
Article:
Milin, P. orcid.org/0000-0001-9708-7031, Divjak, D., Dimitrijević, S. et al. (1 more author) (2016) Towards cognitively plausible data science in language research. Cognitive Linguistics, 27 (4). pp. 507-526. ISSN 0936-5907

https://doi.org/10.1515/cog-2016-0055

Towards cognitively plausible data science in language research

Abstract
Over the past 10 years, Cognitive Linguistics has taken a Quantitative Turn. Yet, concerns have been raised that this preoccupation with quantification and modelling may not bring us any closer to understanding how language works. We show that this objection is unfounded, especially if we rely on modelling techniques based on biologically and psychologically plausible learning algorithms. These make it possible to take a quantitative approach, while generating and testing specific hypotheses that will advance our understanding of how knowledge of language emerges from exposure to usage.

Keywords
naive discrimination learning, memory-based learning, lexical decision, Serbian

Authors
Petar Milin
<p.milin@sheffield.ac.uk>
Department of Journalism Studies
The University of Sheffield, UK
and
Quantitative Linguistics, Department of Linguistics
Eberhard Karls University Tübingen, Germany

Dagmar Divjak
<d.divjak@sheffield.ac.uk>
School of Languages & Cultures
The University of Sheffield, UK

Strahinja Dimitrijević
<strahinja.dimitrijevic@unibl.rs>
Department of Psychology
University of Banja Luka, Bosnia and Herzegovina

R. Harald Baayen
<harald.baayen@uni-tuebingen.de>
Quantitative Linguistics, Department of Linguistics
Eberhard Karls Universität Tübingen, Germany

Acknowledgments
The financial support of the Alexander von Humboldt Foundation (to Harald Baayen and Petar Milin) and the British Academy (to Dagmar Divjak) is gratefully acknowledged. We wish to thank Emmanuel Keuleers for providing generous help in implementing TiMBL, and Svetlana Borojević who provided access to the experimental hardware and software.

1. Introduction

Within cognitive linguistics the number of publications relying on empirical data collections and statistical data modelling has increased spectacularly over the two last decades (c.f., Ellis & Larsen-Freeman, 2006; Gries & Divjak, 2010; Zeschel, 2008). The field now abounds with studies that use statistical classification models to analyse either textual corpus data or behavioural experimental data. The most advanced corpus-based studies rely on a range of statistical analyses, most often regression-based, to model data that has been annotated for a multitude of linguistically relevant parameters (i.e., linguistic abstractions). The goal of these studies is to determine which parameters might be predictive for the form in focus. Think of, for example, the well-known constructional alternation studies by Bresnan, Cueni, Nikitina and Baayen (2007).

The fact that this approach does not implement the cognitive commitment (Lakoff, 1990) on a number of points, relating to different stages of the analysis process, has not yet attracted much attention in the literature (but see Divjak, 2015). We discuss the implications of the way in which the data is annotated, prepared for statistical analysis and, finally, modelled, on the cognitive reality of the resulting linguistic analysis.

First, datasets are typically annotated for various higher-level abstractions (i.e., morpho-syntactic, syntactic, semantic, and discourse-related features) that are believed to be helpful in revealing systematicity in language. These features are, however, often difficult to define and to annotate with high levels of agreement between human annotators. But even if that would not be the case, research has shown that abstract labels do not necessarily yield better modelling results than the actual words used in the sentences (Theijssen, ten Bosch, Boves, Cranen, & van Halteren, 2013).

Second, regression models are characterized by a quirk that seems incompatible with a fundamental property of language: *recurring* information or redundancy. To manage the uncertainty (i.e., entropy) inherent to communication, language encodes bits of information recurrently. Regression, however, requires explanatory predictors not to be collinear, and therefore, redundancy needs to be removed from predictors prior to modelling. This equals removing information that is part and parcel of the system we are trying to learn about statistically.

Third, although regression models have produced classification results that have received support from behavioural studies (for an overview of this relatively recent trend in linguistics see Klavan & Divjak, 2016), the algorithms these models rely on are not based on learning mechanisms but maximize likelihood using optimization techniques. Whether humans do or do not exhibit (near-)optimal behaviour remains a matter of debate (see Kahneman & Tversky, 1984).[1] What is undisputable, however, is that human (and animal) learning unfolds gradually over time, and that the order of exposition matters

[1] Bowers & Davis, 2012 critique the Bayesian approach in psychology and neuroscience that embodies the hypothesis of near-optimal behaviour in humans; see Griffiths, Chater, Norris, & Pouget, 2012 for a response to these critiques.

greatly. Regression-based statistical learning was not designed to take this core aspect of learning into account.

