UNIVERSITETET I TROMSØ UiT

Statistical Tools for Evaluating the Behavior of Rival Forms: Logistic Regression, Tree & Forest, and Naive Discriminative Learning

R. Harald Baayen

University of Tübingen/University of Alberta

Laura A. Janda

CLEAR-group (Cognitive Linguistics – Empirical Approaches to Russian)

University of Tromsø

Other contributors:

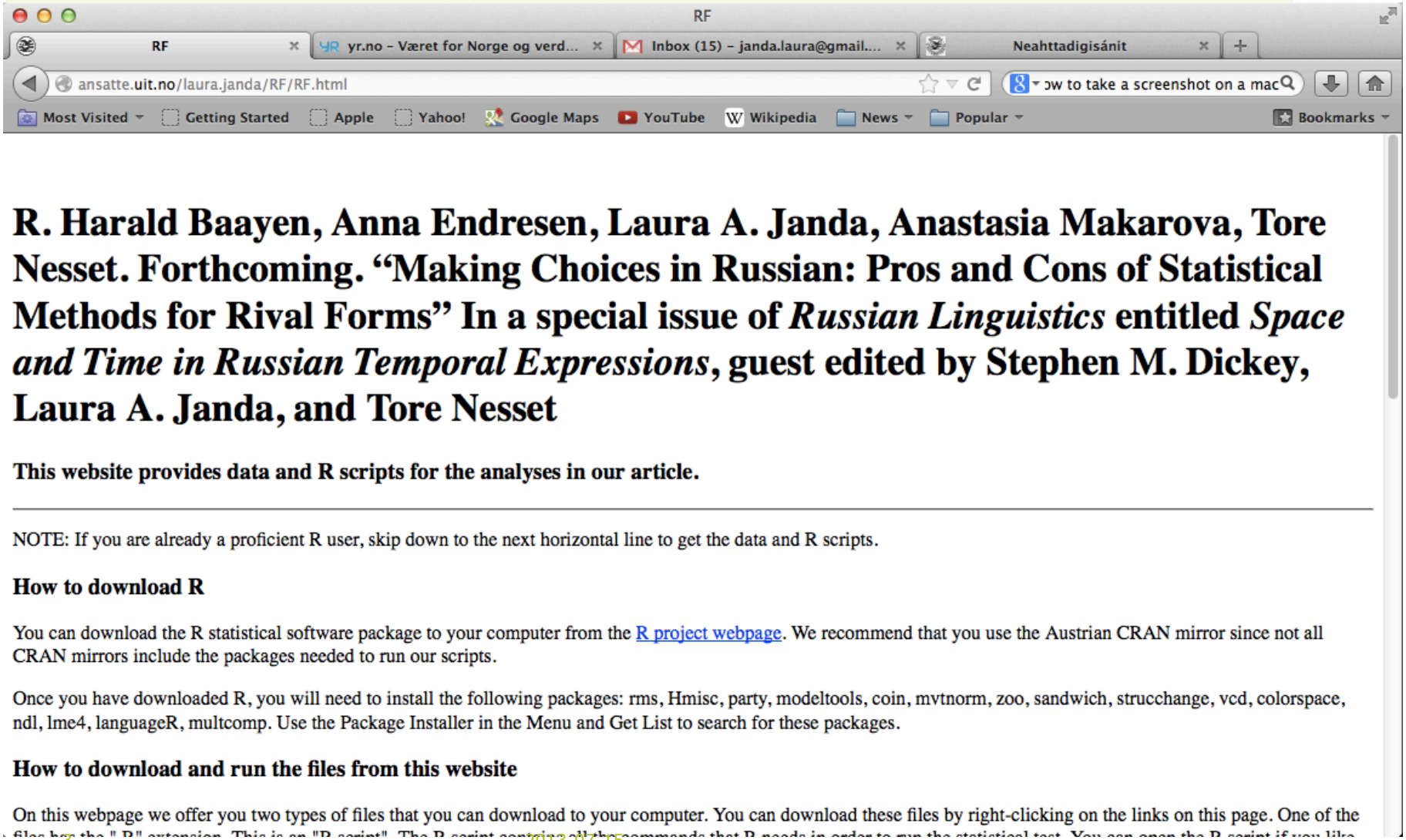
Tore Nesset

Svetlana Sokolova

Anna Endresen

Anastasia Makarova

All data and code are publicly available on this website:



Evaluating the behavior of rival forms

- **Rival forms:**
  - Languages often provide choices of **two or more forms**, let's call them X vs. Y, with (approximately) the **same meaning**
  - Linguists often want to know **what factors influence the choice of rival forms**
- **Traditional approach:**
  - Collect **corpus data** showing the **distribution of the rival forms** and the **other factors**
  - Build a **logistic regression** model with rival forms (form X vs. form Y) as dependent variable and other factors as independent (predictor) variables

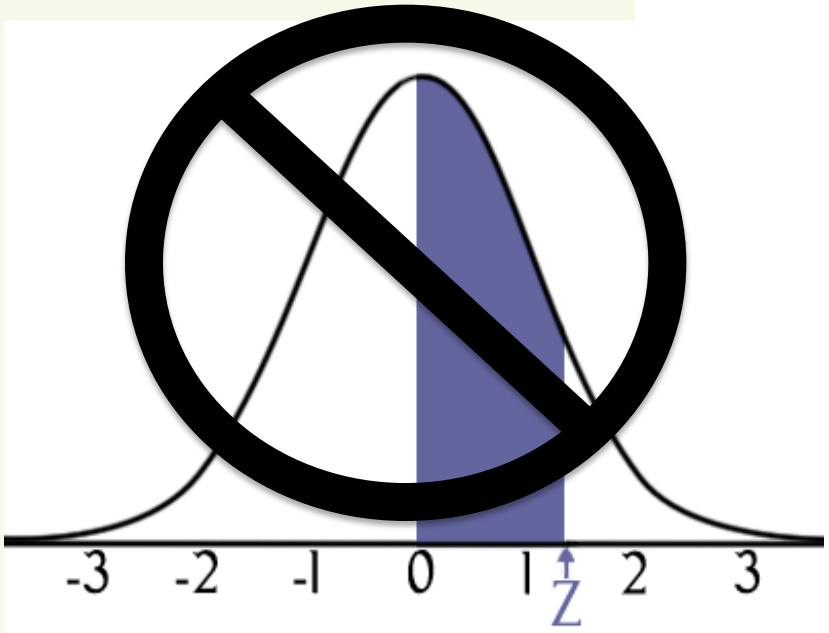
What a logistic regression analysis does

- A logistic regression model predicts the probability that a given value (form X, or alternatively, form Y) for the dependent variable will be used, taking into account the effects of the independent variables
- It can tell us **what the relationship is between the choice of rival forms and other factors**

logistic function

Drawbacks of logistic regression

- Assumes normal distribution (**parametric assumption**), which is usually **not the case with corpus data**
- Assumes that **all combinations** of all variables are represented in the dataset, but **linguistic data often involves paradigmatic gaps**



normal distribution

More drawbacks of logistic regression

- Building the optimal model can be a **laborious task**
- There is a **danger of overfitting** and no **validation** (after corpus data is collected there is usually no opportunity to collect an independent equivalent dataset)
- Results are a table and coefficients that can be **hard to interpret**

Alternatives to regression: tree & forest and naive discriminative learning

- These models:
  - are non-parametric (do not assume a normal distribution)
  - do not assume all combinations of variable levels are represented
  - do not require tedious building and fine-tuning
  - provide opportunities for validation
  - yield results that can be easier to interpret (especially trees)

Drawbacks of tree & forest and naive discriminative learning

- **tree & forest model cannot handle random effects** – mixed effects logistic models are needed
- **tests of significance** (p-values) are available for logistic regression and tree & forest, but **not for naive discriminative learning**
- from a cognitive perspective, **naive discriminative learning** makes sense only when **data represent a speaker's learning experience**

Comparing regression, tree & forest, and naive discriminative learning

- We have tested the performance of regression, tree & forest and naive discriminative learning across **four different datasets** and found:
  - The three models **perform nearly identically** in terms of **classification accuracy**, **indices of concordance** and **measurement of variable importance**

