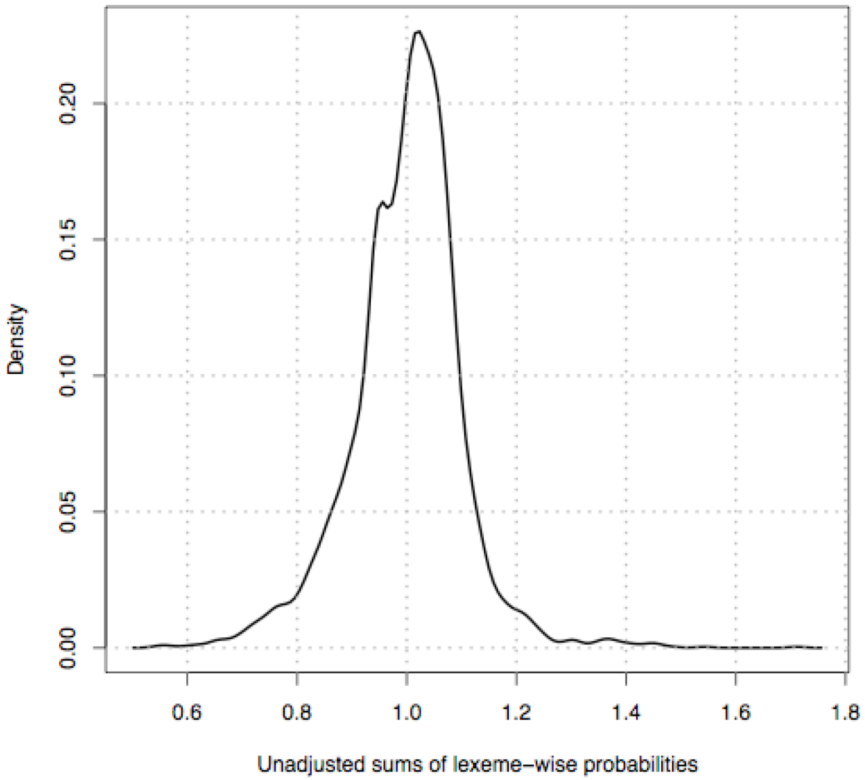
(3.23) $P_k(X) = P(Y=k|X)$, with and $k=\{1, \ldots, K\}$, and $P_{\neg k}(X) = P(Y=\neg k|X) = 1-P_k(X) = 1-P(Y=k|X)$ as the opposite case, that is, the 'rest', so naturally $P_k(X) + P_{\neg k}(X) = 1$ for each binary model.

(3.24) $\log_e\{P_k(X)/[1- P_k(X)]\} = \alpha_k+\beta_k X \Leftrightarrow P_k(X)/P_{\neg k}(X) = \exp(\alpha_k+\beta_k X)$

(3.25) $\beta_k X = \beta_{k,1}X_1 + \beta_{k,2}X_2 + \ldots + \beta_{k,M}X_M$

with classes $k=\{1, \ldots, K\}$, and M explanatory variables $X=\{X_1, \ldots, X_M\}$, parameters $\beta = \{(\beta_{1,1}, \ldots, \beta_{1,M}), (\beta_{2,1}, \ldots, \beta_{2,M}), \ldots, (\beta_{K,1}\ldots, \beta_{K,M})\}$, and constants $\alpha=\{\alpha_1, \ldots, \alpha_K\}$

One-vs-rest


Density
0.00
0.05
0.10
0.15
0.20
0.6
0.8
1.0
1.2
1.4
1.6
1.8
Unadjusted sums of lexeme−wise probabilities

Pairwise contrasts

- **Reasoning:** all outcomes are contrasted pairwise with each other
- {*ajatella* vs. *miettiä*}, {*ajatella* vs. *pohtia*}, {*ajatella* vs. *harkita*}, {*miettiä* vs. *ajatella*}, {*miettiä* vs. *pohtia*}, …
- Number of binary models:
  - Round-robin: *{n(outcomes)·[n/outcomes)-1)]}/2*
  - Double round-robin: *n(outcomes)·[n/outcomes)-1)]*
- Variables and odds sensitive to pairwise differences, but overall may exaggerate these and be difficult to interpret if distinctions are contradictory
  - Overall verb-feature odds can only be approximated as a geometric average of the pairwise odds
- No direct/approximate probability estimates

Pairwise contrasts

(3.26) $P_{k1/k2}(X) = [P(Y=k_1|X) \mid Y=\{k_1, k_2\}]$, and $P_{k2/k1}(X)=1-P_{k1/k2}(X) = 1-[P(Y=k_1|X) \mid Y=\{k_1, k_2\}]$

(3.27) $\log_e[P_{k1/k2}(X) \mid Y=\{k_1, k_2\}] = \alpha_{k1/k2}+\beta_{k1/k2}X$

(3.28) $\beta_{k1/k2}X = \beta_{k1/k2,1}X_1 + \beta_{k1/k2,2}X_2 + \ldots + \beta_{k1/k2,M}X_M$

(3.29) $\beta_{k1,m}\approx(\beta_{k1/k2,m}+\beta_{k1/k3,m}+ \ldots +\beta_{k1/K,m})/(K-1)$, since the geometric average of the binary log-odds is $[e^{\beta(1)}\cdot e^{\beta(2)}\cdot\ldots\cdot e^{\beta(K-1)}]^{1/(K-1)}=e^{[\beta(1)+\beta(2)+\ldots+\beta(K-1)]/(K-1)}$

(3.30) $P_{k1}(X)\approx\{n[P_{k1/k2}(X)>0.5]+n[P_{k2/k1}(X)\leq 0.5]\}/[K\cdot(K-1)]$; N.B. $0\leq P_{k1}(X)\leq 0.5$

with classes $k_1=\{1, \ldots, K\}$, and $k_2=\{1, \ldots, K\}$, with $k_1\neq k_2$, and M explanatory variables $X=\{X_1, \ldots, X_M\}$, parameters $\beta = \{(\beta_{1/2,1}, \ldots, \beta_{1,M}), \ldots, (\beta_{1/K,1}, \ldots, \beta_{1/K,M}), (\beta_{2/1,1}\ldots, \beta_{2/1,M}) \ldots, (\beta_{2/K,1}\ldots, \beta_{2/K,M}) \ldots, (\beta_{K/1,1}\ldots, \beta_{K/1,M}) \ldots, (\beta_{K/K-1,1}\ldots, \beta_{K/K-1,M})\}$, and constants $\alpha=\{\alpha_{k1/k2}, \alpha_{k1/k3}, \ldots, \alpha_{K/K-2}, \alpha_{K/K-1}\}$

