Corpus linguistics and naive discriminative learning

R. Harald Baayen
University of Alberta

Abstract

Three classifiers from machine learning (the generalized linear mixed model, memory based learning, and support vector machines) are compared with a naive discriminative learning classifier, derived from basic principles of error-driven learning characterizing animal and human learning. Tested on the dative alternation in English, using the Switchboard data from (Bresnan, Cueni, Nikitina, & Baayen, 2007), naive discriminative learning emerges with state-of-the-art predictive accuracy. Naive discriminative learning offers a unified framework for understanding the learning of probabilistic distributional patterns, for classification, and for a cognitive grounding of distinctive collexeme analysis.

According to Gries (2011), linguistics is a distributional science exploring the distribution of elements at all levels of linguistic structure. He describes corpus linguistics as investigating the frequencies of occurrence of such elements in corpora, their dispersion, and their co-occurrence properties. Although this characterization of present-day corpus linguistics is factually correct, the aim of the present paper is to argue that corpus linguistics should be more ambitious, and that for a better understanding of the data its current descriptive approach may profit from complementation with cognitive computational modeling.

Consider the dative alternation in English. Bresnan et al. (2007) presented an analysis of the dative alternation in which the choice between the double object construction (*Mary gave John the book*) and the prepositional object construction (*Mary gave the book to John*) was modeled as a function of a wide range of predictors, including the accessibility, definiteness, length, and animacy of theme and recepient (see also Ford & Bresnan, 2010). A mixed-effects logistic regression model indicated that their variables were highly succesful in predicting which construction is most likely to be used, with approximately 94% accuracy.

The statistical technique used by Bresnan and colleagues, logistic regression modeling, is but one of many excellent statistical classifiers currently available to the corpus linguist, such as memory based learning (MBL, Daelemans & Bosch, 2005), analogical modeling of language (AML, Skousen, 1989), support vector machines

(SVM, Vapnik, 1995), and random forests (RF, Strobl, Malley, & Tutz, 2009; Tagliamonte & Baayen, 2010). The mathematics underlying these techniques varies widely, from iterative optimization of the model fit (regression), nearest-neighbor similarity-based inference (memory based learning), kernel methods (support vector machines), and recursive conditioning with subsampling (random forests). All these statistical techniques tend to provide a good description of the speaker-listener's knowledge, but it is unlikely that they provide a good characterization of how speaker-listeners actually acquire and use this knowledge. Of these four techniques, only memory-based learning, as a computational implementation of an exemplar-based model, may arguably reflect human performance.

A first question addressed in the present study is whether these different statistical models provide a correct characterization of the knowledge that a speaker has of how to choose between these two dative constructions. A statistical model may faithfully reflect a speaker's knowledge, but it is also conceivable that it underestimates or overestimates what native speakers of English actually have internalized. This question will be addressed by comparing statistical models with a model based on principles of human learning.

A second question concerns how frequency of occurrence and co-occurrence frequencies come into play in human classification behavior as compared to machine classification. For machine classification, we can easily count how often a linguistic element occurs, and how often it co-occurs with other elements. The success of machine classification in reproducing linguistic choice behavior suggests that probabilities of occurrence are somehow available to the human classifier. But is frequency of (co-)occurrence available to the human classifier in the same way as to the machine classifier? Simple frequency of occurrence information is often modeled by means of some 'counter in the head', implemented in cognitive models in the form of 'resting activation levels', as in the interactive activation models of McClelland and Rumelhart (1981); Coltheart, Rastle, Perry, Langdon, and Ziegler (2001); Van Heuven, Dijkstra, and Grainger (1998), in the form of frequency based rankings (Murray & Forster, 2004), as a unit's verification time (Levelt, Roelofs, & Meyer, 1999), or in the Bayesian approach of Norris, straightforwardly as a unit's long-term a-priori probability (Norris, 2006; Norris & McQueen, 2008). A potential problem that arises in this context is that large numbers of such 'counters in the head' are required, not only for simple or complex words, but also for hundreds of millions of word n-grams, given recent experimental results indicating human sensitivity to n-gram frequency (Arnon & Snider, 2010; Tremblay & Baayen, 2010). Moreover, given the tendency of human memory to merge, or blend, previous experiences, it is rather unlikely that the human classifier has at its disposal exactly the same frequency information that we make available to our machine classifiers.

To address these questions, the present study explores what a general model of human learning may offer corpus linguistics as a computational theory of human classification.

Frequency	Definiteness of Theme	Pronominality of Theme	Construction
7	definite	non-pronominal	NP NP
1	definite	pronominal	NP NP
28	indefinite	non-pronominal	NP NP
1	indefinite	pronominal	NP NP
3	definite	non-pronominal	NP PP
4	definite	pronominal	NP PP
6	indefinite	non-pronominal	NP PP
0	indefinite	pronominal	NP PP

Table 1: Example instance base for discriminative learning with the Rescorla-Wagner equations, with as cues the definiteness and pronominality of the theme, and as outcome the construction (double object, NP NP, versus prepositional object, NP PP).

Naive Discriminative Learning

In psychology, the model of Wagner and Rescorla (1972) is one of the most influential and fruitful theories of animal and human learning (Miller, Barnet, & Grahame, 1995; Siegel & Allan, 1996). Its learning algorithm is closely related to the connectionist delta-rule (cf. Gluck & Bower, 1988; Anderson, 2000) and to the Kalman filter (cf. Dayan & Kakade, 2001), and can be viewed as an instantiation of a general probabilistic learning mechanism (see, e.g., Chater, Tenenbaum, & Yuille, 2006; Hsu, Chater, & Vitányi, 2010).

The Rescorla-Wagner equations

Rescorla and Wagner formulated a set of equations that specify how the strength of association of a cue in the input to a given outcome is modified by experience. By way of example, consider the instance base in Table 1, which specifies for the four combinations of the pronominality and definiteness of the theme (*the book* in *John gave the book to Mary*) which construction is used (the double object construction, NP NP, or the prepositional object construction, NP PP). The eight possible combinations occur with different frequencies, modeled on the data of Bresnan et al. (2007). The cues in this example are the values for definiteness and pronominality. The outcomes are the two constructions. There are in all 50 learning trials, more than half of which pair an indefinite non-pronominal theme with the double object construction (e.g., *John gave a book to Mary*).

The Rescorla-Wagner equations implement a form of supervised learning. It is assumed that the learner predicts an outcome given the available cues. Depending on whether this prediction is correct, the weights (association strengths) from the cues to the outcomes are adjusted such that at subsequent trials, prediction accuracy will improve.

Let $\textsc{present}(C,t)$ denote the presence of a cue C (definiteness, pronominality) and $\textsc{present}(O,t)$ the presence of outcome O (construction) at time t, and let $\textsc{absent}(C,t)$ and $\textsc{absent}(O,t)$ denote their 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 $t+1$ by means of the recurrence relation

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

which simply states that the association strength at time $t+1$ is equal to its previous association strength at time t modified by some change change in association strength ΔV_i^t, defined as

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

Standard settings for the parameters are $\lambda = 1$, $\alpha_1 = \alpha_2 = 0.1$, $\beta_1 = \beta_2 = 0.1$. If a cue is not present in the input, its association strength is not changed. When the cue is present, the change in association strength depends on whether or not the outcome is present. Association strengths are increased when cue and outcome co-occur, and decreased when the cue occurs without the outcome. Furthermore, when more cues are present simultaneously, adjustments are more conservative. In this case, we can speak of cue competition.

Figure 1 illustrates, for a random presentation of the 50 learning trials, how the association strengths (or weights) from cues to outcomes develop over time. As indefinite nonpronominal themes dominate the instance base, and strongly favor the double object construction, the weights from the cues `indefinite` and `non-pronominal` to the construction `NP NP` increase steadily during the learning process.