In other words, by most measures tree & forest and naive discriminative learning models perform just **as well as regression**

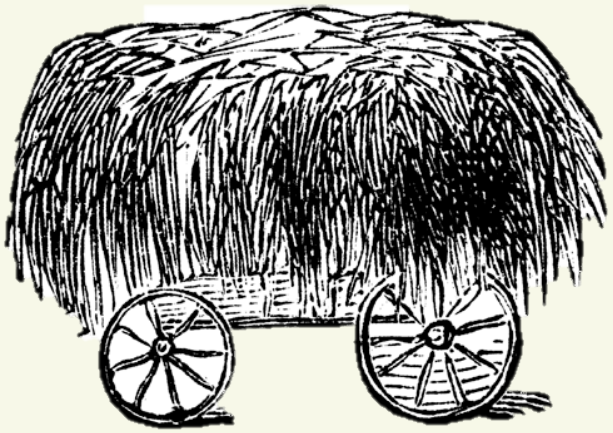
Our four datasets (all Russian)

- The Locative Alternation in Russian
- Distribution of prefixes *pere-* vs. *pre-*
- Distribution of prefixes *o-* vs. *ob-*
- Loss of *–nu* suffix in verbs

Different datasets may be better served by different models
See Baayen et al. forthcoming

The Locative Alternation in Russian

- **Theme**-object construction
  - *gruzit' **seno** na telegu*
  - [load **hay-ACC** onto wagon-ACC]
  - 'load **hay** onto the wagon'
- **Goal**-object construction
  - *gruzit' **telegu** senom*
  - [load **wagon-ACC** hay-INST]
  - 'load **the wagon** with hay'



- Variables: VERB (prefixes), (passive) PARTICIPLE, REDUCED
- VERB: unprefixed *gruzit'* or prefixed: *nagruzit', zagruzit', pogruzit'*
- PARTICIPLE:
  - **Theme**-object: ***seno*** *gruženo na telegu* '**hay** is loaded onto the wagon'
  - **Goal**-object: ***telega*** *gružena senom* '**the wagon** is loaded with hay'
- REDUCED:
  - **Theme**-object*: gruzit'* ***seno*** 'load **the hay**'
  - **Goal**-object: *gruzit'* ***telegu*** 'load **the wagon**'

The Locative Alternation in Russian

RIVAL FORMS: the two constructions, theme-object vs. goal-object

DEPENDENT VARIABLE:

CONSTRUCTION:

theme-object vs. goal-object

INDEPENDENT VARIABLES:

VERB:

zero (for the unprefixed verb *gruzit'*) vs. *na-* vs. *za-* vs. *po-*

PARTICIPLE:

yes vs. no

REDUCED:

yes vs. no

DATA: 1920 sentences from the Russian National Corpus

Logistic regression model
See Handout

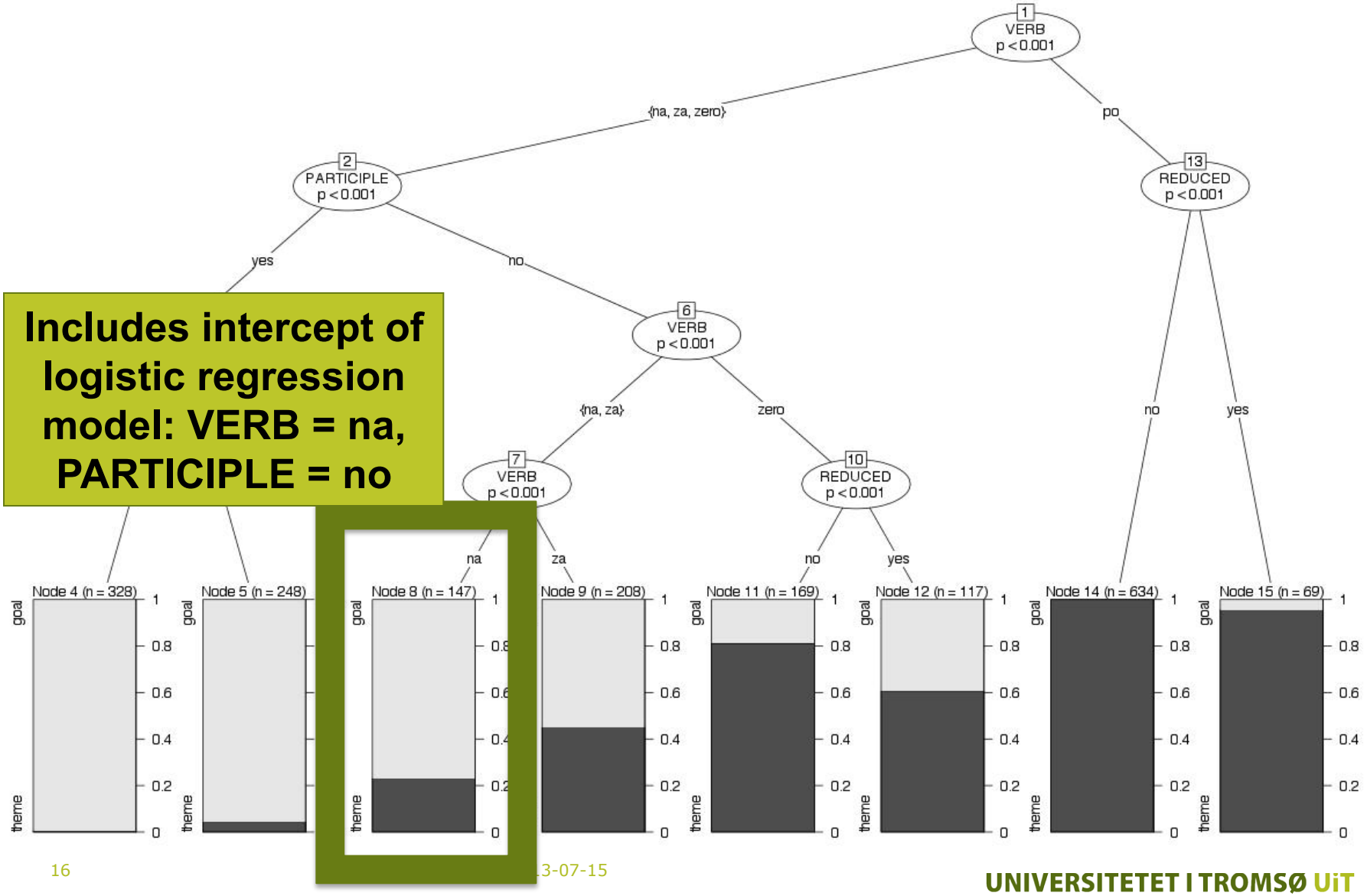
Optimal model: CONSTRUCTION~VERB +REDUCED+PARTICIPLE+VERB*PARTICIPLE

The model estimates how the log of the number of theme constructions divided by the log of the number of goal constructions depends on the predictors. The coefficient of the estimate (Coeff.) is POSITIVE if the combination of factors predicts more theme constructions, but NEGATIVE if they predict more goal constructions