Baseline vs. *One-vs-rest* vs. *Pairwise* contrasts

Feature/Verb	miettiä	pohtia	harkita
Z_ANL_SG12	2	1/5	1/1
Z_ANL_SG3	2	1	2
Z_ANL_PL12	3	1	2
SX_AGE.SEM_INDIVIDUAL	1/1	1/1	1/2
SX_AGE.SEM_GROUP	2	5	4

Feature/Verb	ajatella	miettiä	pohtia	harkita
Z_ANL_SG12	(1)	2	1/6	(1/1)
Z_ANL_SG3	1/2	2	(1)	(1)
Z_ANL_PL12	1/2	2	(1/1)	(1)
SX_AGE.SEM_INDIVIDUAL	1	(1)	(1/1)	1/2
SX_AGE.SEM_GROUP	1/3	1/2	3	2

Lexemes/Contextual features	ajatella	miettiä	pohtia	harkita
0_ANL_SINGURAL-1|2	0	+	-	0
0_ANL_SINGULAR-3	+-	-	+	0
0_ANL_PLURAL-1|2	-	0	+	0
SX_AGENT+SEM_INDIVIDUAL	0	0	+	-
SX_AGENT+SEM_GROUP	---	+-	+++	+-

Nested dichotomy

- **Reasoning:** Polytomous setting is partitioned into a successive set of dichotomies (Fox 1997)
  - Partitioning should be clearly naturally motivatable
- E.g. {*ajatella* vs. {*miettiä* vs. {*pohtia* vs. *harkita*}}
- Number of binary models: *n(outcomes)-1*
  - N.B. number of partitions: *T(1)=1;*
    T[n(outcomes)]= 2 · n(outcomes-3) · T(n(outcomes)-1)
- Overall variable odds can be generated as a product of the sequence of odds
- Direct probability estimates can be calculated exactly as a product of the sequence of probabilities in the appropriate partitions
  - No normalization is necessary

Nested dichotomy

- Consider e.g. the partition

  {*ajatella* vs. {*miettiä* vs. {*pohtia* vs. *harkita*}}}

  - The probability of the outcome *Y*=*harkita* for some given context and features (represented as *X*) is thus $P_{\{h\}|\{a,m,p\}}(Y=harkita|X)$

    $P_{\{m,p,h\}|\{a\}}(Y=\{miettiä, pohtia, harkita\}|X)$

    · $P_{\{p,h\}|\{m\}}(Y=\{pohtia, harkita\}|X)$

    · $P_{\{h\}|\{p\}}(Y=\{harkita\}|X)$

Ensemble of nested dichotomies

- **Reasoning:** Sample a set of partitions, when no obviously natural partitioning of the outcomes exists, and average over the results (Frank & Kramer 2004)
  - All partitions are considered equally likely, and may each represent fault-lines among the outcomes specific to one or more among the variables
  - 20 randomly sampled partitions sufficient
- Number of binary models: *20·[n(outcomes)-1]*
- Overall variable odds may be approximated as an average of the aggregate odds of the constituent partitioned models; the same applies for outcome-specific probability estimates

Summary overview – heuristics for polytomous logistic regression

Heuristic/ characteristics	Multinomial (baseline category)	One-vs-rest	Pairwise	Nested dichotomy	Ensemble of nested dichotomies
Number of constituent binary models	$n_{lex}-1$	n_{lex}	$n_{lex}\cdot(n_{lex}-1)/2$ (round-robin) $n_{lex}\cdot(n_{lex}-1)$ (double-round-robin)	$n_{lex}-1$	~20 partitions (each with $n_{lex}-1$)
Lexeme-specific odds-ratios for feature variables	No (Every lexeme against the baseline)	Yes (Every lexeme against the rest)	No (Approximation by geometric averages of binary odds-ratios)	Yes (Products of binary odds-ratios)	Yes (Averages of products of binary odds-ratios)
Probability estimates for lexemes (i.e., outcomes)	Direct	Direct $P_{lex/rest}(X)$	No	Direct (Product of probabilities at nodes in partition tree)	Direct (Average of products of probabilities at nodes in partition tree)
Selection of lexeme in prediction	Probability-based $\arg_{lex}\max(P_{lex}\|X)$	Probability-based $\arg_{lex}\max(P_{lex}\|X)$	Voting $\arg_{lex}\max\{n[P_{lex1/lex2}(X)>0.5] + n[P_{lex2/lex1}(X)\leq 0.5]\}$	Probability-based $\arg_{lex}\max(P_{lex}\|X)$	Probability-based $\arg_{lex}\max(P_{lex}\|X)$
Other	Necessity of baseline category	May not discover pairwise distinctions	May exaggerate pairwise distinctions, and the behavior with contradictory distinctions is problematic	Selection of single appropriate partition may be difficult or impossible	-

Comparisons of heuristics – model fit

Heuristic	R_L^2	**Recall (%)**	$\lambda_{prediction}$	$\tau_{classification}$
one-vs-rest	0.313	64.60	0.370	0.490
pairwise	NA	64.63	0.370	0.490
(simultaneous) multinomial	0.316	64.89	0.375	0.494
ensemble of nested dichotomies (END)	0.315	64.78	0.373	0.493
"best" nested dichotomies: {A, {H, {M, P}}} and {P, {A, {M, H}}}	NA	64.66	NA	NA
"worst" nested dichotomy: {{A, P}, {M, H}}	NA	63.66	NA	NA

