# Statistical classification and principles of human learning

Antti Arppe[1 2] Harald Baayen[2]

[1]Department of Modern Languages
University of Helsinki

[2]Department of Linguistics
University of Alberta

4th Conference on Quantitative Investigations in Theoretical Linguistics

# Outline

# Theoretical questions - generalizing or piecemeal learning?

- Various statistical machine-learning techniques may seemingly faithfully and accurately mimic overall human linguistic behavior (e.g. in terms of choices in produced texts and utterances).
- But do the premises of these machine-learning techniques and the resultant internal representations correctly reflect those of human learning processes and cognitive structures?
- Most multivariate statistical/computational methods optimize over the entire accumulated data, assuming the maximization of likelihood with optimization algorithms – but how cognitively realistic is this assumption?
- Might human learning rather fundamentally operate incrementally, absorbing (new) information in a piecemeal fashion?

# Theoretical questions - frequencies and probabilities

- The good performance of machine learning techniques in representing human linguistic behavior suggests that their fundamental characteristic – keeping track co-occurrence frequencies and associated probabilities – should also somehow be an integral component of also human learning. But need this necessarily hold?
- How may it be possible that the brain appears to be sensitive and receptive to assimilating probabilistic information in linguistic usage – but not internally representing it in the same way as machine-learning methods?
- A Metaphor – bicycle riding: do people apply the calculation of Newtonian physics or something much simpler?

# Objectives





- Compare the performance of several well-established machine-learning classifiers and a new parameter-free model of naive discriminative learning based on principles of a human learning process
- If they fare equally well, what could that reveal us about the nature of human learning – in comparison to machine-learning?

# Linguistic phenomenon and data





- Lexical choice of near-synonymous words in context
- The four most frequent Finnish verbs denoting think: *ajatella*, *miettiä*, *pohtia*, and *harkita* (Arppe 2008; Arppe & Järvikivi 2007 [QITL1]; Arppe 2006 [QITL2])
- Altogether 3,404 instances in Finnish newspaper and Internet newsgroup discussion (SFNET) text
- Analyzed in term of the morphological and syntactic structure of the verbs and their context – supplemented with semantic and structural subclassifications

# Linguistic variables



- In all 6000 distinct contextual features (including lexemes) observable in the contexts – 46 selected for this study:
  - 10 morphological features of the verb or verb chain
  - 6 semantic characterizations of the verb chain
  - 10 syntactic argument types
  - 20 combinations of syntactic arguments + semantic subclassifications
  - (Random effects: Register, Subsection, Author)

# Multivariate statistical methods

- Polytomous logistic regression (Arppe 2008)
    - iterative optimization of model fit (in terms of maximum likelihood) over entire data
- Polytomous mixed-effects logistic regression (Arppe, in prep.)
    - Poisson reformulation
- Support vector machine (Vapnik 1995)
    - kernel methods
- Memory-based learning (Daelemans & Bosch 2005)
    - nearest-neighbor similarity-based inference (incorporating exemplars from the entire data)
- Random forests (Breiman 2001)
    - recursive conditioning with sumbsampling (sets of conditional inference trees)
- Naive Discriminative Learning (Baayen et al. 2011)

# Multivariate statistical methods





- Polytomous logistic regression (Arppe 2008)
  - iterative optimization of model fit (in terms of maximum likelihood) over entire data
- Polytomous mixed-effects logistic regression (Arppe, in prep.)
  - Poisson reformulation
- Support vector machine (Vapnik 1995)
  - kernel methods
- Memory-based learning (Daelemans & Bosch 2005)
  - nearest-neighbor similarity-based inference (incorporating exemplars from the entire data)
- Random forests (Breiman 2001)
  - recursive conditioning with sumbsampling (sets of conditional inference trees)
- Naive Discriminative Learning (Baayen et al. 2011)

# Naive Discriminatory learning

- Based on the *Rescorla-Wagner* equations (1972)
- Proven to be surprisingly fruitful in human and animal learning (Miller, Barnet & Grahame 1995)
- Basically models *incremental* learning in response to co-occurrences of outcomes and cues – adjusts weights for associations of such outcomes and cues with each new experience
- Association weights in the end result of a learning process (representing a saturated "stable" state) can be estimated with *equilibrium equations* (Danks 2003)
- Baayen et al. (2011) have incorporated these equilibrium equations into a general discriminative learning model – naive in the sense of naive Bayesian classifiers