The equilibrium equations for the Rescorla-Wagner equations

The Rescorla-Wagner equations have recently turned out to be of considerable interest for understanding child language acquisition, see, for instance, Ramscar and Yarlett (2007); Ramscar, Yarlett, Dye, Denny, and Thorpe (2010); Ramscar, Dye, Popick, and O'Donnell-McCarthy (2011). For corpus linguistics, the equilibrium equations for the Rescorla-Wagner equations developed by Danks (2003) are of key interest. Danks was able to derive a set of equations that define the association strengths (weights) from cues to outcomes for the situation in which these strengths no longer change, i.e., for the adult state of the learner. It can be shown that when

$$V_i^{t+1} = V_i^t, \text{ or, equivalently,} \tag{3}$$

$$V_i^{t+1} - V_i^t = 0, \tag{4}$$

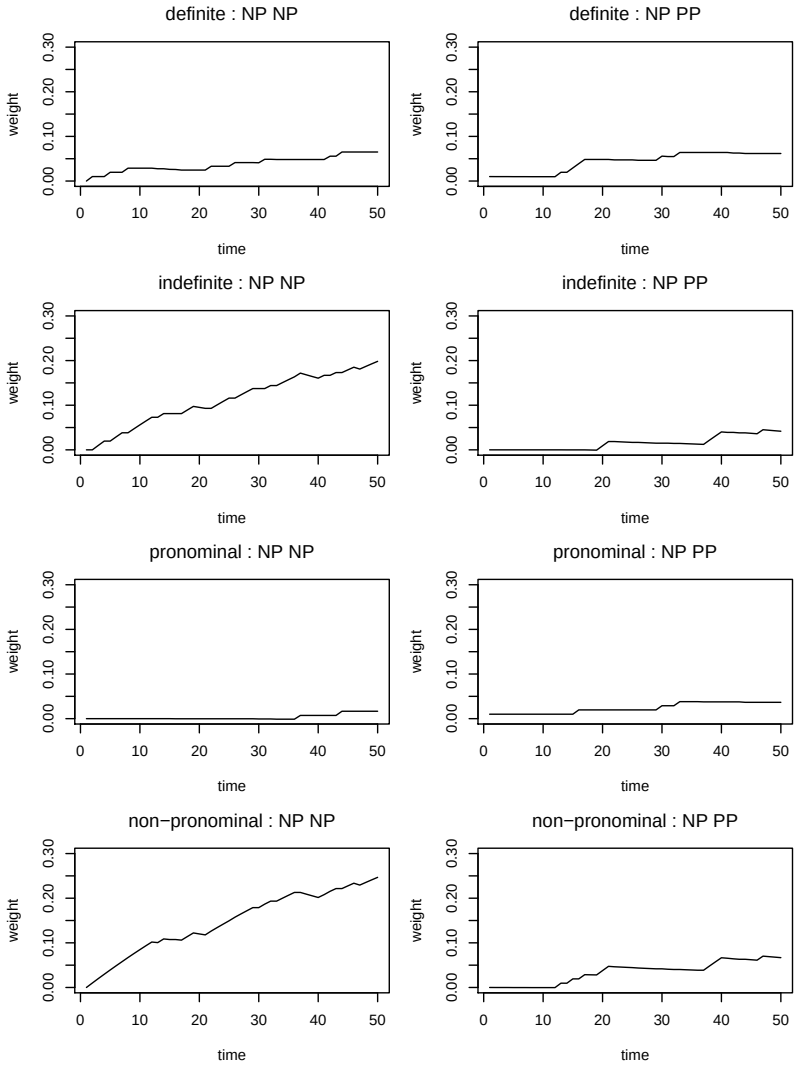

Figure 1. Development of the association strengths (weights) from cues (definite/indefinite/pronominal/non-pronominal) to outcomes (NP NP/NP PP) given the instance base summarized in Table 1. The 50 instance tokens were presented for learning once, in random order.

the weights to the outcomes can be estimated by solving the following set of equations, with $\boldsymbol{W}$ the matrix of unknown weights:[1]

$$\boldsymbol{CW} = \boldsymbol{O}. \tag{5}$$

[1]Equation (5) is formulated using notation from matrix algebra. The following example illustrates the principle of the calculations involved. $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} v & w \\ x & y \end{pmatrix} = \begin{pmatrix} av + bx & aw + by \\ cv + dx & bw + dy \end{pmatrix}$.

In (5), $\boldsymbol{C}$ is the matrix of conditional probabilities of the outcomes. It is obtained by first calculating the matrix $\boldsymbol{M}$ listing the frequencies with which cues co-occur:

$$\boldsymbol{M} = \begin{pmatrix} & \texttt{indefinite} & \texttt{pronominal} & \texttt{nonpronominal} & \texttt{definite} \\ \texttt{indefinite} & 35 & 1 & 34 & 0 \\ \texttt{pronominal} & 1 & 6 & 0 & 5 \\ \texttt{nonpronominal} & 34 & 0 & 44 & 10 \\ \texttt{definite} & 0 & 5 & 10 & 15 \end{pmatrix}. \tag{6}$$

As can be verified by inspecting Table 1, the cue `indefinite` occurs 35 times, the combination of `indefinite` and `pronominal` occurs once, `indefinite` co-occurs 34 times with `non-pronominal`, and so on. From this matrix, we derive the matrix of conditional probabilies of cue j given cue i:

$$\boldsymbol{C} = \begin{pmatrix} & \texttt{indefinite} & \texttt{pronominal} & \texttt{nonpronominal} & \texttt{definite} \\ \texttt{indefinite} & 0.50 & 0.01 & 0.49 & 0.00 \\ \texttt{pronominal} & 0.08 & 0.50 & 0.00 & 0.42 \\ \texttt{nonpronominal} & 0.39 & 0.00 & 0.50 & 0.11 \\ \texttt{definite} & 0.00 & 0.17 & 0.33 & 0.50 \end{pmatrix}. \tag{7}$$

The probability of `indefinite` given `indefinite` is $35/(35+1+34+0) = 0.5$, that of `indefinite` given `pronominal` is $1/(1+6+0+5) = 0.083$, and so on.

The matrix $\boldsymbol{W}$ is the matrix of association strengths (weights) from cues (rows) to outcomes (columns) that we want to estimate. Finally, the matrix $\boldsymbol{O}$,

$$\boldsymbol{O} = \begin{pmatrix} & \texttt{NP NP} & \texttt{NP PP} \\ \texttt{indefinite} & 0.41 & 0.09 \\ \texttt{pronominal} & 0.17 & 0.33 \\ \texttt{nonpronominal} & 0.40 & 0.10 \\ \texttt{definite} & 0.27 & 0.23 \end{pmatrix} \tag{8}$$

lists the conditional probabilities of the constructions (columns) given the cues (rows). It is obtained from the co-occurrence matrix of cues ($\boldsymbol{M}$) and the co-occurrence matrix of cues and constructions $\boldsymbol{N}$,

$$\boldsymbol{N} = \begin{pmatrix} & \texttt{NP NP} & \texttt{NP PP} \\ \texttt{indefinite} & 29 & 6 \\ \texttt{pronominal} & 2 & 4 \\ \texttt{nonpronominal} & 35 & 9 \\ \texttt{definite} & 8 & 7 \end{pmatrix}. \tag{9}$$

For instance, the probability of the double object construction given (i) the `indefinite` cue is $29/(35+1+34+0) = 0.414$, and given (ii) the `pronominal` cue it is $2/(1+6+0+5) = 0.167$. The set of equations (5) can be solved using the generalized

	indefinite non-pronominal	indefinite pronominal	definite non-pronominal	definite pronominal
NP NP	0.84	0.49	0.65	0.3
NP PP	0.16	0.51	0.35	0.7

Table 2: Probabilities of the two constructions following from the equilibrium equations for the Rescorla-Wagner model.

inverse, which will yield a solution that is optimal in the least-squares sense, resulting in the weight matrix

$$\boldsymbol{W} = \begin{pmatrix} & \texttt{NP NP} & \texttt{NP PP} \\ \texttt{indefinite} & 0.38 & 0.12 \\ \texttt{definite} & 0.19 & 0.31 \\ \texttt{nonpronominal} & 0.46 & 0.04 \\ \texttt{pronominal} & 0.11 & 0.39 \end{pmatrix}. \tag{10}$$

The support for the two constructions given a set of input cues is obtained by summation over the association strengths (weights) of the active cues in the input. For instance, for indefinite non-pronominal themes, the summed support for the NP NP construction is $0.38 + 0.46 = 0.84$, while the support for the NP PP construction is $0.12 + 0.04 = 0.16$. Hence, the probability of the double object construction equals $0.84/(0.84{+}0.16){=}\ 0.84$, and that for the prepositional object construction is 0.16. (In this example, the two measures of support sum up to one, but this is not generally the case for more complex data sets.) One can think of the weights being chosen in such a way that, given the co-occurrences of cues and outcomes, the probability of a construction given the different cues in the input is optimized.

We can view this model as providing a re-representation of the data: Eight frequencies (see Table 1) have been replaced by eight weights, representing 50 trials of learning. The model does not work with exemplars, nevertheless, its weights do reflect exemplar frequencies. For instance, the probabilities of the double object construction in Table 2 are correlated with the original frequencies ($r_s = 0.94, p = 0.051$). It is worth noting that the probabilities in Table 2 are obtained with a model that is completely driven by the input, and that is devoid of free parameters — the learning parameters of the Rescorla-Wagner equations (2) drop out of the equilibrium equations.