Comparisons of heuristics – model fit

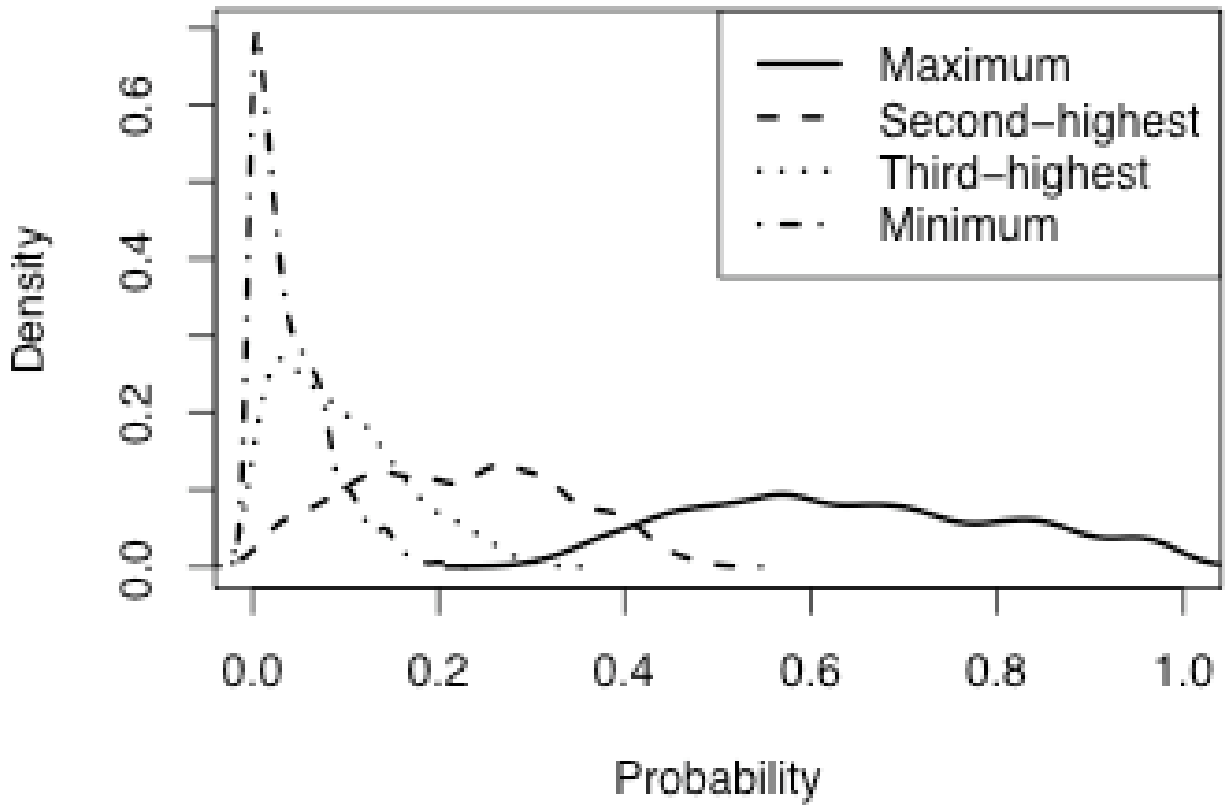
Heuristic	R_L^2	Recall (%)	$\lambda_{prediction}$	$\tau_{classification}$
one-vs-rest	0.287 (0.264, 0.300)	63.80 (63.07, 64.51)	0.355 (0.343, 0.368)	0.479 (0.468, 0.489)
pairwise	NA	63.79 (62.87, 64.57)	0.355 (0.339, 0.369)	0.478 (0.465, 0.490)
(simultaneous) multinomial	0.292 (0.276, 0.302)	63.78 (62.96, 64.51)	0.355 (0.340, 0.368)	0.478 (0.466, 0.489)
ensemble of nested dichotomies (END)	0.294 (0.277, 0.305)	63.89 (63.10, 64.63)	0.357 (0.343, 0.370)	0.480 (0.468, 0.490)
"best" nested dichotomy: {A, {H, {M, P}}}	NA	63.65 (62.87, 64,37)	NA	NA
"worst" nested dichotomy: {A, {P, {M, H}}}	NA	63.01 (61.93, 63.84)	NA	NA

Comparisons of heuristics – overlap of outcome selections

Heuristics	pairwise	multinomial (simultaneous)	ensemble of nested dichotomies
one-vs-rest	3279 (96.3%)	3325 (97.7%)	3360 (98.7%)
pairwise	-	3313 (97.3%)	3312 (97.3%)
multinomial (simultaneous)	-	-	3344 (98.2%)

Results – overall probabilities estimated by the full model

Figure 1. Probabilities of lexemes per each context.



Results – probabilities

- only 258 (7.6%) instances for which $P_{max}(L|C)>0.90$
- as many as 764 (22.4%) of the minimum estimated probabilities per instance are practically nil with $P_{min}(L|C)<0.01$
- the other way around, for 2640 (77.6%) instances the mimimum estimated probability $P_{min}(L|C)\geq 0.01$
  - i.e. representing an expected possibility of occurrence at least once every hundred times or even more often in a similar context.

Model fit revisited – Proportions vs. Probabilities

- Another frequent feature combination in data:
  - *n*{Z_ANL_IND, Z_ANL_THIRD, SX_AGE.SEM_GROUP, SX_PAT.SEM_ACTIVITY}=17
- Observed frequencies

ajatella	miettiä	pohtia	harkita
0	1	4	12

- Observed proportions

ajatella	miettiä	pohtia	harkita
0.0	0.06	0.24	0.71

- Estimated probabilities

ajatella	miettiä	pohtia	harkita
0.06	0.06	0.41	0.46

Model fit – Proportions vs. Probabilities

- Still another frequent feature combination in data:
  - *n*{Z_PHR_CLAUSE, SX_PAT.SEM_ABSTRACTION}=31
- Observed frequencies

ajatella	miettiä	pohtia	harkita
9	4	10	8

- Observed proportions

ajatella	miettiä	pohtia	harkita
0.29	0.13	0.32	0.26

- Estimated probabilities

ajatella	miettiä	pohtia	harkita
0.33	0.12	0.39	0.15

Exemplary contexts of usage – *ajatella*

A:#1 (7/2) P(*ajatella*|*Context*)=**1** P(*miettiä*|*Context*)=0 P(*pohtia*|*Context*)=0 P(*harkita*|*Context*)=0	*Miten*MANNER+GENERIC ***ajattelit***INDICATIVE +SECOND, COVERT, AGENT+INDIVIDUAL *erota*PATIENT+INFINITIVE … *jostain SAKn kannattajasta?* [sfnet] [3066/politiikka_9967] 'How did you **think** to differ at all from some dense supporter of class-thinking in SAK?'