# Rescorla-Wagner (1972) equations



Let present$(X, t)$ denote the presence of a cue (predictor value) or outcome (one of the four Finnish think verbs) $X$ at time $t$, and absent$(X, t)$ denote its absence at time $t$.
The Rescorla-Wagner equations specify the association strength $V_i^{t+1}$ of cue $C_i$ with outcome $O$ at time $t+1$ using a recurrence equation, as follows:

$$V_i^{t+1} = V_i^t + \Delta V_i^t. \tag{1}$$

The change in association strength $\Delta V_i^t$ defined as

$$\Delta V_i^t = \begin{cases} 0 & \text{if absent}(C_i, t) \\ \alpha_i \beta_1 \left(\lambda - \sum_{\text{present}(C_j,\, t)} V_j\right) & \text{if present}(C_j, t) \text{ \& present}(O, t) \\ \alpha_i \beta_2 \left(0 - \sum_{\text{present}(C_j,\, t)} V_j\right) & \text{if present}(C_j, t) \text{ \& absent}(O, t) \end{cases} \tag{2}$$

# Rescorla-Wagner equations





Represent incremental learning and subsequently on-going adjustments to an accumulating body of knowledge:
Changes in association strengths:

- If a cue is not present in the input, no change
- Increased when the cue and outcome co-occur
- Decreased when the cue occurs without the outcome
- The more cues are present simultaneously, the smaller the adjustments are

# Danks (2003) equilibrium equations



$$\Pr(O|C_i) - \sum_{j=0}^{n} \Pr(C_j|C_i)V_j = 0 \qquad (3)$$

- make it possible to estimate the weights for an 'adult/stable' system by solving the above set of equations using the co-occurrence vector of a specific outcome (verb) given the different predictor values and the co-occurrence matrix of predictor values.
- provide a convenient short-cut to calculating the consolidated cue-outcome association weights resulting from incremental learning
- the learning parameters ($\lambda$, $\alpha_i$, $\beta_i$) of the Rescorla-Wagner equations drop out of the equilibrium equations

# Danks equilibrium equations



Alternatively can be formulated with matrix notation:

$$CW = O \tag{4}$$

where:

- $C$ is the matrix of conditional probabilities cues, given other cues
- $W$ is the matrix of unknown weights, representing outcome-cue associations, to be estimated; and
- $O$ is the matrix of conditional probabalities of outcomes, given some set of cues.

$W$ can be solved using the generalized inverse, yielding a solution that is optimal in the least-squares sense.

# Stable adult state – simple example case



The association of semantic subtypes of Agents and Patients with the occurrence of Finnish think verbs:

| | *Lexeme* | *Agent* | *Patient* |
|---|---|---|---|
| 1 | *pohtia* | *None* | *Abstraction* |
| 2 | *harkita* | *Group* | *Activity* |
| 3 | *miettia* | *Individual* | *DirectQuote* |
| 4 | *miettia* | *Individual* | *IndirectQuestion* |
| 5 | *ajatella* | *Individual* | *etta.CLAUSE* |
| 6 | *ajatella* | *Individual* | *Abstraction* |
| ... | ... | ... | ... |
| 3404 | *ajatella* | *None* | *Abstraction* |

# Stable adult state – simple example case

Matrix $M$ of cue co-occurrences for Agent and Patients of Finnish think verbs:

$$M = \left\{ \begin{array}{lrrrr} & \mathit{Agent} & \mathit{Agent} & \mathit{Agent} & \ldots \\ & \mathit{Group} & \mathit{Individual} & \mathit{NoAgent} & \ldots \\ \mathit{AgentGroup} & 256 & 0 & 0 & \ldots \\ \mathit{AgentIndividual} & 0 & 2251 & 0 & \ldots \\ \mathit{AgentNoAgent} & 0 & 0 & 897 & \ldots \\ \mathit{PatientAbstraction} & 70 & 392 & 236 & \ldots \\ \mathit{PatientActivity} & 90 & 225 & 174 & \ldots \\ \mathit{PatientCommunication} & 1 & 30 & 11 & \ldots \\ \mathit{PatientDirectQuote} & 1 & 106 & 0 & \ldots \\ \mathit{Patientetta.CLAUSE} & 7 & 324 & 65 & \ldots \\ \mathit{PatientDirectQuestion} & 37 & 330 & 71 & \ldots \\ \mathit{PatientIndividualGroup} & 3 & 77 & 29 & \ldots \\ \mathit{PatientInfinitive} & 3 & 33 & 5 & \ldots \\ \mathit{PatientParticiple} & 5 & 53 & 15 & \ldots \\ \mathit{PatientNoPatient} & 39 & 681 & 291 & \ldots \end{array} \right\} \tag{5}$$