If we want empirical evidence to accrue and alter the way in which we think about language we should consider modelling techniques that implement principles of human behaviour, and of learning in particular. Such models have been used to model language data within traditions that are close in spirit to Cognitive Linguistics. Examples include Parallel-Distributed Processing or Connectionist Modelling (PDP: Plaut & Gonnerman, 2000; Rumelhart & McClelland, 1986; Seidenberg & Gonnerman, 2000), Analogical Modelling (AM: Skousen, 1989), Memory-based Learning (TiMBL: Daelemans & Van den Bosch, 2005), and more recently Naive Discriminative Learning (NDL: Baayen, Milin, Filipović Đurđević, Hendrix, & Marelli, 2011). The performance of several of these models has been compared (see Eddington, 2000 for a connectionist, an analogical and a memory-based model on the English past tense; Theijssen et al., 2013 compared logistic regression, Bayesian networks and Memory-based learning in predicting the English dative alternation; Baayen, 2011 compared NDL with TiMBL, Logistic Mixed-Effects Regression, Classification Trees & Random Forests, and Support Vector Machines – SVM, on the English dative alternation; Baayen, Endresen, Janda, Makarova, & Nesset, 2013 compared the same set of techniques on four different morphological alternations in Russian). Important for the current paper is the finding that the classification accuracy of NDL was outperformed only by SVM.

In the following sections we reflect on the cognitive commitment at the stage of data annotation, preparation and modelling and explore the possibility of using biologically and psychologically motivated modelling as a tool for designing behavioural experiments and as a guide to an in-depth discussion of the findings. The NDL approach, which will serve not only as a computational model but also as a theoretical framework, enables us to consider the impact of introducing radically usage-based patterns and associated cognitively plausible abstractions into linguistics proper.

2. Learning Theory

Usage-based linguistics is predicated upon the premise that the knowledge of language emerges from exposure to usage. With our linguistic abilities believed to be rooted in general cognitive abilities, this leaves a prominent role to be played by learning. Vigorously exiled from the linguistic landscape by Chomsky's (1959) criticism of Skinner's "Verbal Behavior" (1957), Learning Theory is still to make a full come-back onto the linguistic scene.[2]

Within psycholinguistics, a simple principle of learning, formally expressed in the Rescorla and Wagner rule (1972), has been attracting attention. This error-driven learning mechanism governs success in adaptation to an environment by iteratively correcting erroneous predictions for upcoming

[2] See MacCorquodale, 1970; Andresen, 1991; Virués-Ortega, 2005 for a discussion of Chomsky's misinterpretation of some of Skinner's crucial arguments.

events. In a nutshell, the Rescorla-Wagner rule defines how a system (an animal, human or a computer device) learns from its own errors in order to adapt to the task at hand.

Core components in this learning system are input *cues* and their *weight* in predicting learning *outcomes*. These weights are repeatedly updated as experience accumulates. Over time, some cues become discriminative (i.e., predictive) for an outcome, while many become irrelevant. The system is parsimonious in the sense that, for each outcome, only a handful of cues develop strong positive or negative connection weights to outcomes. If a given cue is consistently present when an outcome is present, their connection is strengthened. However, if a given cue is repeatedly present when the outcome is absent, the weight on the connection between them is weakened. This dynamic ensures minimal error in prediction given all prior experience. As the number of available cues increases, the amount by which the weight on its connection to an outcome can increase is affected. The more cues are present, the smaller the increase and the greater the decrease in weights will be. This reflects the competition between cues. The strengthening of weights reflects learning, and the weakening of links captures unlearning (for details see Baayen et al., 2011; Milin, Feldman, Ramscar, Hendrix, & Baayen, subm.).

The Rescorla-Wagner model was conceived to account for a range of learning phenomena that are also valuable for understanding life-long language learning. The *blocking* phenomenon (Kamin, 1969), for example, explains why an association between a cue and an outcome (the conditioned and unconditioned stimuli in traditional learning theory terminology) is impaired if that same outcome had already been paired with another cue: the second cue will not facilitate prediction of the outcome and for that reason the cue will be ignored. Blocking has been used to explain L2 acquisition (Ellis, 2006a), as well as phenomena of early language acquisition such as overgeneralization of irregular plurals (Ramscar & Yarlett, 2007) and difficulties in acquiring grammatical gender in L2 (Arnon & Ramscar, 2012).