Baayen, Milin, Filipovic Durdjevic, Hendrix, and Marelli (2011) made use of discriminative learning to model visual lexical decision and self-paced reading latencies in Serbian and English. They obtained excellent fits to empirical latencies, both in terms of good correlations at the item level, as well as in terms of the relative importance and effect sizes of a wide range of lexical distributional predictors. Simulated latencies correctly reflected morphological family size effects as well as whole-word

frequency effects for complex words, without any complex words being represented in the model as individual units. Their model also predicts word n-gram frequency effects (see also Baayen & Hendrix, 2011). It provides a highly parsimonious account of morphological processing, both in terms of the representations it assumes, and in terms of the extremely limited number of free parameters that it requires to fit the data. For monomorphemic words, the model is essentially parameter free, as in the present example for the dative alternation.

Baayen et al. (2011) refer to the present approach as *naive* discriminative learning, because the probability of a given outcome is estimated independently from all other outcomes. This is a simplification, but thus far it seems that this simplification does not affect performance much, just as often observed for *naive Bayes* classifiers, while making it possible to obtain model predictions without having to simulate the learning process itself.

The question to which we now turn is to what extent naive discriminative learning provides a good fit to corpus data. If the model provides decent fits, then, given that it is grounded in well-established principles of human learning, and given that it performs well in simulations of human processing costs at the lexical level, we can compare discriminative learing with well-established statistical methods in order to answer the question of whether human learning is comparable, superior, or inferior to machine learning. We explore this issue by a more comprehensive analysis of the dative alternation data.

Predicting the dative alternation

From the `dative` dataset in the `languageR` package (Baayen, 2009), the subset of data points extracted from the Switchboard corpus were selected for further analysis. For this subset of the data, information about the speaker is available. In what follows, the probability of the prepositional object construction is taken as the response variable. Software for naive discriminative classification is available in the `ndl` package for R, available at `www.r-project.org`. Example code is provided in the appendix.

Prediction accuracy

A discriminative learning model predicting construction (double object versus prepositional object) was fitted with the predictors Verb, Semantic Class, and the Animacy, Definiteness, Pronominality, and Length of recipient and theme. As the model currently requires discrete cues, as a workaround, the length of recipient and theme were split into three ranges: length 1, lengths 2–4, and lengths exceeding 4. These three length levels were used as cues, instead of the original numerical values. As Bresnan et al. (2007) did not observe significant by-speaker variability, speaker is not included as a predictor in our initial model. (Models including speaker as predictor will be introduced below.)

To evaluate goodness of fit, we used two measures, the index of concordance C and the model's accuracy. The index of concordance C is also known as the receiver operating characteristic curve area 'C' (see, e.g. Harrell, 2001). Values of C exceeding 0.8 are generally regarded as indicative of a succesful classifier. Accuracy was defined here as the proportion of correctly predicted constructions, with as cut-off criterion for a correct prediction that the probability for the correct prediction exceed 0.5. According to these measures, the naive discriminative learning model performed well, with $C = 0.97$ and an accuracy of 0.92.

To place the performance of naive discriminative learning (NDL) in perspective, we compared it with memory based learning (MBL), logistic mixed-effects regression (GLMM), and a support vector machine with a linear kernel (SVM). The index of concordance obtained with MBL, using TiMBL version 6.3 (Daelemans, Zavrel, Sloot, & Bosch, 2010), was C =0.89. Its accuracy was 0.92. TiMBL was supplied with speaker information.

A logistic mixed-effects regression model, fitted with the LME4 package for R (D. Bates & Maechler, 2009), with both Speaker and Verb as random-effect factors did not converge. As the GLMM did not detect significant speaker-bound variance, we therefore fitted a model with verb as only random-effect factor, including length of theme and recipient as (numerical) covariates. The index of concordance for this model was $C = 0.97$, accuracy was at 0.93. The regression model required 18 parameters (one random-effect standard deviation, an intercept, and 16 coefficients for slopes and contrasts) to achieve this fit. A support vector machine, provided with access to Speaker information, and fitted with the `svm` function in the E1017 package for R (Dimitriadou, Hornik, Leisch, Meyer, & Weingessel, 2009), yielded $C = 0.97$ with accuracy at 0.93, requiring 524 support vectors.

From this comparison, naive discriminative learning emerges as more or less comparable in classificatory accuracy to existing state-of-the-art classifiers. It is outperformed in both C and accuracy only by the support vector machine, the currently best-performing classifier available. We note here that the NDL classifier used here is completely parameter-free. The weights are fully determined, and only determined, by the corpus input. There are no choices that the user could make to influence the results.

Since speaker information was available to TiMBL and to the SVM, we fitted a second naive discriminative learning model to the data, this time including speaker as a predictor. The index of concordance increased slightly to 0.98, and accuracy to 0.95. Further improvement can be obtained by by allowing pairs of predictor values to function as cues, following the naive discriminative reader model of Baayen et al. (2011). They included both letters and letter bigrams as cues, the former representing static knowledge of which letters are present in the input, the latter representing information about sequences of letters. Analogously, pairs of features, e.g., semantic class `p` combined with a `given theme`, can be brought into the learning process. This amounts to considering (when calculating the conditional co-occurrence matrix $\boldsymbol{C}$

	all data		10-fold cross-validation	
	C	Accuracy	C	Accuracy
SVM	0.98	0.95	0.95	0.91
TiMBL	0.89	0.92	0.89	0.92
GLMM	0.97	0.93	0.96	0.92
NDL (verb)	0.97	0.92	0.89	0.85
NDL (verb+speaker)	0.98	0.95	0.93	0.89
NDL-2 (verb+speaker)	0.99	0.96	0.94	0.91

Table 3: Index of concordance C and accuracy for all data (left) and average across 10-fold cross-validation.

not only pairwise co-occurrences of cues, but also the co-occurrences of triplets and quadruplets of cues. Within the framework of naive discriminative learning, this is the functional equivalent of interactions in a regression model. In what follows, NDL-2 refers to a model which includes pairs of features for all predictors, excluding however pairs involving Verb or Speaker. With this richer representation of the input, the index of concordance increased to 0.99 and accuracy to 0.96.

However, we now need to assess whether naive discriminative learning achieves this good performance at the cost of overfitting. To assess this possibility, we made use of 10-fold cross-validation, using exactly the same folds for each of the classifiers. The right half of Table 3 summarizes the results. In cross-validation, naive discriminative learning performs less well than the SVM and the GLMM, but similar to TiMBL. Fortunately, concordance and accuracy remain high.

We are now in the position to tentatively answer our first question, of whether machine learning outperforms human learning. If naive discriminative learning is indeed a reasonable approximation of human learning, then the answer is that human learning builds a representation of past experience comparable to that of other machine learning techniques. However, for generalization to unseen, new data, human classification seems thus far to be outperformed, albeit only slightly, by some of the best machine classifiers currently available.

Effect sizes and variable importance

One of the advantages of regression models for linguistic analysis is that the estimated coefficients offer the researcher insight into what forces shape the probabilities of a construction. For instance, a pronominal theme is assigned a β weight of 2.2398 on the log odds scale, indicating that pronominal themes are much more likely to be expressed in a prepositional object construction than in a double object construction. This kind of information is more difficult to extract from a support vector machine or from a memory based model, for which one has to inspect the support vectors or the similarity neighborhoods respectively. Interestingly, the weights of the