# Stable adult state – simple example case

Matrix $C$ of conditional probabilities of $cue_j$ given $cue_i$:

$$C = \left\{ \begin{array}{lrrrr} & Agent & Agent & Agent & \ldots \\ & Group & Individual & NoAgent & \ldots \\ AgentGroup & 0.50 & 0.00 & 0.00 & \ldots \\ AgentIndividual & 0.00 & 0.50 & 0.00 & \ldots \\ AgentNoAgent & 0.00 & 0.00 & 0.50 & \ldots \\ PatientAbstraction & 0.14 & 0.09 & 0.13 & \ldots \\ PatientActivity & 0.18 & 0.05 & 0.10 & \ldots \\ PatientCommunication & 0.00 & 0.01 & 0.01 & \ldots \\ PatientDirectQuote & 0.00 & 0.02 & 0.00 & \ldots \\ Patientetta.CLAUSE & 0.01 & 0.07 & 0.04 & \ldots \\ PatientIndirectQuestion & 0.07 & 0.07 & 0.04 & \ldots \\ PatientIndividualGroup & 0.01 & 0.02 & 0.02 & \ldots \\ PatientInfinitive & 0.01 & 0.01 & 0.00 & \ldots \\ PatientParticiple & 0.01 & 0.01 & 0.01 & \ldots \\ PatientNoPatient & 0.08 & 0.15 & 0.16 & \ldots \end{array} \right\} \tag{6}$$

# Stable adult state – simple example case



Matrix $N$ lists the co-occurrences of outcomes (columns: Verbs) and cues (rows: Agents & Patients):

$$N = \begin{array}{lrrrr} & ajatella & harkita & miettia & pohtia \\ AgentGroup & 37 & 64 & 36 & 119 \\ AgentIndividual & 1047 & 198 & 632 & 374 \\ AgentNoAgent & 408 & 125 & 144 & 220 \\ PatientAbstraction & 192 & 57 & 190 & 259 \\ PatientActivity & 83 & 213 & 72 & 121 \\ PatientCommunication & 6 & 7 & 19 & 10 \\ PatientDirectQuote & 2 & 0 & 41 & 64 \\ Patientetta.CLAUSE & 317 & 8 & 48 & 23 \\ PatientIndirectQuestion & 38 & 26 & 242 & 132 \\ PatientIndividualGroup & 87 & 7 & 11 & 4 \\ PatientInfinitive & 37 & 3 & 0 & 1 \\ PatientParticiple & 65 & 5 & 0 & 3 \\ PatientNoPatient & 665 & 61 & 189 & 96 \end{array} \tag{7}$$

# Stable adult state – simple example case



Matrix $O$ lists the conditional probabilities of the outcomes (columns: Verbs) given cues (rows: Agents & Patients), derived from $M$ and $N$:

$$O = \begin{array}{l} \\ AgentGroup \\ AgentIndividual \\ AgentNoAgent \\ \ldots \end{array} \left\{ \begin{array}{rrrr} ajatella & harkita & miettia & pohtia \\ 0.07 & 0.12 & 0.07 & 0.23 \\ 0.23 & 0.04 & 0.14 & 0.08 \\ 0.23 & 0.07 & 0.08 & 0.12 \\ \ldots & \ldots & \ldots & \ldots \end{array} \right\} \quad (8)$$