The Rescorla-Wagner equations provide various parameters for differentiating the *salience* of cues and outcomes. There are parameters which specify the salience of an input cue i (α_i), and parameters for the maximum learnability of an outcome j (λ_j). Furthermore, the importance or strength of correct (β_1) vs. incorrect (β_2) predictions can be weighted differentially. Although in a typical simulation run these parameters are set to their default values, they allow for a principled account of various learning "peculiarities" (see Ghirlanda, 2005 for how a simple error-driven learning model, formally equivalent to the Rescorla-Wagner model, can account for a range of intricate learning phenomena). For example, in a series of experiments and modelling studies (Jordanov, Nešković, & Milin, 2015; Nešković, Jordanov, & Milin, 2015) it was shown that a change in the salience of a crucial learning cue could explain an unexpected pattern of results in a variant of the object naming task, designed to demonstrate the *highlighting* effect in learning. The effect itself distinguishes early vs. late learning, and perfect vs. imperfect learning cues (i.e., those that predict one and only one or more possible outcomes), and shows a prediction preference for early-learned imperfect cues and late-learned

perfect cues. Highlighting was also used to explain the *cognate* effect in L1/L2 lexical processing (Anđel, Radanović, Feldman, & Milin, 2015).

In recent years, support has been accumulating for error-driven learning as an explanatory model accounting for a wide range of language phenomena (Ellis, 2006a, 2006b, 2012; Ramscar & Yarlett, 2007; Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010; Ramscar & Dye, 2011; Dye, Milin, Futrell, & Ramscar, 2016). Naive Discrimination Learning (Baayen et al., 2011) provides a computational framework for error-driven discrimination of potentially very large numbers of outcomes given potentially also large numbers of cues. Computations scale up and can be run on large data sets, including corpora with billions of words.

As for any computational model, the representations chosen for cues and outcomes are crucial for the model's performance (c.f., Gallistel, 2008). Models for lexical processing typically made use of large numbers of simple cues, such as letter pairs or letter triplets, but cues can also be words, acoustic features, or constructional properties. Likewise, outcomes can range from lexical and grammatical features to idioms and constructions. Different networks can be combined, as in the study of Milin, Divjak and Baayen (subm.) which modelled both bottom-up orthographic learning and top-down semantic learning in sentence reading. NDL has proven successful in modelling the processing of a wide range of language phenomena from inflections to phrasal effects (Baayen et al., 2011), and in explaining the effects of priming and form neighbourhood (Milin et al., subm.), as well as frequency and age-of-acquisition (Baayen, Milin, & Ramscar, 2016).

Note, however, that NDL should not be mistaken for a classifier: it was not designed to compete with the state-of-the-art machine learning classification techniques (despite the fact that it achieves comparable results, as shown by Baayen, 2011 and Baayen et al., 2013). Instead of relying on optimization algorithms to maximize prediction accuracy, NDL is conceived to mimic human learning, including the restrictions on memory and learning that set human learning apart from machine learning. NDL, which could be viewed as a method for doing incremental regression (for discussion see Evert & Arppe, 2015), offers the advantages of being exquisitely sensitive to the order of learning events, while at the same time allowing researchers to consider many collinear predictors simultaneously.

3. Comprehension of “easy” and “difficult” plural nouns in Serbian: A lexical decision experiment

As our case study, we will present a TiMBL model (Daelemans & Van den Bosch, 2005) which produces novel inflected word forms in Serbian, relying on similar (i.e., neighbouring) word forms. TiMBL has been used to model allomorphy in the Serbian instrumental singular (Milin, Keuleers, & Filipović Đurđević, 2011) and outperformed Analogical Modelling (Skousen, 1989) in handling allomorphy in the Croatian instrumental singular and genitive plural (Lečić, 2016). It also showed good performance in producing a range of Serbian inflected word forms from their lemmata (Dimitrijević, 2015).