Exemplary contexts – *miettiä*

M:#2 (7/1) P(*ajatella*|*Context*)=0.018 P(*miettiä*|*Context*)=**0.878** P(*pohtia*|*Context*)=0.084 P(*harkita*|*Context*)=0.02	*Vilkaise*CO-ORDINATED_VERB(+MENTAL) *joskus*FREQUENCY(+SOMETIMES) *valtuuston esityslistaa ja* ***mieti***(IMPERATIVE+)SECOND,COVERT, AGENT+INDIVIDUAL *monestako*PATIENT +INDIRECT_QUESTION *asiasta sinulla on jotain tietoa.* [sfnet] 'Glance sometimes at the agenda for the council and **think** on how many issuesyou have some information.'

Exemplary contexts – *pohtia*

P:#1 (6/3) P(*ajatella*|*Context*)=0.036 P(*miettiäContext*)=0.071 P(*pohtia*|*Context*)=**0.852** P(*harkita*|*Context*)=0.041	*Suomessa*LOCATION(+LOCATION) *kansalaisjärjestöt*AGENT+GROUP ***pohtivat***INDICATIVE+THIRD+PLURAL … *auttamisen periaatteita*PATIENT+NOTION … *eettisessä neuvottelukunnassa*LOCATION(+GROUP). [1259/ hs95_10437] 'In Finland civic organizations are **pondering** the principles of novel forms of assistance (e.g. the idenfication of an A-subscriber) in the so-called ethical advisory board of telephone assistance.'

Exemplary contexts – *harkita*

H:#1 (7/2) P(*ajatella*|*Context*)=0.025 P(*miettiäContext*)=0.115 P(*pohtia*|*Context*)=0.135 P(*harkita*|*Context*)=**0.725**	*Monen puoluetoverinkin mielestäMETA … Kauko JuhantalonAGENT+INDIVIDUAL olisiCONDITIONAL+THIRD pitänytVERB_CHAIN +NECESSITY* ***harkita*** *tarkemminMANNER +POSITIVE(<THOROUGH)* *ehdokkuuttaan.* [275/hs95_2077] 'In the opinion of many fellow party members, for instance Kauko Juhantalo should have **considered** more carefully his candidacy.'

Variation – “wrong” choice

H:#2 (8/2) P(*ajatella*|*Context*)=0.025 P(*miettiä*|*Context*)=0.125 P(*pohtia*|*Context*)=0.125 P(*harkita*|*Context*)=**0.725**	*Tarkastusviraston mielestäMETA tätä ehdotustaPATIENT+ACTIVITY olisiCONDITIONAL+THIRD, COVERT syytäVERB_CHAIN+NECESSITY* ***pohtia*** *tarkemminMANNER+POSITIVE.* [766/hs95_7542] ‘In the opinion of the Revision Office there is reason to **ponder** this proposal more thoroughly.’

Variation – equiprobable choice

8/1 (0.044) P(*ajatella*|*Context*)=**<u>0.301</u>** P(*miettiä*|*Context*)=0.272 P(*pohtia*|*Context*)=0.215 P(*harkita*|*Context*)=0.212	*Aluksi harvemmin, mutta myöhemmin tyttö alkoi viettää öitä T:n luona ja vuoden tapailun päätteeksi* *P*AGENT+INDIVIDUAL *sanoi, että voisi*CONDITIONAL +THIRD,VERB-CHAIN+POSSIBILITY,COVERT ***ajatella*** *asiaa*PATIENT+ABSTRACTION(<NOTION) *vakavamminkin*MANNER+POSITIVE. *[sfnet]* [50/ ihmissuhteet_8319] '... P said that [he] could **think** about the matter more seriously [perhaps]'

Synonymy – or not?

- *Aluksi harvemmin, mutta myöhemmin tyttö alkoi viettää öitä T:n luona ja vuoden tapailun päätteeksi P sanoi, että voisi **ajatella** asiaa vakavamminkin.*
  - Possibility to have an attitude/opinion concerning the 'issue' (*asia*)

  *... P sanoi, että voisi **miettiä** asiaa vakavamminkin.*
  - Actually give some occasional thought to the 'issue', without any expression of its duration, intensity

  *... P sanoi, että voisi **pohtia** asiaa vakavamminkin.*
  - Give the 'issue' serious, considerable and lengthy consideration

  *... P sanoi, että voisi **harkita** asiaa vakavamminkin.*
  - Consider the 'issue' with respect to making a decision one way or another concerning it
- None of these can be resolved on the basis of the immediate sentence context alone
- Might be deducible from prior passages in the text or extralinguistic knowledge about the context and/or the participants in the linguistic exchange

Results - discussion

- The recall rate seems to reach a ceiling at ~65%, and appears indifferent to whether some individual group of variables is left out
  - Do we yet lack some necessary variables or variable types?
  - Are some of the characteristics embedded in the synonymous lexeme itself, and not manifest – nor expressable – in any overt way in the immediate context (though possibly in the entire text or overall extralinguistic context)
  - Does this level represent the maximum that can be reached with the descriptive apparatus and associated variables of traditional grammatical analysis?
  - Might the remaining one-third represent to some extent cases of “true” synonymy and interchangeability?
- The results support Bresnan’s (2007) probabilistic view of the relationship between linguistic usage and the underlying linguistic system
  - Few choices are categorical, given the known context (feature cluster) that can be analytically grasped and identified
  - Rather, most contexts exhibit various degrees of variation as to their outcomes, resulting in proportionate choices on the long run
  - The question remains to what extent we are able o model this variation on the basis of current converntional linguistic theories
- These should be tested by comparing the predicted probabilities with selection in forced-choice experiments as well as acceptability ratings

The End

- Thank you!
- Questions, comments, suggestions?!?