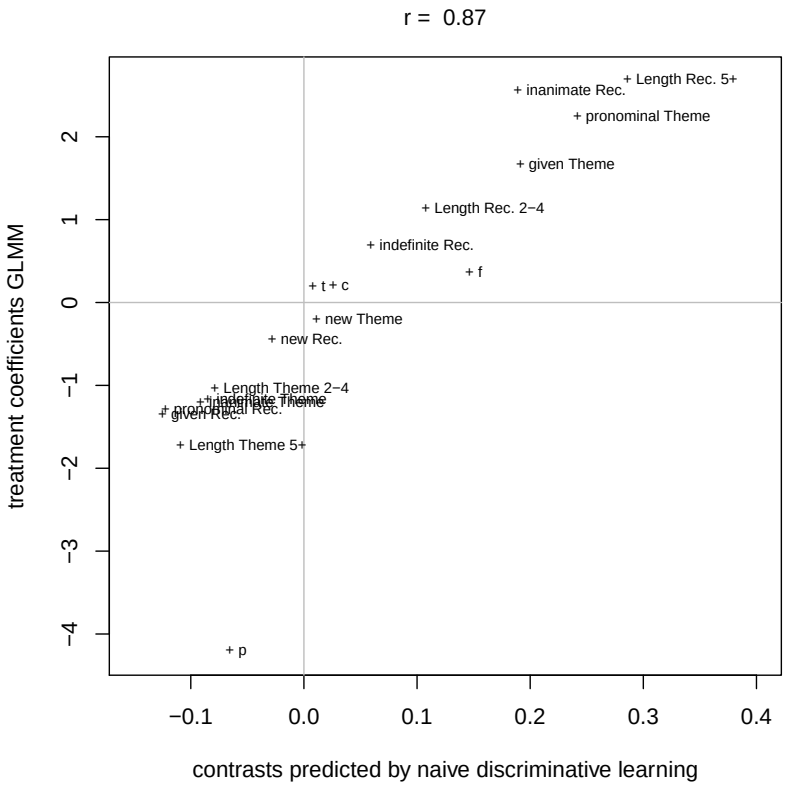

Figure 2. Treatment contrasts generated from the association strengths of the naive discriminative learner (horizontal axis) and the treatment contrasts of a generalized linear mixed-effects model (vertical axis). Semantic class: reference level `abstract` (*give it some thought*); `c`: communication (*tell, give me your name*); `f`: future transfer of possession (*owe, promise*); `p`: prevention of possession (*cost, deny*); `t`: transfer of possession (*give an armband, send*). Reference levels for the other predictors are `animate`, `definite`, `accessible`, `'non-pronominal`, and `Length 1`.

naive discriminative learner provide the same kind of information as the coefficients of the regression model. For instance, in the model with verb (and without speaker), a non-pronominal theme has a negative weight equal to -0.046 for the prepositional object construction, whereas a pronominal theme has a positive weight of 0.203. The difference between the two, henceforth the NDL treatment contrast, is 0.248. This difference should be similar to the treatment contrast for the pronominality of the theme, which is defined as the difference (on the logit scale) between a pronominal theme and the reference level of the non-pronominal theme. When we plot the NDL treatment contrast together with the treatment coefficients of the logistic regression model, we find that the two enter into a strong correlation, r = 0.87 ($t(16) = 7.18$, $p = 0$), as can be seen in Figure 2.

For sparse data, the naive discriminative learner tends to be more conservative

than the GLMM. The `+p` data point in the lower left of Figure 2 represents the 'prevention of possession' semantic class, which supports 182 instances with the double object construction and only one case with the prepositional object construction. The logistic regression model concludes that a prepositional object construction is extremely unlikely, assigning `+p` verbs a negative weight of no less than -4. The naive discriminative learner is assigning this type of verb a larger, though still small, probability.

In order to assess the importance of a predictor for classification accuracy across the very different classifiers considered above, we permute the values of the predictor in order to break its potential relation with the dependent variable. We then inspect to what extent classification accuracy decreases. The greater the decrease in classification accuracy, the greater the importance of the predictor. This non-parametric approach is inspired by how variable importance is assessed for random forests, which are also non-parametric classifiers (see, e.g., Strobl et al., 2009). Figure 3 summarizes the results for the regression model, the support vector machine, and for naive discriminative learning.

First consider variable importance for the regression model, summarized in the upper left panel. The pronominality of the theme emerges as the most important predictor for regression accuracy, followed by verb, and at a distance, by the definiteness of the theme. Semantic class has a negative score, indicating that random permutation of its values resulted in slightly improved accuracy. By chance, the random reordering of values resulted in a configuration that affords a slightly better model to be fitted. This may arise when the values of an irrelevant predictor are reshuffled. By mirroring the minimal score to the right of zero, we obtain an interval that characterizes irrelevant predictors (see, e.g., Strobl et al., 2009, for this logic in the context of random forests). For the regression model, this interval contains, in addition to Semantic Class, the predictors Animacy of Theme and Definiteness of Recipient.

The support vector machine comes to rather different conclusions. Its classification accuracy is very sensitive to having access to verb and speaker information. Accuracy is also afffected negatively by removal of the predictors specifying the pronominality of theme and recipient, as well as the length of the recipient.

Predictors marked as important by naive discriminative learning are the animacy of the recipient, the length of the theme, the pronominality of the theme, the identity of the verb, and the accessibility of the theme. Speaker is characterized as having no importance, in accordance with the GLMM but contrary to the results obtained with the SVM.

For all models, overall accuracy (which is in the nineties) is hardly affected by permuting the values of a single predictor. This especially striking for the naive discriminative learning model with cue pairs (lower right panel), for which the reductions in accuracy are an order of magnitude smaller than those for the other models (note the different scale on the horizontal axis in the lower right panel of Figure 3). Apparently, this model is exceptionally robust against noise predictors.

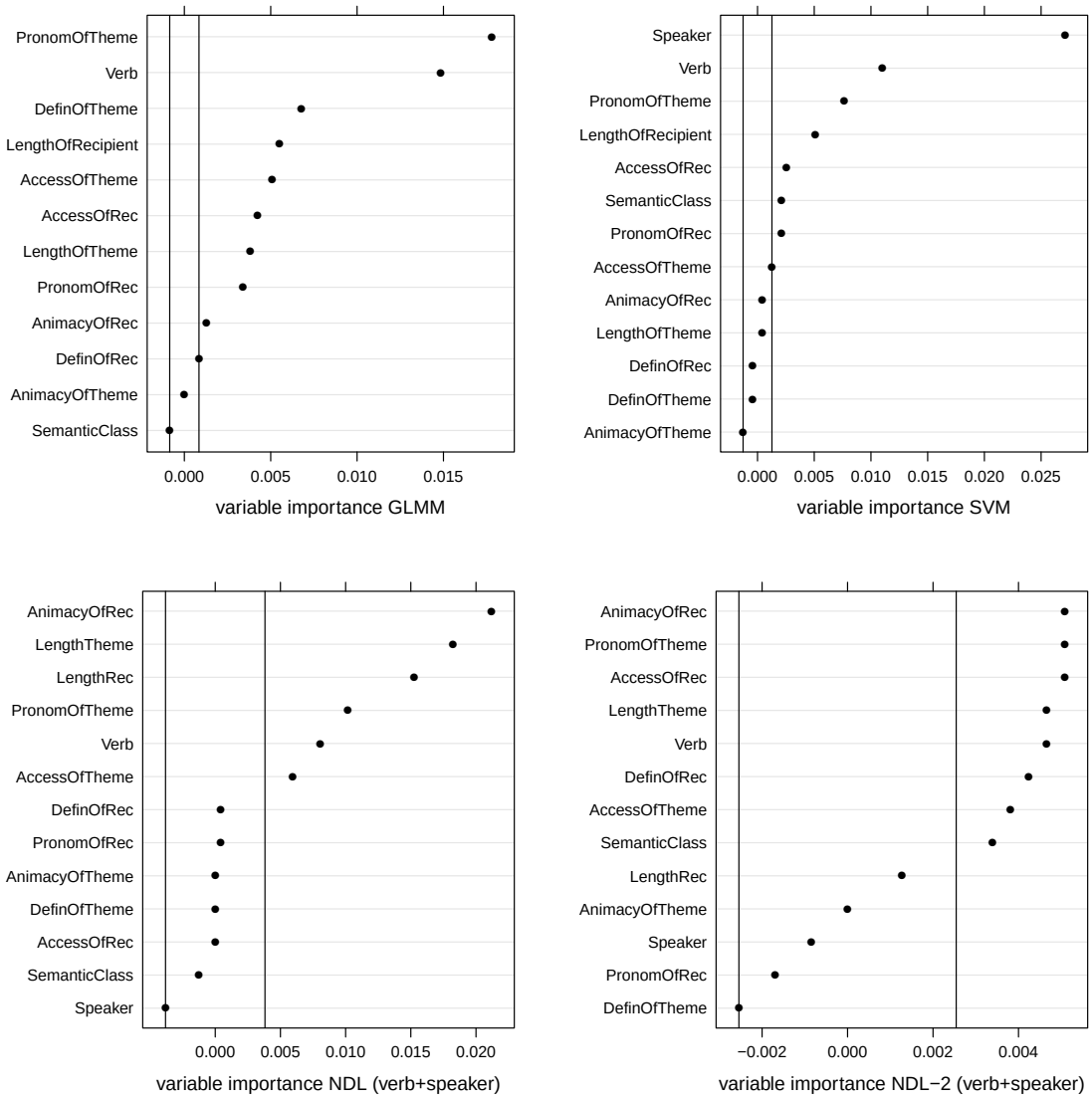

Figure 3. Permutation accuracy importance: the reduction in accuracy for predicting the prepositional object construction when a predictor is randomly permuted, for mixed-effects logistic regression (upper left), a support vector machine (upper right), naive discriminative learning (lower left), and naive discriminative learning with feature pairs (lower right).