Training. TiMBL was trained on a sample of 89,024 different word tokens (pronouns and open-class content words), retrieved from the manually lemmatized and morpho-syntactically annotated *Frequency dictionary of contemporary Serbian* (Kostić, 1999).[3] Coded exemplars, the building blocks of the learning memory, contain three types of information: (1) the syllabic structure of the lemma; (2) a morpho-syntactic tag; (3) a class label with the inflectional suffix (e.g., "-ih"). Table 1 shows two coded exemplars using a syllable-based alignment method in which components are right-aligned. (1) represents the four last syllables of the given lemma, each consisting of onset, nucleus and coda. Lemmata consisting of fewer than four syllables (like the two presented in Table 1) would have the leftmost positions marked with "=", signalling the non-availability of an element. Similarly, (3) is right-aligned from 0 to 9, with the suffix attached at the end; the numbers flag alternations by position. This coding scheme has proven to work well for TiMBL (see Keuleers & Daelemans, 2007).

Table 1. Lemma, word form and exemplar structure.

Lemma	**Word form**	**Exemplar coding**		
		(1)	**(2)**	**(3)**
polagan	polaganih[a]	=, =, =, p, o, =, l, a, =, g, a, n	201221	9876543210ih
poseban	posebnih[b]	= ,=, =, p, o, =, s, e, =, b, a, n	201221	987654320ih

[a] slow; [b] special

The algorithm was trained on coded exemplars available in the memory and tested for the production of novel (i.e., unseen) forms, given its k-nearest neighbours. Details of this procedure are provided in Appendix A. Lemma and word form were always excluded from training and evaluation. The overall success rate of the TiMBL simulation was 89%. This is a conservative estimate, obtained by counting as errors all grammatically acceptable alternate forms such as doublets ("vukovi" – "vuci"), dialectal variants (ekavica: "mleko" – ijekavica: "mlijeko") etc.

Stimuli. Experimental items were selected from all masculine nouns for which a nominative plural was produced in the TiMBL simulation run. They were split into two groups of correct ("easy") and incorrect ("difficult") productions, and the items in these groups were matched in number for nominative plural formation. Frequency counts (retrieved from Kostić, 1999) were entered as a covariate in the statistical model. The final list of items consisted of 60 "easy" and 60 "difficult" nouns. *Wuggy* (Keuleers & Brysbaert, 2010) generated 120 pseudowords, matching the words from the list in length and phonotactics.

[3] The algorithm was trained using the *k-nearest neighbours* method, with *modified value difference* as similarity metric (MVDM: Cost & Salzberg, 1993), and *k* set to 7 as default neighbourhood size. An element's importance was determined with the *information gain ratio* (Quinlan, 2014). For further technical details we refer to Dimitrijević (2015).

Participants. 39 students (31 females) from The University of Banja Luka (Bosnia and Herzegovina) participated in the experiment in partial fulfilment of the course requirements. All were native speakers of Serbian with normal or normal-to-corrected vision and no known reading or speech disorders.

Procedure. The experiment began with 8 practice trials, followed by 240 experimental trials, randomized for each participant. Standard experimental procedures for administering a lexical decision task were adhered to. 13.72% of the data had to be removed from further analyses, leaving N = 4,048 datapoints.

Results. We made use of Generalized Additive Mixed Modelling (GAMM: Wood, 2006), as implemented in the **mgcv** package for **R** Statistical Environment (Wood, 2011; R CoreTeam, 2014).

A GAMM showed significant random effects of participants and items (respectively: F = 48.763, p < 0.0001; F = 6.719, p < 0.0001). The main experimental factor – word difficulty ("easy" vs. "difficult") as derived from TiMBL – made a significant contribution (t = 2.584, p = 0.0098), and entered into interactions with both form frequency and lemma frequency. Interestingly, while form frequency is significant for the "easy" words (F = 4.565, p < 0.0327), lemma frequency is predictive in case of the "difficult" words (F = 27.276, p < 0.0001). Both interactions are presented in Figure 1. Participants experience more difficulties *recognizing* those items that the TiMBL model found difficult to *produce*.