# Stable adult state – simple example case



Estimated matrix $W$ representing the associations of Outcomes given Cues:

$$W = \left\{ \begin{array}{lrrrr} & \textit{pohtia} & \textit{harkita} & \textit{miettia} & \textit{ajatella} \\ \textit{AgentIndividual} & 0.10 & 0.07 & 0.21 & 0.38 \\ \color{red}{\textit{AgentGroup}} & \color{red}{0.37} & \color{red}{0.13} & \color{red}{0.06} & \color{red}{0.20} \\ \textit{AgentNoAgent} & 0.18 & 0.08 & 0.11 & 0.40 \\ \color{red}{\textit{PatientAbstraction}} & \color{red}{0.22} & \color{red}{0.00} & \color{red}{0.11} & \color{red}{-0.10} \\ \textit{PatientActivity} & 0.07 & 0.35 & 0.00 & -0.19 \\ \textit{PatientDirectQuote} & 0.49 & -0.07 & 0.17 & -0.36 \\ \textit{PatientIndirectQuestion} & 0.16 & -0.02 & 0.37 & -0.28 \\ \textit{Patientetta.CLAUSE} & -0.06 & -0.05 & -0.07 & 0.42 \\ \textit{PatientCommunication} & 0.11 & 0.09 & 0.27 & -0.24 \\ \textit{PatientInfinitive} & -0.11 & 0.00 & -0.19 & 0.53 \\ \textit{PatientNoAgent} & -0.04 & -0.01 & 0.01 & 0.28 \\ \textit{PatientIndividualGroup} & -0.09 & -0.01 & -0.08 & 0.42 \\ \textit{PatientParticiple} & -0.10 & -0.01 & -0.18 & 0.52 \end{array} \right\} \tag{9}$$

# Stable adult state – simple example case



- Support for any one of the four near-synonymous outcome alternatives given a set of active cues apparent in a context is obtained by summation of the respective weights ($W$)
- The corresponding probabilities for each outcome are calculated by dividing each outcome-specific support by the sum total of support for all outcomes with the contextual cues in question.

$$P(pohtia|\{AgentGroup, PatientAbstraction\})$$
$$= (0.37 + 0.22)/(0.37 + 0.13 + 0.06 + 0.20 + 0.22 + 0 + 0.11 - 0.10)$$
$$= 0.596$$

$$P(harkita|\{AgentGroup, PatientAbstraction\})$$
$$= (0.13 + 0)/(0.37 + 0.13 + 0.06 + 0.20 + 0.22 + 0 + 0.11 - 0.10) = 0.131$$

$$P(mietti|\{AgentGroup, PatientAbstraction\})$$
$$= (0.06 + 0.11)/(0.37 + 0.13 + 0.06 + 0.20 + 0.22 + 0 + 0.11 - 0.10) = 0.172$$

$$P(ajatella|\{AgentGroup, PatientAbstraction\})$$
$$= (0.20 - 0.10)/(0.37 + 0.13 + 0.06 + 0.20 + 0.22 + 0 + 0.11 - 0.10) = 0.101$$

# Representation of the stable end state

So – as simple as that?

$$CW = O \tag{10}$$

That the solution to the stable end state can be represented with a matrix equation does not mean that what our brains are calculating matrix algebra – an incremental learning process according to the Rescorla-Wagner equations simply results in an accumulated, consolidated body of knowledge which happens to be representable with a matrix notation!

# Representation of the stable end state



So – as simple as that?

$$CW = O \tag{10}$$

That the solution to the stable end state can be represented with a matrix equation does not mean that what our brains are calculating matrix algebra – an incremental learning process according to the Rescorla-Wagner equations simply results in an accumulated, consolidated body of knowledge which happens to be representable with a matrix notation!

# Comparison of statistical methods – Classification Accuracy & Recall

| | $\lambda_{prediction}$ | $\tau_{classification}$ | Accuracy |
|---|---|---|---|
| Polytomous logistic regression (One-vs-rest) | 0.368 | 0.488 | 0.645 |
| Polytomous mixed logistic regression (Poisson reformulation) | | | |
| 1\|Section | 0.360 | 0.482 | 0.640 |
| 1\|Author | 0.358 | 0.481 | 0.640 |
| 1\|Section + 1\|Author | 0.358 | 0.481 | 0.640 |
| Support Vector Machine | 0.340 | 0.466 | 0.629 |
| Memory-Based Learning (TiMBL) | 0.286 | 0.422 | 0.599 |
| Random Forests | 0.326 | 0.455 | 0.621 |
| Naive Discriminative Learning | 0.346 | 0.471 | 0.632 |