The minor effect of variable permutation also indicates that, apparently, individual predictors are not that important. This is in all likelihood a consequence of the correlational structure characterizing the predictor space. For the dative set, each of the predictors listed in Table 4 can be predicted from the other predictors, with 2 up to 6 of the other predictors having significant coefficients ($p < 0.05$), and with prediction accuracies up to 95%. Although this kind of rampant collinearity can pose

serious problems for statistical analysis (in fact, a conditional variable importance measure (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008) for random forests would be a better choice than the straightforward permutation measure used above), it probably provides exactly the redundancy that makes human learning of language data robust. The improvement in classification accuracy of the naive discriminative learner when provided with feature pairs instead of single features as cues provides further support for the importance of redundancy. By making richer co-occurrence information available to the model, classification accuracy increases. The other side of the same coin is that permuting one predictor's values leaves prediction accuracy virtually unchanged: The 'functional' burden of individual predictors is small.

	Accuracy	Number of Significant Predictors
Animacy of Recipient	0.93	3
Definiteness of Recipient	0.91	2
Pronominality of Recipient	0.91	5
Accessibility of Recipient	0.95	4
Length of Recipient	0.33	3
Animacy of Theme	0.03	4
Definitiness of Theme	0.79	4
Pronominality of Theme	0.90	4
Accessibility of Theme	0.76	6
Length of Theme	0.04	2

Table 4: Prediction accuracy and number of significant predictors for (logistic) regression models predicting one predictor from the remaining other predictors.

Non-normal speaker variability

From a methodological perspective, it is noteworthy that Figure 3 clarifies that the importance of individual predictors is evaluated rather differently by the different models. The information gain ratios used by TiMBL to evaluate exemplar similarity, not shown here, provide yet another, and again different, ranking of variable importance. In the light of this diversity, one would hope that the variable importance suggested by models that are cognitively more realistic are closer to the truth. Whether this is indeed the case for naive discriminative learning awaits further validation, perhaps through psycholinguistic experimentation.

In what follows, we focus on one particularly salient difference, the discrepancy between the SVM and the other models when it comes to the importance of Speaker. Figure 4 visualizes the distributions of the contributions of the verb and speaker weights to the probability of the prepositional object construction in NDL-2, as well as the random intercepts for the verbs in the generalized linear mixed model.

The left panels show estimated probability density functions, the right panels the corresponding quantile-quantile plots.

The top panels present the NDL-2 weights for the associations of verbs to the prepositional object construction in the naive discriminative learning model. These weights follow, approximately, a normal distribution. The central panels graph the distribution of the random intercepts for the verbs in the GLMM, these also roughly follow a normal distribution. The NDL-2 verb weights and the GLMM random intercepts for verbs correlate well, r = 0.77 ($t(36) = 7.18$, $p = 0$), indicating that the two models are representing the same variation in a very similar way.

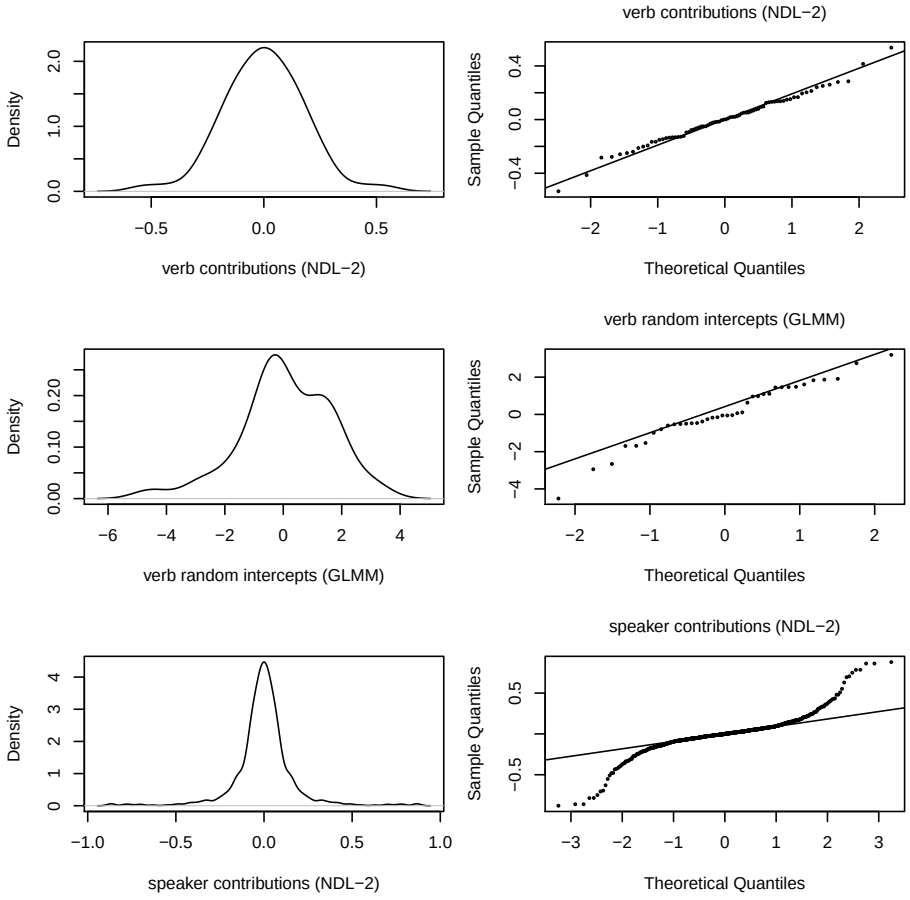

Figure 4. Distributions of the contributions of the individual verbs (top) and speakers (bottom) to the likelihood of the prepositional object construction, and the by-verb random intercepts in the generalized linear mixed model (center panels).

The bottom panels summarize the distribution of the association strengths from speakers to the prepositional object construction in the NDL. These weights are characterized by a symmetrical distribution that, however, deviates markedly from normality. There are too many very small weights close to zero, combined with long but

slim tails with outliers. This is, at least in part, due to the sparsity of information on individual speakers (the median number of observations for Speaker is 4, less than half of the median for Verb, 10.4).

The generalized linear mixed model builds on the assumption that random intercepts follow a normal distribution. For the speakers, this assumption is clearly violated. The mixed-effects model either fails to detect non-normally-distributed speaker variability, or infers that including speaker as random-effect factor does not lead to improved prediction. As the GLMM slightly outperforms the SVM under cross-validation, it seems likely that the SVM may be overfitting the data. The permutation variable importance for speaker in the naive discriminative learning models points in the same direction.

Returning to the difference between machine learning and human learning, the performance of naive discriminative learning suggests that human learning might be sensitive to variation (such as variation coming with individual speakers) that machine learning would back off from. However, for the human learner, thanks to the highly redundant nature of the set of predictors, the consequences of human overfitting seem negligible.

Naive discriminative learning and distinctive collexeme analysis

We have seen that naive discriminative learning provides a statistical tool for classification that, at least for the present data set, performs comparably to other state-of-the-art statistical classifiers. Crucially, naive discriminative classification is theoretically motivated as the end state of human discriminative learning. Over time, very simple adjustments to the association strengths of verbs to constructions result in excellent classification performance. The aim of this section is to show that within this new approach, a measure for distinctive collexeme analysis can be straightforwardly formulated.

Distinctive collexeme analysis (Gries & Stefanowitsch, 2004) quantifies to what extent a word is attracted to a particular construction. For instance, for the verb *take*, a contingency table (Table 5) serves as the input to a Fisher exact test of independence. The p-value produced by this test is log-transformed. The absolute value of the resulting measure is used to gauge attraction to or repulsion from a given construction. For *take*, distinctive collexeme strength is 35.7, indicating extremely strong attraction to the prepositional object construction. (Here, and in what follows, the focus is on the prepositional object construction.)

From a statistical perspective, it is somewhat odd to derive a measure from a p-value. An alternative approach is to make use of a measure from information theory, the Kullback-Leibler divergence, also known as relative entropy. Relative entropy specifies the difference between two probability distributions. The first probability distribution, p, concerns the probabilities of the two constructions for the verb *take*. The second probability distribution, q, specifies the probabilities of the two construc-

	NP NP	NP PP
take	2	56
other verbs	1857	445

Table 5: Contingency table for distinctive collexeme analysis of *take*.

	p	q
double object construction	2/(2+56)	(2+1857)/(2+56+1857+445)
prepositional object construction	56/(2+56)	(56+445)/(2+56+1857+445)

Table 6: The probability distribution p and q required for the calculation of the relative entropy for *take*.

tions in general. From Table 5 these probabilities can be obtained straightforwardly, as shown in Table 6. Given the two distributions p and q, their relative entropy is defined as

$$\text{RE(p,q)} = \sum_i p_i \log_2 \frac{p_i}{q_i}, \tag{11}$$

which for *take* evaluates to 1.95.

