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

R. Harald Baayen & Laura A. Janda
University of Tübingen/University of Alberta & University of Tromsø

We wish to expand the repertoire of statistical tools for usage-based analysis of linguistic data. Our data and R-scripts will be available on a website so attendees may apply these techniques to similar data.

Languages often provide choices among rival forms. For example, in Russian and English we have:

a) theme-object construction
Boris gruzil$_{VERB}$ *jaščiki*$_{THEME}$ *na telegu*$_{GOAL}$
Boris loaded$_{VERB}$ *the boxes*$_{THEME}$ *onto the wagon*$_{GOAL}$
b) goal-object construction
Boris gruzil$_{VERB}$ *telegu*$_{GOAL}$ *jaščikami*$_{THEME}$
Boris loaded$_{VERB}$ *the wagon*$_{GOAL}$ *with boxes*$_{THEME}$.

Linguists often want to explore how various factors contribute to the choice of rival forms based on corpus distributions.

We compare the performance of statistical tools on a dataset representing the distribution of CONSTRUCTION (theme-object vs. goal-object) for 1920 examples of Russian *gruzit'* 'load' in the Russian National Corpus (www.ruscorpora.ru; Sokolova 2012). This dataset represents three factors: 1) VERB: the form of the verb (unprefixed or prefixed in *na-, za-,* or *po-*); 2) PARTICIPLE: either active or passive voice, as in *Irina*$_{GOAL}$ *šla nagružennaja*$_{VERB}$ *sumkami*$_{THEME}$ 'Irina$_{GOAL}$ walked, loaded$_{VERB}$ with bags$_{THEME}$'; 3) REDUCED: either both theme and goal present, or one omitted, as in *Mužiki gruzili*$_{VERB}$ *les*$_{THEME}$ 'The men loaded$_{VERB}$ timber$_{THEME}$'.

Logistic regression, the traditional way to analyze a choice between rival forms, has drawbacks: it makes assumptions about the distribution of the data, the analyst must go through a tedious trial-and-error process to discover the optimal model (cf. Gries 2009), and training is needed to interpret the results.

The "tree & forest" (classification trees and random forests, cf. Strobl et al. 2009) model is more straightforward to use and provides an intuitive diagram of the outcomes yielded by various combinations of predictor values, along with measures of the relative strength of factors. This alternative avoids assumptions about the distribution of the data and eliminates the search for an optimal model. In addition, this model can provide cross-validation by repeatedly partitioning the data.

A third tool, "naive discriminative learning" (ndl; Baayen 2011), shares many of the advantages of the tree & forest model, but instead provides a quantitative model for how the brain chooses between rival forms, using estimated weights and equilibrium equations based on extensive results from animal and human learning (Danks 2003, Wagner & Rescorla 1972). This is not an alternative statistical technique, but a way of explaining how learners might acquire usage of rival forms based on distributional patterns. Ndl addresses the cognitive foundation of our linguistic probabilistic knowledge.

We show that logistic regression, tree & forest and ndl 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. This means that tree & forest and ndl 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.