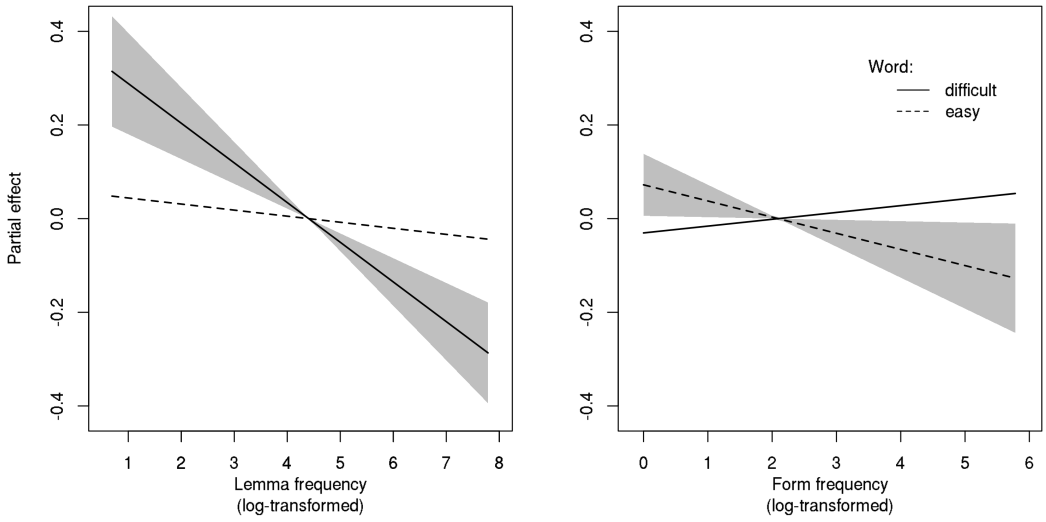

Figure 1. Lemma frequency by word difficulty (left panel) and form frequency by word difficulty (right panel) interaction. Confidence intervals are presented only for the significant effect of interactions.

Strikingly, TiMBL's inflectional class probabilities turn out to be predictive in production and comprehension, i.e., for lexical decision latencies. The question, however, remains whether we are justified to conclude, based on the convergence between the model's predictions and subjects' responses, that the way in which TiMBL learns from data mimics the way humans learn? Memory-

based learning offers an attractive framework for the exemplar-based modelling of language and language processing, and the TiMBL implementation provides researchers with a powerful toolkit for detecting which properties of exemplars guide prediction (see, e.g., Krott, Baayen, & Schreuder, 2001). However, the very techniques that facilitate these predictions (e.g., information gain weights or modified value differences) stem from probability theory and are based on conditional probabilities that assume, contrary to fact (see Kamin, 1969; Ramscar & Yarlett, 2007; Ellis, 2006a), that blocking should not occur. Whereas in some fields or for some applications one may not want to work with algorithms that are subject to blocking, respecting blocking is an important desideratum for the reverse engineering of human language processing. From this perspective, it is interesting that in evolutionary biology, the Rescorla-Wagner learning rule has been found to be superior in cross-generational performance than more advanced classifiers (c.f., Trimmer, McNamara, Houston, & Marshall, 2012).

In our analysis of reaction times for decisions regarding the lexicality of Serbian plural nouns, we observed that predictions about a word's plural form, generated by memory-based learning, explain some of the variance in reaction times. In the next section, we introduce some measures obtained through naive discriminative learning that extend our understanding of the variance observed in the reaction times.

4. Taking a Naive Discrimination Learning perspective

NDL was trained on a 300 million word Serbian subtitle corpus (Tiedemann, 2012). Subtitle corpora are easy to obtain, and constitute a particular register in which short, frequent, easy, and emotional words are over-used compared to spontaneous conversational speech and written registers (Baayen, Milin, & Ramscar, 2016). This constellation of properties yields frequency counts that have been found to be particularly well-suited for predicting reaction times in visual lexical decision tasks. Simple letter triplets (i.e., trigraphs) serve as orthographic input cues, while space-separated letter sequences – actual word forms in our present implementation – are learning outcomes. As laid out in Milin et al. (subm.), these word forms are referred to as *lexomes*, and are conceptualized as pointers that give access to locations in high-dimensional semantic co-occurrence space, where meaning is not fixed and encapsulated but distributed and dynamic and construed as the message unfolds.[4] Table 2 presents examples of the input and output representations used.

Table 2. Input cues and learning outcomes for the NDL model.

Form based input cues	**Outcome**	**Lemma**
#po, pol, ola, lag, aga, gan, ani, nih, ih#	polaganih	polagan[a]
#po, pos, ose, seb, ebn, bni, nih, ih#	posebnih	poseban[b]

[a] slow; [b] special

[4] Compare with Beard's (1977; 1981) *separation hypothesis*, and Aronoff's (1994) definition of a *lexeme*.