Alternatively, the ΔP measure (Allan, 1980; Ellis, 2006) can be used. This measure comes from learning and conditioning theory in psychology, where it has been found to be useful to probe cue learnability. Given a contingency table m cross-tabulating for the presence and absence of a given cue C and outcome O,

$$m = \begin{pmatrix} & O & -O \\ C & a & b \\ -C & c & d \end{pmatrix} \tag{12}$$

this one-way dependency statistic is defined as

$$\begin{aligned} \Delta P &= \Pr(O|C) - P(O|-C) \\ &= a/(a+b) - c/(c+d) \\ &= (ad-bc)/[(a+b)(c+d)]. \end{aligned} \tag{13}$$

ΔP ranges between -1 and 1, and represents the difference between two conditional probabilities, the probability of the outcome given the cue, and the probability of the outcome in the absence of the cue. For the data in Table 5, ΔP for the cue *take* and the outcome NP NP is -0.77, indicating that the cue *take* decreases the probability of the double object construction. Conversely, ΔP for the cue *take* and the prepositional object construction is 0.77, indicating that *take* is a reliable cue for this construction.

Yet another option for quantifying a verb's preference for a construction is to use the random intercepts of the generalized linear mixed model. For *take*, this

random intercept (the adjustment of the baseline log-odds for the prepositional object construction) is 2.75, again indicating that the use of this verb is biased towards the prepositional object construction.

Finally, we can also use the association strength of a verb to a construction as estimated by naive discriminative learning as a measure for distinctive collexeme strength. In the model with both Verb and Speaker, the association strength (weight) to the prepositional object construction for *take* is 0.13. The verb *promise* has the largest negative association strength for the prepositional object construction (-0.28), and the verb *read* the largest (0.54).

Figure 5 presents a scatterplot matrix for the five measures for distinctive collexeme analysis, calculated across all verbs. First note that all measures enter into positive correlations that are consistently significant according to the non-parametric Spearman correlation test. The standard measure of Collexeme Strength is most clearly correlated with the relative entropy measure. ΔP correlates well with Relative Entropy, with the Random Intercepts, and with the Cue Strengths. Furthermore, the random intercepts of the GLMM and the verb-to-construction association strengths of the NDL are strongly correlated. The Random Intercepts and the Cue Strengths emerge as less prone to generate extreme outliers. For instance, whereas *take* is an extreme outlier on the scale of the Collexeme Strength and Relative Entropy measures, it is well integrated within the cloud of data points for the Random Intercepts and Cue Strengths.

What this survey of measures suggests is that corpus linguistics has a range of measures for verb-specific constructional preferences at its disposal that probably all do a decent job of highlighting verbs with strong constructional biases. The Cue-to-Construction Strength measure, however, is particularly interesting and promising, in that it is derived from the Rescorla-Wagner equations, described by Ellis (2006) as "the most influential formula in the history of conditioning theory". As a speaker/listener becomes more and more proficient in a language, the association strengths of words to constructions become more and more fine-tuned to the distributional properties of the language. For a verb such as *take*, the speaker/listener comes to expect the prepositional object construction. Sentences such as *We hope he took his mother the ingredients to bake a Simnel Mothering Cake* (`stpauls-healdsburg.org/wp-content/uploads/.../2010/201004-stpauls.pdf`) then come as a surprise, violating the expectation of a prepositional object construction, but at the same time constituting a learning experience with concomitant adjustments of the association strengths of this verb to the double object construction.

General Discussion

Corpus linguistics is generally conceived of as a descriptive subdiscipline of linguistics. As increasingly powerful and realistic models of human learning and cognition become available, however, corpus linguistics can begin to take on the challenge of not only describing distributional patterns in corpora, but also of explaining the

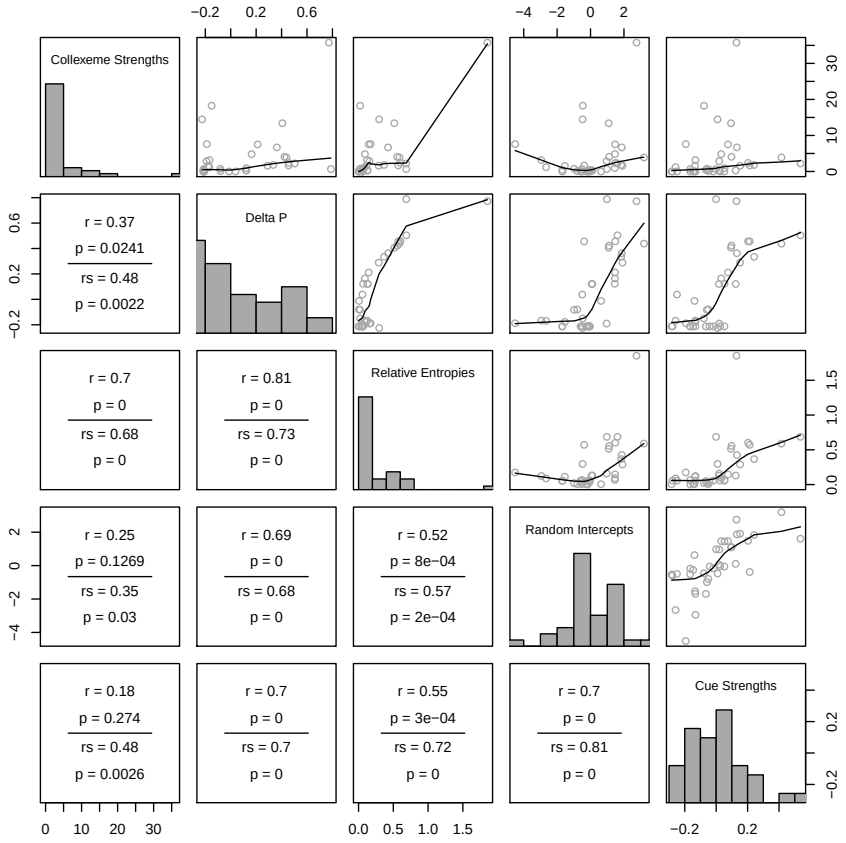

Figure 5. Different measures of collexeme strength and their pairwise correlations (Pearson and Spearman).

consequences of the observed distributional patterns for human learning and linguistic choice behavior.

Over the last decades, the statistical evaluation of distributional patterns has become increasingly important in corpus linguistics. Statistical models provide excellent insight into the quantitative structure of distributional patterns, but it is unclear to what extent such models provide an adequate characterization of the speaker-listener's actual knowledge. Moreover, the way in which statistical models derive a quantitative characterization of distributional patterns will, in general, be very different from how the speaker-listener acquires this knowledge.

As a first step towards a better cognitive grounding of quantitative analysis in corpus linguistics, the present study introduces a classifier grounded in naive discriminative learning. Using the data of Bresnan et al. (2007) on the dative alternation in spoken English as a case study, we have been able to show that, in theory, human classification can achieve nearly the same high level of accuracy as current state-of-the-art

machine-learning techniques.

We have to be careful with this conclusion, however. First, this study has examined only one data set. Naive discriminative learning may not perform as well on other more complex data sets. Second, the validity of naive discriminative learning as a model for how speaker-listeners acquire and represent probabilistic knowledge depends on the validity of the Rescorla-Wagner equations. These equations specify learning under optimal conditions, without noise factors such as lack of attention, incomplete assessment of relevant cues, and incomplete knowledge of the targeted outcomes. The present results for naive discriminative learning therefore probably represent an upper bound for human performance. Third, although it is well known that dopamine neurons display a short-latency, phasic reward signal that indicates the difference between actual and predicted rewards (Schultz, 2002; Daw & Shohamy, 2008), providing a neuro-biological justification for the hypothesis that learning is error-driven, it is well-known that the Rescorla-Wagner equations, however fruitful, do not cover all aspects of learning (Miller et al., 1995; Siegel & Allan, 1996).

Although naive discriminative classification performs well for the present data set, the conclusion that machine classification and human classification would be equivalent is not warranted. An examination of variable importance across models suggests that although statistical models can achieve comparable performance, they may do so by assigning predictors rather different explanatory relevance. There is a surprising and from a statistical perspective disquieting lack of convergence in the variable importance assigned to the predictors for the dative constructions across the support vector machine model, the logistic regression model, and the naive discriminative learner (Figure 3). In the face of such diversity, one would hope that that a statistical classifier derived from principles of human learning may provide superior estimates of variable importance for human-produced quantitative data. Without experimental support, unfortunately, this remains a conjecture at best.