Tree & forest

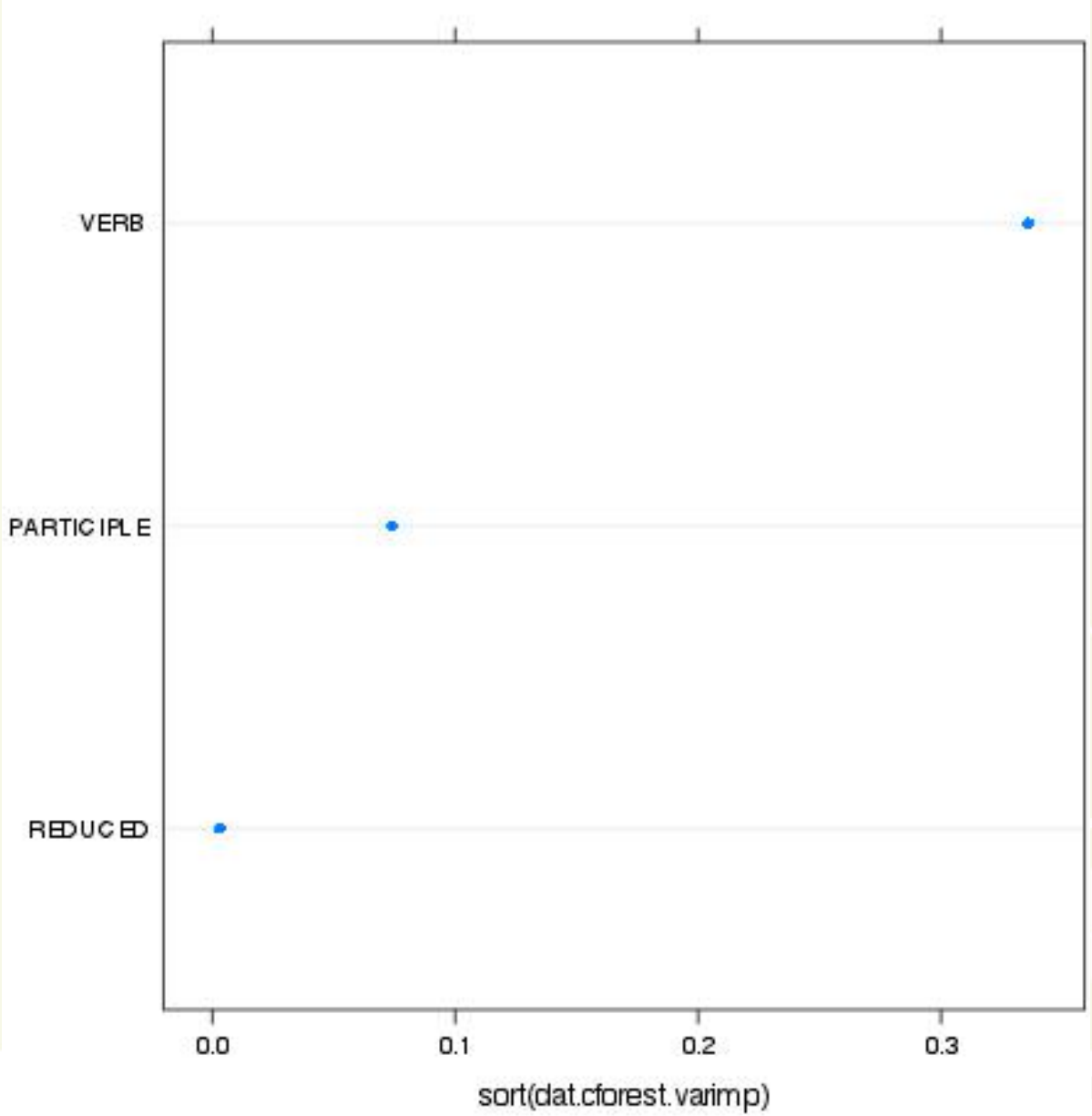
- **Classification and regression tree** (CART) uses recursive partitioning to yield a classification tree showing the best sorting of observations separating the values for the dependent variable
  - optimal algorithm for predicting an outcome given the predictor values
- **Random forest** uses repeated bootstrap samples drawn with replacement from the dataset such that in each repetition some observations are sampled and serve as a training set and other observations are not sampled, so they can serve for validation of the model
  - predictor variables are also randomly removed from repetitions, making it possible to measure variable importance