The model was trained on each of the 300 million words, one after the other, adjusting the weights from the cues to all outcomes using the Rescorla-Wagner learning rule. Training results in a cues-by-outcomes matrix. This matrix specifies, for any given cue and outcome pair, how well the cue supports the outcome. Given a word's input cues (**#po**, **pol**, **ola**, **lag**, **aga**, **gan**, **ani**, **nih**, and **ih#**), the sum of the connection weights from these cues to the outcome **polaganih** defines the outcome's activation. For understanding lexicality decisions, two "Grapheme-to-Lexome" (G2L) measures were found to be particularly important:

1. **Diversity**. The G2L-Diversity is the sum of the absolute values of the activations of all possible outcomes, given a set of input cues. Input cues that activate many different outcomes give rise to a highly diverse activation vector, which in turn indicates a high degree of uncertainty about the intended outcome.
2. **G2L-Prior**. The G2L-Prior is the sum of the absolute values of the weights on the connections from all cues to a given outcome. This measure, which is independent of the actual cues encountered in the input, reflects the prior availability of an outcome, its entrenchment in the learning network.

We ran two sets of statistical models using GAMMs (Wood, 2006), one in which the NDL measures were used to explain the TiMBL generated probabilities for the produced inflected forms (Section 4.1), and one in which they were used to explain the RT latencies from the lexical decision experiment directly (Section 4.2). The two learning-based measures were rank-transformed to facilitate statistical modelling.

4.1. Explaining TiMBL probabilities with discrimination learning measures

The GAMM model fitted to the TiMBL probabilities indicated that TiMBL probability increases linearly with G2L-Diversity (F = 6.869, p = 0.0099), as illustrated in Figure 2. With this single predictor the model accounts for 5.5% of explained deviance on N = 120 word items.

G2L-Diversity captures the dispersion of lexomes that are co-activated by the input cues (trigraphs). Lexomes that are irrelevant will have activations close to zero, and will not contribute to the diversity. Simply, letter triplets that are shared by many lexomes will boost their co-activation, which will be captured by higher values of G2L-Diversity. Thus, G2L-Diversity is an indirect measure of the number of near-neighbours of a given word form. For TiMBL to predict a plural with accuracy, it is important to have many exemplars that are

near-neighbours. Having more such neighbours allows TiMBL to make more precise predictions about the most likely shape of the word's plural form.[5]

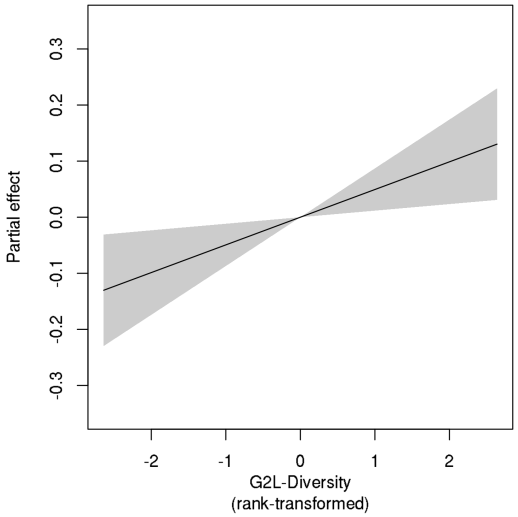

Figure 2. The effect of word-form G2L-Diversity fitted to the TiMBL generated probabilities.

However, a state of affairs that is optimal for the selection of a plural form in language production may be disadvantageous in language comprehension. Specifically, the co-activation of the many potential plural forms together with the intended plural form creates uncertainty, and this has been found to give rise to elongated reaction times in the visual lexical decision task. We will explore this in the next section.

4.2. Modelling lexicality decisions with discrimination learning measures

We fitted a GAMM to the visual lexical decision latencies with as predictors Word Difficulty, G2L-Diversity and G2L-Prior. The GAMM included random intercepts for participants and for items ($F = 45.908$, $p < 0.0001$; $F = 6.077$, $p < 0.0001$, respectively), as well as a main effect of word difficulty ("easy" vs. "difficult": $t = 2.616$, $p = 0.0089$). The predictor "word difficulty" was included because this two-level factor played a crucial role in the design of the experiment (reported in Section 3).[6] As illustrated in Figure 3, a greater G2L-Prior or stronger entrenchment in the learning network afforded shorter reaction times ($F = 51.221$, $p < 0.0001$), whereas, as predicted, a greater G2L-Diversity

[5] When NDL is used as a classifier, and a network is trained to predict the most likely plural form from letter trigraphs alone, an accuracy rate of nearly 72% is achieved. By adding the morpho-syntactic information that was made available to TiMBL the accuracy of the NDL predictions increases to 84%, which approaches the 89% accuracy of TiMBL. See, however, our discouragement regarding using NDL for classification problems in Section 3.