The association strengths from verbs to constructions emerge in the naive discriminative learning approach as a natural alternative for quantifying distinctive collexeme strength. Although the five measures considered in this study (Collexeme strength, ΔP, Relative Entropy, Random Intercepts, and Cue Strength) are all correlated and useful as measures of collexeme strength, it is only the Cue Strength measure that is fully grounded in learning theory. It offers an important advantage compared to the other measure originating in psychology, ΔP. While ΔP is appropriate for 2 by 2 contingency tables (Allan, 1980), the Cue Strength measure handles n by 2 contingency tables appropriately. Crucially, the Cue Strength measure takes into account that many different cues may compete for a given outcome. Consider, for instance, the expression in the second row of equation (2) above,

$$\lambda - \sum_{\text{PRESENT}(C_j,\, t)} V_j.$$

When many cues are present simultaneously, the sum over cues will be larger, hence a larger number is subtracted from λ, and as a consequence, the cue-to-outcome asso-

ciation strength will increase with a smaller amount. Furthermore, when estimating the equilibrium association strength, the co-occurrence frequencies of the individual cues and outcomes are taken into account. By contrast, the ΔP measure ignores all other cues that can co-occur with a given outcome.

The naive discriminative learning model compresses experience with 2360 verb tokens, each characterized by 14 values (construction, verb, speaker, and 11 predictors) into a matrix of cue-to-construction association strengths with dimensions 865 by 2, a reduction from $2360 \times 14 = 33040$ values to only 1730 values, which amounts to a reduction by almost a factor 20. This reduced representation of past experience in terms of cue-to-construction strengths is reminiscent of connectionist models. The discriminative learning approach shares with the connectionist models of Seidenberg and McClelland (1989) and Harm and Seidenberg (2004), as well as with the Competition Model (E. Bates & MacWhinney, 1987; MacWhinney, 2005) the axiom that learning and generalization is driven by the distributional properties of the input. The discriminative learning model differs from the abovementioned connectionist models in terms of its architecture, which is much simpler. It does not make use of subsymbolic, distributed, representations, and it dispenses with hidden layers of all kinds. As a consequence, it is extremely parsimonious in free parameters: The only free parameter in the present study is whether to make use of single features or of feature-pairs. Computation of the weight matrix is also computationally much more efficient than in connectionist models. Computational effeciency also compares very favorably with random forests (Breiman, 2001; Strobl et al., 2009), a high-performance non-parametric classifier that, unfortunately, is extremely slow for data with factors such as speaker and verb that have very large numbers of levels.

Note that the discriminative learner approach offers the possibility of gauging not only verb-related constructional preferences, but also speaker-related constructional preferences, by means of the weights on the connections from speakers to constructions.

The way in which knowledge is represented in naive discriminative learning differs from other (non-connectionist) computational models for linguistic generalization. In exemplar-based approaches, it is assumed that in the course of experience, exemplars are stored in memory. Prediction is based on similarity neighborhoods in exemplar space. Data Oriented Parsing (Bod, 2006), Analogical Modeling of Language (Skousen, 1989), and Memory Based Learning (Daelemans & Bosch, 2005) provide examples of this general approach.

An important advantage of exemplar-based approaches is that the generalization process is simple and remarkably accurate in its predictions, as witnessed for the present data set by the classification results obtained with TiMBL, using its out-of-the-box default settings of parameters. An important disadvantage is that exemplars must be assumed to be available in memory, which may be unrealistic for human language processing. For example, recent studies suggest that the frequency with which a given sequence of words occurs in the language is predictive for how quickly

such a sequence is processed (Arnon & Snider, 2010; Tremblay & Baayen, 2010). This frequency effect persists for non-idiomatic sequences and for sequences that are incomplete phrases (as, e.g., *the president of the*). The assumption that shorter n-grams are stored in memory implies that hundreds of millions of exemplars would be remembered. This seems unrealistic. While naive discriminative learning shares with memory based learning the premise that each exemplar is important and contributes to learning, unlike memory-based learning, it does not need to posit that individual exemplars 'exist' independently in memory: Exemplar information is merged into the weights.

Instead of calculating predictions over an acquired instance space at run time, as in memory-based learning, one can instead seek to construct rule systems or constraint systems that capture the quantitative forces shaping behavior without having to store exemplars. The Gradual Learning Algorithm of Stochastic Optimality Theory (Boersma & Hayes, 2001) and the Mimimum Generalization Learner (Albright & Hayes, 2003) are examples of this approach. These rule-based approaches do not run into the problem that the instance base can become extremely voluminous, but they are challenged by frequency effects documented for linguistic units such as regular complex words and n-grams. Rule-based approaches tend to ignore these frequency effects, leaving them aside as an unsolved issue supposedly irrelevant to understanding the nature of generalization in human cognition. Rule-based approaches are also challenged by a proliferation of rules necessary to capture the fine details of the quantitative patterns in the data. Naive discriminative learning, by contrast, dispenses with the necessity of positing that the speaker-listener deduces large and complex rule sets from the input. Excellent classification accuracy can be obtained without storing exemplars and without rule induction or deduction.

In summary, the potential importance of naive discriminative learning for corpus linguistics is that it offers a unified framework for learning, for classification, and for distinctive collexeme (and distinctive collocutor) analysis. It is conceivable that variable importance is more adequately assessed by means of discriminative learning. Furthermore, naive discriminative learning may detect non-normally distributed variability where classic mixed models cannot do so. Finally, in discriminative learning, single cues make only modest contributions to classification accuracy. The present case study suggests that cues for outcomes tend to be highly interdependent and to a considerable extent predictable from each other. As such, they constitute a rich and redundant feature space in which a highly context-sensitive error-driven learning algorithm such as defined by the Rescorla-Wagner equations functions well, unhampered by issues of collinearity that plague (parametric) regression models.

Assuming that naive discriminative learning is on the right track as a characterization of human learning and categorization, many important questions remain unanswered. One such question is how speakers/listeners become knowledgeable about the cues and outcomes on which naive discriminative classification is based. Another question is why the language input to the model typically displays high-dimensional

correlational structure, as exemplified by the dative alternation data. Although intercorrelated, redundant feature spaces are apparently relatively easy to learn, at least under ideal conditions, it remains unclear why the data take the distributional forms typically attested in corpora. Furthermore, our use of the equilibrium equations for the Rescorla-Wagner equations assumes that the adult system would be completely stable and not subject to further change, which is only approximately correct.

The Rescorla-Wagner characterization of discriminative learning is in all likelihood incomplete, in that it does not do justice to tiny biases favoring outcomes that are cognitively easier to process (such as given information preceding new information). Within a speech community, such small biases would, under favorable circumstances, gain momentum, leading to locally optimal, 'functional' distributional patterns. Under this scenario, predictors such as animacy, definiteness, and information status would not shape an individual speaker's production as, for instance, in the variable rule approach of Cedergren and Sankoff (1974). Instead of a given utterance being governed by a probabilistic set of cognitive constraints operating at the level of an individual's brain, an utterance would be shaped by past experience under error-driven discriminative learning, much as described above for the dative alternation. However, tiny cognitive biases, neglected in the current formulation of the naive discriminative learner, would over time give rise to a speech community the utterances of which would then reflect, to some extent, varying from speech community to speech community, these very cognitive biases.

A challenge for corpus linguistics is to develop multi-agent computational models demonstrating that indeed tiny cognitive biases in discriminative learning can generate the kind of grammars and their trajectories of diachronic change that we find in human speech communities. With efficient algorithms such as provided by the equilibrium equations, realistic computational methods are coming within reach.

References

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*, 119–161.

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.

Anderson, J. R. (2000). *Learning and memory: An integrated approach*. New York: Wiley.

Arnon, I., & Snider, N. (2010). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Journal of Memory and Language*, *62*, 67–82.

Arppe, A., & Baayen, R. H. (2011). ndl: Naive discriminative learning: an implementation in r [Computer software manual]. Available from `http://CRAN.R-project.org/package=ndl` (R package version 0.4)

Baayen, R. H. (2009). languageR: Data sets and functions with "analyzing linguistic data: A practical introduction to statistics". [Computer software manual]. Available from `http://CRAN.R-project.org/package=languageR` (R package version 0.955)

Baayen, R. H., & Hendrix, P. (2011). Sidestepping the combinatorial explosion: Towards a

processing model based on discriminative learning. *Empirically examining parsimony and redundancy in usage-based models, LSA workshop, January 2011*.