Results (statistical)

- Measures of overall fit (Menard 1995)
  - recall rate = 65.6%
  - R_L^2 = 0.325
  - $\lambda_{prediction}$=0.387
  - $\tau_{classification}$=0.504
- Measures of model validation
  - 1000 repetitions of training the model with a simple bootstrap resample and then testing the model against the entire data, on the basis of which we finally calculate a mean and the 95% Confidence Intervals of the model statistics
  - recall rate = 63.8% (63.07-64.51%)
  - $R_{L\,(TEACH)}^2$ = 0.325 (0.307, 0.342)
  - $R_{L\,(TEST)}^2$ = 0.287 (0.264,0.300)
  - $\lambda_{prediction}$=0.355 (0.343, 0.368)
  - $\tau_{classification}$=0.479 (0.468, 0.489)
- Compare these with the 58-59% recall rate reported by Arppe (2006) using only semantic classifications of nominals (WordNet)

Results - comparison of models with different sets of feature categories I

Feature set composition	**Recall (%)**	R_L^2	$\lambda_{prediction}$	$\tau_{classification}$
Verb-chain general morphological features (10) as well as those node-specific features which are not subsumed by the verb-chain general ones (17)	47.71	0.100	0.069	0.247
Syntactic argument types, *without* their semantic and structural classifications	50.18	0.098	0.113	0.282
Extralinguistic features alone (2)	47.21	0.057	0.060	0.240

Results - comparison of models with different sets of feature categories II

Feature set composition	**Recall (%)**	R_L^2	$\lambda_{prediction}$	$\tau_{classification}$
Full model with verb-chain general morphological features (10) and their semantic classifications (6) together with syntactic argument types alone (10) or their selected or collapsed subtypes (20)	64.60	0.313	0.370	0.490
Full model with verb-chain general morphological features (10) and their semantic classifications (6) together with syntactic argument types alone (10) or their subtypes (20) as well as extra-linguistic features (2)	65.57	0.325	0.387	0.504

Results - relative importance of feature categories in final model

Feature variable category	Mean odds in favor	Mean odds against	Mean aggregate odds
Verb chain morphology	2.02 (1.48)	0.52~1:1.91 (0.70~1:1.44)	1.96 (1.46)
Verb chain semantics	2.66 (1.62)	0.24~1:4.24 (0.12~1:8.59)	3.17 (3.48)
Syntactic argument types (alone)	2.69 (1.84)	0.32~1:3.14 (0.47~1:2.11)	2.92 (1.99)
Syntax arguments + semantic/structural subtypes	3.71 (2.57)	0.21~1:4.70 (0.06~1:18)	4.13 (7.89)
Extralinguistic features	1.68 (1.68)	0.47~1:2.13 (0.56~1:1.80)	1.86 (1.74)

Results (linguistic) - Odds(lexeme <-- feature)

Lexeme/ Features	Strongest odds in favor of the lexeme	Strongest odds against the lexeme
ajatella	SX_MAN.SEM_GENERIC (23) SX_MAN.SEM_AGREEMENT (16) SX_VCH.SEM_ACCIDENTAL (5.6) SX_PAT.INFINITIVE (5.3) SX_PAT.PARTICIPLE (5.3)	SX_PAT.DIRECT_QUOTE (0.013~1:75) SX_PAT.INDIRECT_QUESTION (0.07~1:14) SX_PAT.SEM_COMMUNICATION (0.1~1:9.6) SX_DUR (0.12~1:8.4) SX_PAT.SEM_ACTIVITY (0.14~1:7.1)
miettiä	SX_PAT.INDIRECT_QUESTION (4.2) SX_DUR (3.4) SX_PAT.DIRECT_QUOTE (3) SX_PAT.SEM_COMMUNICATION (2.8) SX_QUA (2.6)	SX_MAN.SEM_AGREEMENT (0.07~1:14) SX_MAN.SEM_GENERIC (0.15~1:6.8) SX_MAN.SEM_FRAME (0.28~1:3.6) SX_AGE.SEM_GROUP (0.52~1:1.9) SX_LX_että_CS.SX_PAT (0.52~1:1.9)

Odds (lexeme <-- feature) (cont'd)

Lexeme/ Features	Strongest odds in favor of the lexeme	Strongest odds against the lexeme
pohtia	SX_PAT.DIRECT_QUOTE (8.1) SX_AGE.SEM_GROUP (4.2) SX_PAT.SEM_ABSTRACTION (4.1) SX_LOC (3.7) SX_PAT.SEM_COMMUNICATION (3)	SX_MAN.SEM_AGREEMENT (0.22~1:4.5) SX_MAN.SEM_NEGATIVE (0.22~1:4.6) SX_SOU (0.29~1:3.5) Z_ANL_FIRST (0.29~1:3.5) SX_PAT.SEM_INDIV..._GROUP (0.3~1:3.4)
harkita	SX_PAT.SEM_ACTIVITY (9) SX_CND (2.9) Z_ANL_KOND (2.3) SX_MAN.SEM_POSITIVE (1.8) SX_META (1.6)	SX_SOU (0.13~1:7.5) SX_VCH.SEM_TEMPORAL (0.15~1:6.5) SX_GOA (0.21~1:4.7) SX_LX_että_CS.SX_PAT (0.25~1:4) SX_MAN.SEM_FRAME (0.27~1:3.8)

Results (linguistic) - Odds (feature --> lexeme)