[6] Strictly speaking, it is possible that NDL is (dis)advantaged because the dichotomy that characterizes the TiMBL results may conflict with the predictions that NDL would make. For example, items characterized by high values of activation diversity would be less probable (see Figure 2).

($F = 20.356$, $p < 0.0001$) or co-activation of many possible plurals together with the intended plural gave rise to longer reaction times due to the increased uncertainty that comes with co-activation. Both NDL measures reported here as significant have been found to show a similar trend in previous studies involving lexical decision latencies (Baayen, Milin, & Ramscar, 2016; Milin et al., subm.).

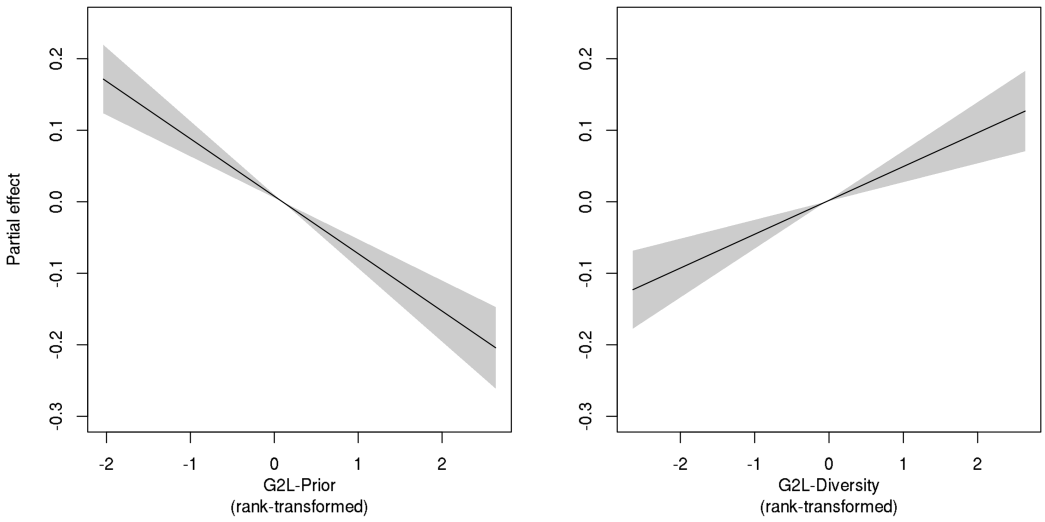

Figure 3. Smooths in the generalized additive model fitted to the lexical decision data, using discrimination-based predictors. Left panel: G2L-Prior, right panel: G2L-Diversity.

The GAM model with learning-based predictors (G2L-Prior and G2L-Diversity) fits the lexical decision times better, and achieves this better fit with fewer parameters, than the model reported in Section 3, which made use of form and lemma frequency as predictors in interaction with Word Difficulty (*ML*: 278.40 vs. 289.20; *AIC*: 337.32 vs. 347.04, number of parameters: 8 vs. 12; see Appendix B for details). This finding illustrates the importance of taking into account the simple but foundational principles of error-driven learning, as formalized in the Rescorla-Wagner rule.

In sum, the GAM models presented in Sections 4.1 and 4.2 reveal that TiMBL and NDL rely on different learning principles to account for the behavioural response data. TiMBL assigns higher probabilities to forms belonging to lemmas with letter trigraphs that yield more diverse activations. Those trigraphs belong to a rich exemplar space in the memory on which TiMBL bases its predictions. Given this, it would be expected that higher probabilities would result in shorter response latencies, yet NDL's G2L-Diversity was in fact positively correlated with RTs, indicating inhibition, i.e. slower recognition.

Recall, however, that TiMBL probabilities are intended to capture the likelihood of a form's occurrence in production. Under such circumstances, dense neighbourhoods might be desirable, and NDL diversity captures this trend indirectly. Conversely, in comprehension and in particular when making lexicality judgments, the diversity of trigraphs may well be hurtful as our results demonstrate (see also Baayen, Milin, & Ramscar, 2016; Milin et al., subm.).

5. Conclusion

In this position statement, we set out to demonstrate that the recent trend in usage-based linguistics to turn towards quantification and modelling does not divert attention from what really matters. Quite the contrary: reliance on cognitively plausible algorithms, in particular those based on principles of animal and human learning, advances understanding of how knowledge of language emerges from exposure to usage.