**Table:** Classification diagnostics for five models fitted to the Finnish data set ($n = 3404$).

# Cross-validation of statistical methods

| | PLR | SVM | TiMBL | NDL |
|---|---|---|---|---|
| Mean | 0.630 | 0.629 | 0.597 | 0.586 |
| 1 | 0.639 | 0.622 | 0.584 | 0.592 |
| 2 | 0.691 | 0.674 | 0.621 | 0.624 |
| 3 | 0.572 | 0.572 | 0.575 | 0.569 |
| 4 | 0.581 | 0.575 | 0.554 | 0.557 |
| 5 | 0.575 | 0.581 | 0.589 | 0.554 |
| 6 | 0.638 | 0.641 | 0.621 | 0.626 |
| 7 | 0.676 | 0.688 | 0.624 | 0.591 |
| 8 | 0.662 | 0.662 | 0.609 | 0.588 |
| 9 | 0.621 | 0.635 | 0.579 | 0.565 |
| 10 | 0.641 | 0.641 | 0.612 | 0.591 |

**Table:** Results of 10-fold cross-validation of four methods using the Finnish data set ($n = 3404$).

N.B. The Machine-learning methods were applied using their standard settings.

# Overview of underlying model complexity

- Polytomous mixed logistic regression: 189-190 coefficients
  - 4 outcomes X (46 coefficients + Intercept) + 1-2 random effects
- Support vector machine: parameter-free(*)
  - (2578 support vectors)
- Random forest: parameter-free(*)
- Memory-based learning (TiMBL): parameter-free(*)
  - (3404 exemplars)
- Naive discriminative learning: parameter-free(*)
  - 4 outcomes X 68 binary cue occurrence values
  - = 272 association weights

(*) *Parameter-free* in the sense that the method does not presuppose some predefined model with specified parameters/coefficients that are to be estimated.

# Overview of underlying model complexity

- Polytomous mixed logistic regression: 189-190 coefficients
  - 4 outcomes X (46 coefficients + Intercept) + 1-2 random effects
- Support vector machine: parameter-free$^{(*)}$
  - (2578 support vectors)
- Random forest: parameter-free$^{(*)}$
- Memory-based learning (TiMBL): parameter-free$^{(*)}$
  - (3404 exemplars)
- Naive discriminative learning: parameter-free$^{(*)}$
  - 4 outcomes X 68 binary cue occurrence values
  - = 272 association weights

$^{(*)}$ *Parameter-free* in the sense that the method does not presuppose some predefined model with specified parameters/coefficients that are to be estimated.

# Overview of underlying model complexity





- Polytomous mixed logistic regression: 189-190 coefficients
  - 4 outcomes X (46 coefficients + Intercept) + 1-2 random effects
- Support vector machine: parameter-free$^{(*)}$
  - (2578 support vectors)
- Random forest: parameter-free$^{(*)}$
- Memory-based learning (TiMBL): parameter-free$^{(*)}$
  - (3404 exemplars)
- Naive discriminative learning: parameter-free$^{(*)}$
  - 4 outcomes X 68 binary cue occurrence values
  - = 272 association weights

$^{(*)}$ *Parameter-free* in the sense that the method does not presuppose some predefined model with specified parameters/coefficients that are to be estimated.