Contextual feature	Lexemes with strong odds in favor	Lexemes with neutral odds	Lexemes with strong odds against
SX_AGE.SEM_INDIVIDUAL	-	pohtia (1.6), miettiä (0.98), ajatella (0.85), harkita (0.69)	-
SX_AGE.SEM_GROUP	pohtia (4.2)	harkita (1.1)	miettiä (0.52), ajatella (0.2)
Z_ANL_FIRST	-	harkita (1.9), miettiä (1.8), ajatella (0.86)	pohtia (0.29)
Z_ANL_SECOND	miettiä (2.4)	ajatella (0.69), harkita (0.68)	pohtia (0.42)
Z_ANL_THIRD	-	harkita (1.6), miettiä (1.3), pohtia (0.99), ajatella (0.63)	-
Z_ANL_PLUR	pohtia (1.6)	harkita (1.2), ajatella (1.1)	miettiä (0.59)
Z_ANL_PASS	pohtia (1.9)	harkita (1.1), miettiä (0.89), ajatella (0.63)	-
Z_ANL_COVERT	-	miettiä (1.2), ajatella (1.1), harkita (0.79), pohtia (0.77)	-

Results (linguistic) - Estimated probabilities – categorical choice

Rank	A	M	P	H	Sentences
A:#1	1	0	0	0	*Miten*$_{MANNER+GENERIC}$ ***ajattelit***$_{INDICATIVE+SECOND,\ COVERT,\ AGENT+INDIVIDUAL}$ *erota*$_{PATIENT+INFINITIVE}$ *mitenkään jostain SAKn umpimielisistä luokka-ajattelun kannattajasta?* [3066/politiikka_9967] 'How did you **think** to differ at all from some dense supporter of class-thinking in SAK?'

Results (linguistic) - Estimated probabilities - dispersion

Rank	A	M	P	H	Sentences
M:#2	0.018	**0.878**	0.084	0.02	*Vilkaise*$_{CO\text{-}ORDINATED\_VERB(+MENTAL)}$ *joskus*$_{FREQUENCY(+SOMETIMES)}$ *valtuuston esityslistaa ja* ***mieti***$_{(IMPERATIVE+)SECOND,\ COVERT,\ AGENT+INDIVIDUAL}$ *monestako*$_{PATIENT\ +INDIRECT\_QUESTION}$ *asiasta sinulla on jotain tietoa.* [2815/politiikka_728] ‘Glance sometimes at the agenda for the council and **think** on how many issues you have some information.’

Results (linguistic) - Estimated probabilities – “wrong” selection

Rank	**A**	**M**	**P**	**H**	**Sentences**
H:#2 (8/2)	0.025	0.125	**0.125**	**0.725**	*Tarkastusviraston mielestä*$_{META}$ *tätä ehdotusta*$_{PATIENT+ACTIVITY}$ *olisi*$_{CONDITIONAL+THIRD}$ *syytä*$_{VERB\_CHAIN+NECESSITY}$ ***pohtia*** *tarkemmin*$_{MANNER+POSITIVE}$. [766/ hs95_7542] ‘In the opinion of the revision office there is reason to **ponder** this proposal more thoroughly.’

Results (linguistic) - Estimated probabilities – P(all)~0.25 -> synonymy?

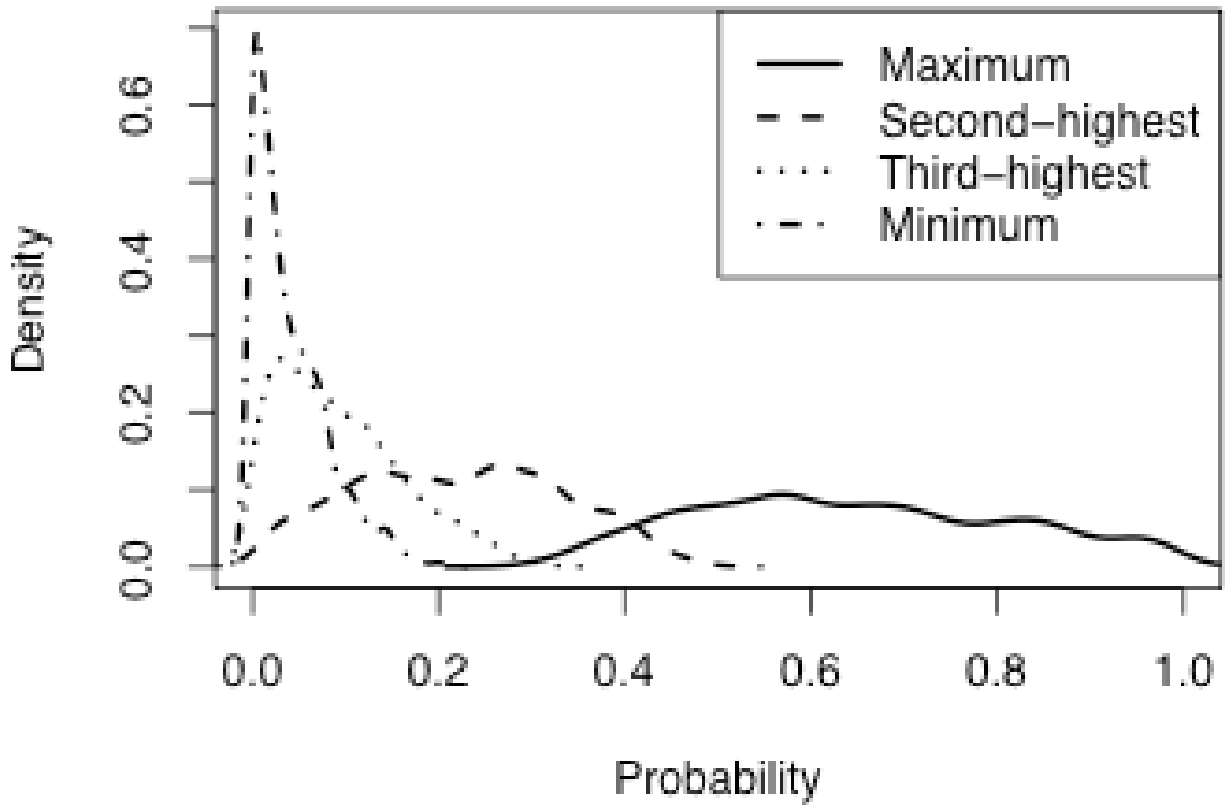
σ(P)	A	M	P	H	Sentences
0.044	**0.301**	0.272	0.215	0.212	*Aluksi harvemmin, mutta myöhemmin tyttö alkoi viettää öitä T:n luona ja vuoden tapailun päätteeksi* $P_{AGENT+INDIVIDUAL}$ *sanoi, että voisi*$_{CONDITIONAL+THIRD,VERB-CHAIN+POSSIBILITY,COVERT}$ ***ajatella*** *asiaa*$_{PATIENT+NOTION}$ *vakavamminkin*$_{MANNER+POSITIVE}$. [50/ihmissuhteet_8319] '... P said that [she] could think about the matter more seriously [perhaps]'

Synonymy – or not?