Tree & forest: CART



Includes intercept of logistic regression model: VERB = na, PARTICIPLE = no

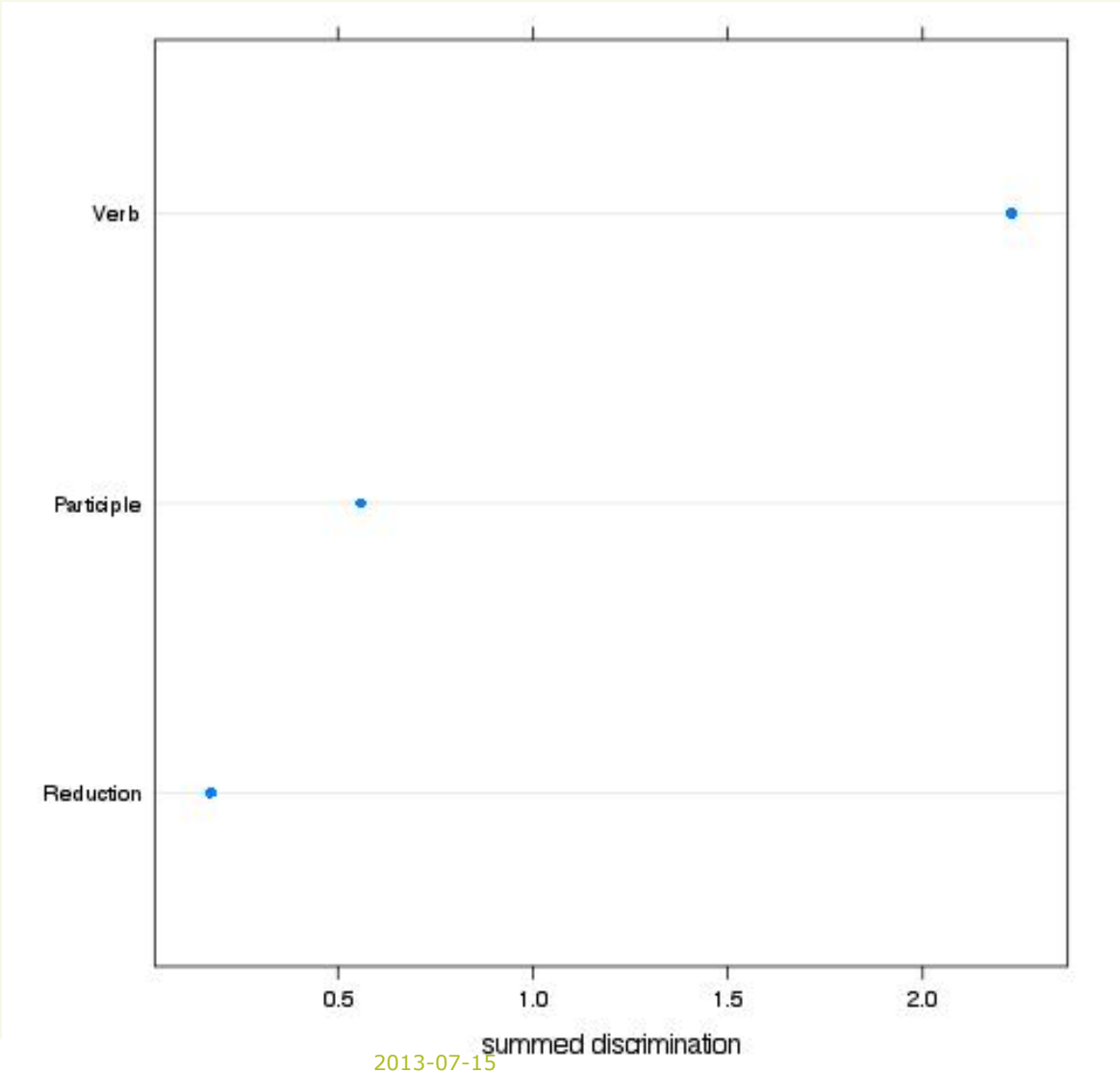
Tree & forest: variable importance



Naive discriminative learning

- Naive discriminative learning is a quantitative model for how choices are made, making use of a system of weights that are estimated using equilibrium equations
  - See Table 3 on handout
- Naive discriminative learning partitions the data into ten subsamples, nine of which serve as the training set, reserving the tenth one to serve for validation
  - This process is repeated ten times so that each subsample is used for validation.

Naive discriminative learning: variable importance



Comparison of performance

- Logistic regression
  - C = index of concordance = 0.96
  - Accuracy = 89%
- Tree & forest
  - C = index of concordance = 0.96
  - Accuracy = 89%
- Naive discriminative learning
  - C = index of concordance = 0.96
  - Accuracy = 88%

Conclusions

- Logistic regression, tree & forest and naive discriminative learning give very similar results in terms of overall performance and assessment of factors
- In addition to the *gruzit'* 'load' dataset, we have tested these tools on three other datasets and found remarkable convergence between the models
- Tree & forest and naive discriminative learning models can be used reliably to complement logistic regression
- These tools can empower cognitive linguists to find the structure in rival forms data and consider how these patterns are learned