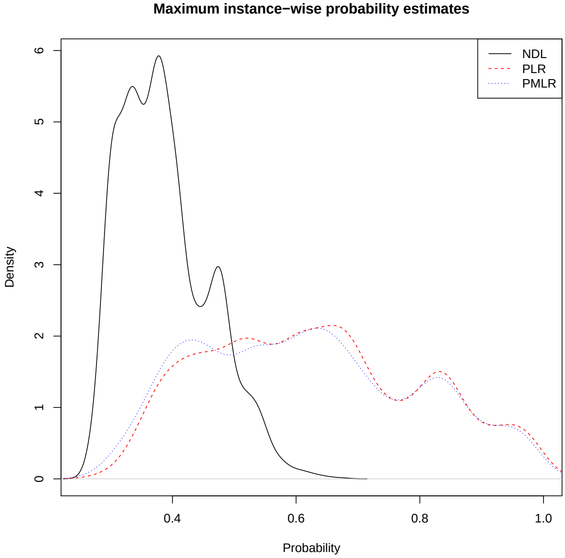
# Predicted outcomes

| | PLR | PMLR | SVM | TiMBL | NDL |
|---|---|---|---|---|---|
| PLR | 1.000 | 0.965 | 0.857 | 0.728 | 0.948 |
| PMLR | 0.965 | 1.000 | 0.872 | 0.731 | 0.938 |
| SVM | 0.857 | 0.872 | 1.000 | 0.740 | 0.874 |
| TiMBL | 0.728 | 0.731 | 0.740 | 1.000 | 0.726 |
| NDL | 0.948 | 0.938 | 0.874 | 0.726 | 1.000 |

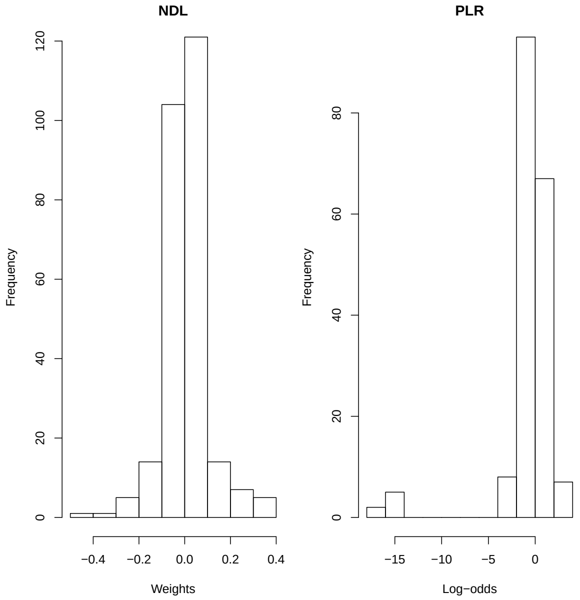
**Table:** Crosstabulation of predicted outcomes for five methods using the Finnish data set ($n = 3404$).

- The predictions of these five statistical methods appear to differ substantially more – implying they model contextual associations divergently
- Cf. different heuristics implementing PLR agreed 96.3%–98.7% of the time with the same data (Arppe 2008)

# Estimated probabilities



# NDL weights vs. PLR log-odds

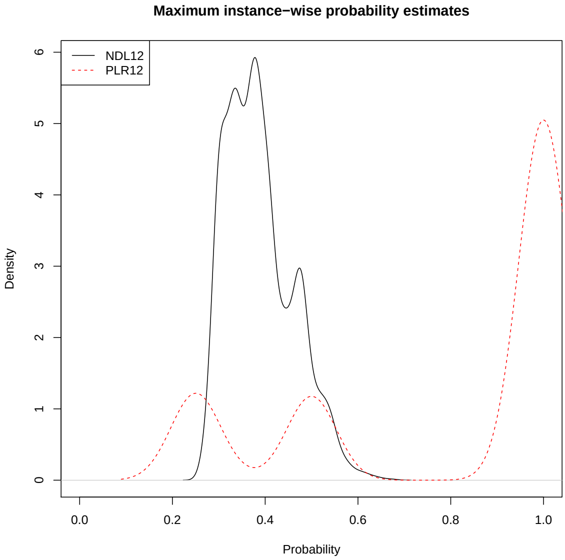


# Adding feature pair co-occurrences to the models

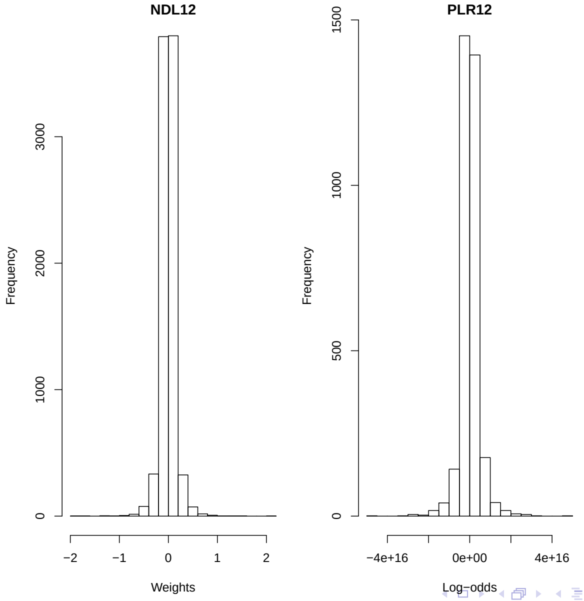
| | $\lambda_{\text{prediction}}$ | $\tau_{\text{classification}}$ | Accuracy |
|---|---|---|---|
| Polytomous logistic regression | | | |
| (single predictors) | 0.368 | 0.488 | 0.645 |
| (+ pairwise interactions) | 0.438 | 0.545 | 0.684 |
| Naive Discriminative Learning | | | |
| (single cues) | 0.346 | 0.471 | 0.632 |
| (+ cue pairs) | 0.569 | 0.651 | 0.758 |