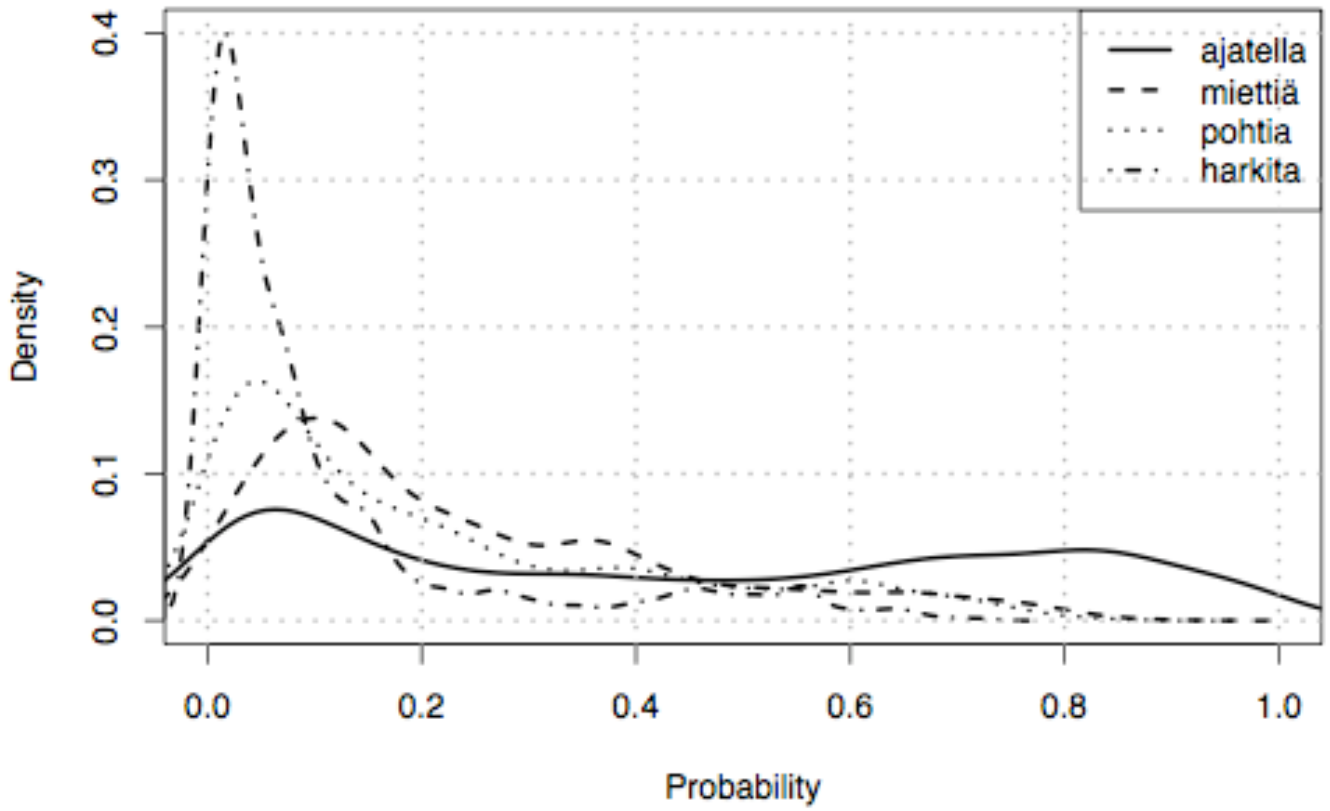
- *Aluksi harvemmin, mutta myöhemmin tyttö alkoi viettää öitä T:n luona ja vuoden tapailun päätteeksi P sanoi, että voisi **ajatella** asiaa vakavamminkin.*
  - Possibility to have an attitude/opinion concerning the 'issue' (*asia*)
- *... P sanoi, että voisi **miettiä** asiaa vakavamminkin.*
  - Actually give some occasional thought to the 'issue', without any expression of its duration, intensity
- *... P sanoi, että voisi **pohtia** asiaa vakavamminkin.*
  - Give the 'issue' serious, considerable and lengthy consideration
- *... P sanoi, että voisi **harkita** asiaa vakavamminkin.*
  - Consider the 'issue' with respect to making a decision one way or another concerning it
- None of these can be resolved on the basis of the immediate sentence context alone
- Might be deducible from prior passages in the text or extralinguistic knowledge about the context and/or the participants in the linguistic exchange

Results - overall probabilities estimated by the full model

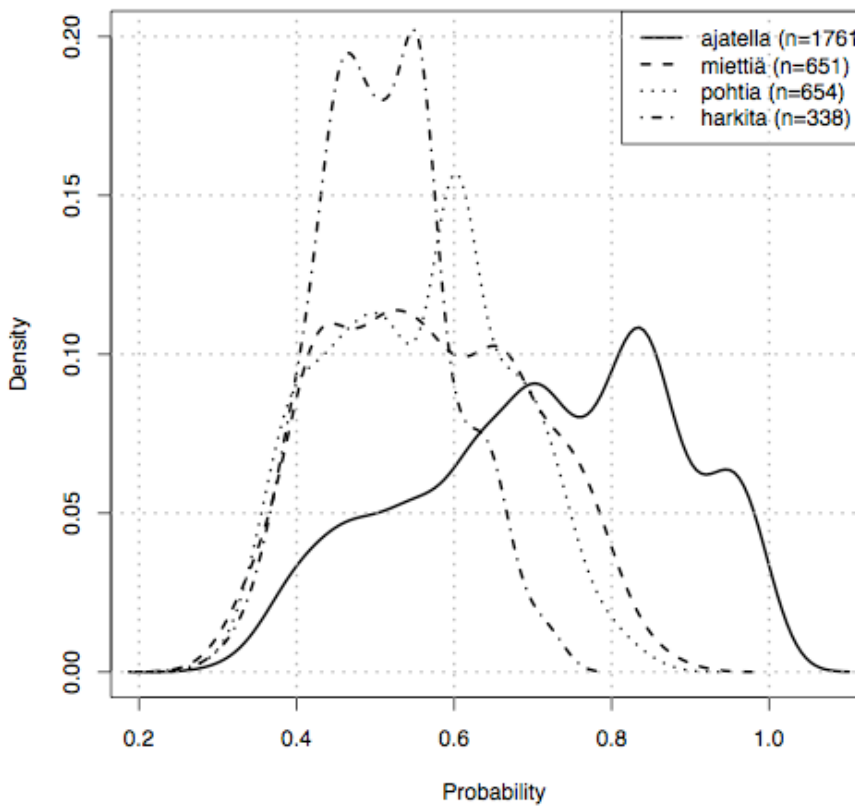
Figure 1. Probabilities of lexemes per each context.