Baayen, R. H., Milin, P., Filipovic Durdjevic, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, in press.

Bates, D., & Maechler, M. (2009). lme4: Linear mixed-effects models using s4 classes [Computer software manual]. Available from `http://CRAN.R-project.org/package=lme4` (R package version 0.999375-31)

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 157–193).

Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review*, *23*(3), 291–320.

Boersma, P., & Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, *32*, 45-86.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Bouma, I. Kraemer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Royal Netherlands Academy of Arts and Sciences.

Cedergren, H., & Sankoff, D. (1974). Variable rules: Performance as a statistical reflection of competence. *Language*, *50*(2), 333–355.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Science*, *10*(7), 287–291.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256.

Daelemans, W., & Bosch, A. Van den. (2005). *Memory-based language processing*. Cambridge: Cambridge University Press.

Daelemans, W., Zavrel, J., Sloot, K. Van der, & Bosch, A. Van den. (2010). *TiMBL: Tilburg Memory Based Learner Reference Guide. Version 6.3* (Technical Report No. ILK 10-01). Computational Linguistics Tilburg University.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, *47*(2), 109–121.

Daw, N., & Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition*, *26*(5), 593–620.

Dayan, P., & Kakade, S. (2001). Explaining away in weight space. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems 13* (pp. 451–457). Cambridge, MA: MIT Press.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2009). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien [Computer software manual]. (R package version 1.5-19)

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*(1), 1–24.

Ford, M., & Bresnan, J. (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language*, *86*(1), 168–213.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology:General*, *117*(3), 227–247.

Gries, S. (2011). Frequency tables: tests, effect sizes, and explorations. In D. Glynn & J. Robinson (Eds.), *Polysemy and synonymy: Corpus methods and applications in Cognitive Linguistics*. Amsterdam & Philadelphia: John Benjamins.

Gries, S., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on alternations. *International Journal of Corpus Linguistics*, *9*(1), 97–129.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720.

Harrell, F. (2001). *Regression modeling strategies*. Berlin: Springer.

Hsu, A. S., Chater, N., & Vitányi, P. (2010). The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis. *Manuscript submitted for publication*.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1-38.

MacWhinney, B. (2005). A unified model of language acquisition. In J. Kroll & A. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 49–67). Oxford University Press.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of the basic findings. *Psychological Review*, *88*, 375-407.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner Model. *Psychological Bulletin*, *117*(3), 363–386.

Murray, W. S., & Forster, K. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, *111*, 721–756.

Norris, D. (2006). The Bayesian Reader: Explaining Word Recognition as an Optimal Bayesian Decision Process. *Psychological Review*, *113*(2), 327–357.

Norris, D., & McQueen, J. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.

Ramscar, M., Dye, M., Popick, H. M., & O'Donnell-McCarthy, F. (2011). The Right Words or Les Mots Justes? Why Changing the Way We Speak to Children Can Help Them Learn Numbers Faster. *PLoS ONE*, to appear.

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, *31*(6), 927–960.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(7), in press.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*(2), 241–263.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psycholgical Review*, *96*, 523–568.

Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review*, *3*(3), 314–321.

Skousen, R. (1989). *Analogical modeling of language*. Dordrecht: Kluwer.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*. Available from `http://www.biomedcentral.com/1471-2105/9/307`

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological Methods*, *14*(4), 26.

Tagliamonte, S., & Baayen, R. (2010). Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Manuscript submitted for publication*.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on Formulaic Language. Acquisition and Communication* (pp. 151–173).

Van Heuven, W. J. B., Dijkstra, A., & Grainger, J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language*, *39*, 458-483.

Vapnik, V. (1995). *The nature of statistical learning theory*. Berlin: Springer-Verlag.

Wagner, A., & Rescorla, R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ii* (pp. 64–99). Appleton-Century-Crofts.

Appendix: naive discriminative classification with the ndl package

The `ndl` package (Arppe & Baayen, 2011), available in the CRAN archives at `www.r-project.org`, provides software for naive discriminative learning for the R statistical programming environment. The following provides an introduction to its basic functionality.

As a first step, we attach the package, extract the `dative` dataset, and remove the data for which no speaker information is available.

```
> library(ndl)
> data(dative)
> dative = dative[!is.na(dative$Speaker), -2]
```

We fit a basic naive discriminative classifier to the data using the standard formula-based interface, where the dot is expanded into all predictors in the `dative` data frame other than the dependent variable (`RealizationOfRecipient`):

```
> dative.ndl = ndlClassify(RealizationOfRecipient ~ ., data = dative)
```

Numeric predictors are converted into factors, by default each factor has two levels. This default can be changed by the user, as explained in the documentation. Models with cue pairs can be specified using the interaction notation for R formulae. For instance,

```
> dative.ndl2 = ndlClassify(RealizationOfRecipient ~ (SemanticClass +
+     LengthOfRecipient + AnimacyOfRec + DefinOfRec + PronomOfRec +
+     LengthOfTheme + AnimacyOfTheme + DefinOfTheme + PronomOfTheme +
+     AccessOfRec + AccessOfTheme) + Verb + Speaker, data = dative)
```

includes pairwise cues for all independent variables, except verb and speaker.

The weight matrix can be extracted from the model object, which is a list:

```
> names(dative.ndl)
[1] "activationMatrix" "weightMatrix"     "cuesOutcomes"     "frequency"
[5] "call"             "formula"          "data"

> head(dative.ndl$weightMatrix)
                                   NP           PP
AccessOfRecaccessible   -0.015468613  0.083503412
AccessOfRecgiven         0.094783250 -0.026748451
AccessOfRecnew          -0.009894431  0.077929230
AccessOfThemeaccessible  0.089768608 -0.021733809
AccessOfThemegiven      -0.093523058  0.161557856
AccessOfThemenew         0.073174656 -0.005139857
```

The association strengths of the individual verbs to the constructions can be accessed as follows:

```
> w = dative.ndl$weightMatrix
> verbs = w[grep("Verb", rownames(w)), ]
> verbs = verbs[order(verbs[, "PP"]), ]
> head(verbs)
                    NP          PP
Verbaward   0.6194557 -0.6146320
Verbbet     0.3843946 -0.3795708
Verbowe     0.3570426 -0.3522188
Verbpromise 0.3425307 -0.3377070
Verbtell    0.3036573 -0.2988335
Verbteach   0.1962304 -0.1914066

> tail(verbs)
                    NP         PP
Verbhand   -0.2039228 0.2087466
Verbbring  -0.2059774 0.2108011
Verbleave  -0.2593050 0.2641288
Verbwrite  -0.4221433 0.4269670
Verbread   -0.4432427 0.4480664
Verbafford -0.6125922 0.6174159
```

A summary method for `ndl` objects is available that provides a wide range of measures of goodness of fit, including

```
> summary(dative.ndl)$statistics$C
[1] 0.9820687

> summary(dative.ndl)$statistics$accuracy
[1] 0.9457627
```

A crosstabulation of observed and predicted values is available with

```
> summary(dative.ndl)$statistics$crosstable
     NP  PP
NP 1821  38
PP   90 411
```

The predicted probabilities of the double object and prepositional object constructions for each row of the `dative` data frame are obtained with

```
> p = acts2probs(dative.ndl$activationMatrix)$p
> head(p)
```

```
           NP        PP
[1,] 0.7582780 0.2417220
[2,] 0.1872549 0.8127451
[3,] 0.5710474 0.4289526
[4,] 0.5707516 0.4292484
[5,] 0.5190592 0.4809408
[6,] 0.4767222 0.5232778
```

```
> tail(p)
```

```
              NP        PP
[2355,] 0.5009516 0.4990484
[2356,] 0.6145346 0.3854654
[2357,] 0.6999555 0.3000445
[2358,] 0.4434956 0.5565044
[2359,] 0.6017827 0.3982173
[2360,] 0.6433302 0.3566698
```

Crossvalidation can be carried out as follows:

```
> dative.ndl.10 = ndlCrossvalidate(RealizationOfRecipient ~ .,
+     data = dative)
```

```
> summary(dative.ndl.10)$statistics.summary["Mean", "C"]
```

```
[1] 0.9265221
```

```
> summary(dative.ndl.10)$statistics.summary["Mean", "accuracy"]
```

```
[1] 0.8889831
```

Permutation variable importance is assessed with

```
> dative.varimp = ndlVarimp(dative.ndl)
```

```
> library(lattice)
> dotplot(sort(summary(dative.ndl)$statistics$accuracy - dative.varimp$accuracy),
+     xlab = "permutation variable importance")
```