**Table:** Classification diagnostics for four models fitted to the Finnish data set ($n = 3404$) [cf. QITL2]

# Estimated probabilities – feature pair co-occurrences



# NDL weights vs. PLR log-odds – feature pair co-occurrences

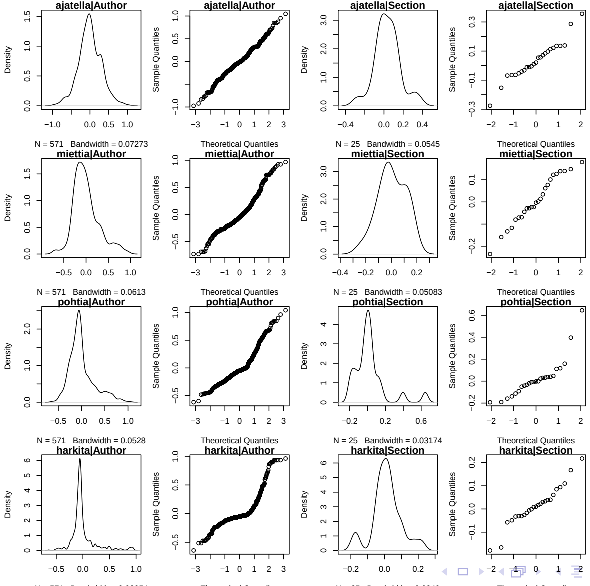


# ndl-Package

```
> library(ndl)
> data(think)

> think.ndl <- ndlClassify(Lexeme ~ Agent * Patient
  + Section, data = think)

> ndlStatistics(think.ndl)$accuracy
[1] 0.6113396

> ndlStatistics(think.ndl)$crosstable
           ajatella harkita miettia pohtia
  ajatella     1263      59     115     55
  harkita       107     182      32     66
  miettia       306      58     305    143
  pohtia        180      84     118    331
```

# NDL and random effects

Whereas `lmer` detects no random effect for *Author* ($n = 571$, *Variance* $= 0$), NDL is able to extract some outcome-variant impact:



# Discussion – Implications

- In overall absorption of knowledge, human learning builds a representation of past experience that is comparable to that of machine-learning techniques – works well with new but familiar input.
- Cross-validation results indicate that human learning performs somewhat less well than machine-learning techniques for unseen, new data – at least initially, though the incremental learning process should soon absorb this new information, too.
- Human learning would appear to overfit accumulated information, but does so in a substantially more well-behaved, robust manner than machine-learning (reflected in the modest association weights).

# Implications

- Timid probability estimates suggest that human learning is more open to variation in terms of its internal representation – speakers are more likely to produce alternative forms (due to noise and whatever confounding factors), as they are not attempting to maximize a likelihood over all accumulated experience.
- Human learning becomes well attuned to familiar patterns (idiosyncracies of often-met people, local dialects, and professional jargon), but is at first at a loss with new, unfamilar patterns, though will quickly adapt to this new information.

# Conclusions

- Naive Discriminative Learning implements the simplest possible mathematical characterization of probabilistic linguistic competence.
- This is compatible with the insight that grammar is usage-based.
- Importantly, usage is acquired piecemeal in a much simpler weight space – new information is integrated immediately by adjusting the accumulated body of knowledge as it is experienced, and this new information is not independently retained.
- The model can get very close to the observed data - as if the speaker accommodates to his/her own linguistic environment (e.g. dialect) – at the expense of being able to use/understand a general norm.