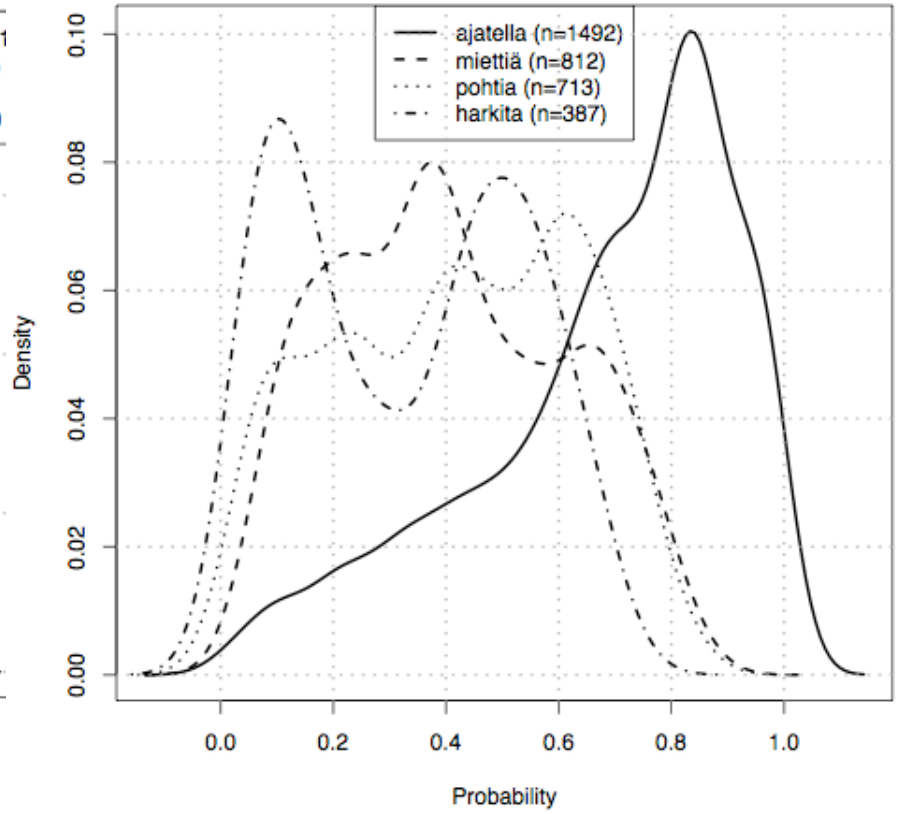


Results - overall probabilities estimated per lexeme by the full model



Results - overall probabilities estimated per lexeme by the full model





Results - probabilities

- maximum probability per all instances and contexts
  - Mean as low as x($P_{max}[L|C]$)=0.636,
  - overall span of maximal values as broad as (0.28, 1.00)
  - 95% CI=(0.369, 0.966).
- second-highest probability estimates per instances
  - mean x($P_{max\text{-}1}[L|C]$)=0.244
  - overall range of (0.000, 0.490)
  - 95% CI=(0.026, 0.415)
- third-highest probability estimates
  - mean x($P_{max\text{-}2}[L|C]$)=0.096
  - overall range of (0.000, 0.307)
  - 95% CI=(0.000, 0.241)
- minimum probability estimates
  - clearly keep some distance from zero as their mean x($P_{min}[L|C]$)=0.043
  - even though their overall range is (0.000, 0.212) as well as 95% CI=(0.000, 0.144)

Results – probabilities

- only 258 (7.6%) instances for which $P_{max}(L|C)>0.90$
- as many as 764 (22.4%) of the minimum estimated probabilities per instance are practically nil with $P_{min}(L|C)<0.01$
- the other way around, for 2640 (77.6%) instances the mimimum estimated probability $P_{min}(L|C)\geq 0.01$
  - i.e. representing an expected possibility of occurrence at least once every hundred times or even more often in a similar context.

Results - discussion

- The recall rate seems to reach a ceiling at ~65%, and appears indifferent to whether some individual group of variables is left out
  - Do we yet lack some necessary variables or variable types?
  - Are some of the characteristics embedded in the synonymous lexeme itself, and not manifest – nor expressable – in any overt way in the immediate context (though possibly in the entire text or overall extralinguistic context)
  - Does this level represent the maximum that can be reached with the descriptive apparatus and associated variables of traditional grammatical analysis?
  - Might the remaining one-third represent to some extent cases of “true” synonymy and interchangeability?
- The results support Bresnan’s (2007) probabilistic view of the relationship between linguistic usage and the underlying linguistic system
  - Few choices are categorical, given the known context (feature cluster) that can be analytically grasped and identified
  - Rather, most contexts exhibit various degrees of variation as to their outcomes, resulting in proportionate choices on the long run
  - The question remains to what extent we are able o model this variation on the basis of current converntional linguistic theories
- These should be tested by comparing the predicted probabilities with selection in forced-choice experiments as well as acceptability ratings

The End

- Thank you!
- Questions, comments, suggestions?!?