We have illustrated how a learning model with straightforward letter triplets as cues and word forms as outcomes generates predictors that outperform classical frequency measures. We note here that it is far from straightforward to generate predictions for lexicality decision times from memory-based learning. Information gain weighting or the modified value difference metric might be used to derive sets of nearest neighbours, but effect sizes of neighbourhood density measures in visual lexical decision are tiny (see Baayen, Milin, & Ramscar, 2016). More important is that memory-based learning assumes that exemplars are available in memory and can be straightforwardly accessed and compared, whereas naive discriminative learning addresses precisely the question of how the brain might access lexical information, and does this without making use of data structures from information science such as hash tables, linked lists, letter trees or information gain trees. Naive discriminative learning not only provides a step forward to answering this fundamental question, but, importantly, also shows that when this fundamental question is addressed, many phenomena reported in the processing literature receive simple yet powerful explanations (see, e.g., Baayen et al., 2011).

NDL resolves the two further concerns that we raised at the outset. Regression analysis works best and is best interpretable when predictors are orthogonal. In language, however, many predictors are highly correlated. For instance, frequent words tend to be polysemous, short, with high-frequency letter pairs, from dense neighbourhoods, and with high-frequency neighbours. Since the Rescorla-Wagner learning rule is applied locally, at the level of individual learning events, collinearity as a technical issue does not arise. Whenever predictors conspire, their joint effect will be absorbed by the learner. This error-driven learning approach makes it possible to gauge the impact of truly usage-based patterns and associated cognitively plausible abstractions.

Second, although the Rescorla-Wagner rule can be viewed as incremental regression (c.f., Widrow & Hoff, 1960), what sets it apart from standard regression is its sensitivity to order in learning. This sensitivity to order allows it to capture the effect of blocking (Kamin, 1969; Rescorla & Wagner, 1972), and to formulate precise predictions about the consequences of order for human learning (Ramscar et al., 2010; Arnon & Ramscar, 2012; Ellis, 2006a).

Last but not least, we have also shown how the discrimination measures that are derived from NDL's activation matrix, constructed on the basis of iterative learning, can be used to interpret the outcomes of other computational

algorithms, in this case MBL (as implemented in TiMBL). The excellent performance of NDL raises hope that we may have access to a "computational model that explains how grammar emerges from usage" (Baayen et al., 2013, p. 288). An approach couched in learning is ideally suited for testing the emergentist perspective on language knowledge that lies at the core of Cognitive Linguistic Theory.

The Naive Discrimination Learning framework provides valuable solutions to concerns that have been voiced in Cognitive Linguistic circles (Divjak, 2015) and are discussed further in this Special Issue (see in particular the contributions by Blumentahl-Dramé and Dąbrowska): it allows linguists to systematically explore the effect of different types of input to the system on the resulting representation and to model spoken or written language in a way that respects principles of human learning, while yielding predictions that are realistic and can be tested experimentally.

Taken together, these points provide a strong argument for adopting discrimination learning as an encompassing explanatory and computational framework that also allows empirical evidence to accrue and alter the way in which we think about language, during acquisition and in representation.

References

Anđel, M., Radanović, J., Feldman, L. B., & Milin, P. (2015). Processing of cognates in Croatian as L1 and German as L2. In *NetWordS* (pp. 182–186). Pisa, Italy.

Andresen, J. (1991). Skinner and Chomsky 30 years later. Or: The return of the repressed. *The Behavior Analyst*, *14*(1), 49–60.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*(3), 292–305.

Aronoff, M. (1994). *Morphology by itself: stems and inflectional classes*. Cambridge, MA: MIT Press.

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*, *11*(2), 295–328.

Baayen, R. H., Endresen, A., Janda, L. A., Makarova, A., & Nesset, T. (2013). Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics*, *37*(3), 253–291.

Baayen, R. H., Milin, P., Filipović Đurđević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–481.

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, to appear. Retrieved from: http://www.tandfonline.com/doi/abs/10.1080/02687038.2016.1147767

Beard, R. (1977). On the extent and nature of irregularity in the lexicon. *Lingua, 42*, 305-341.

Beard, R. (1981). On the question of lexical regularity. *Journal of Linguistics, 17*(1), 31–37.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389–414.

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). Amsterdam, The Netherlands: Royal Netherlands Academy of Arts and Sciences.

Chomsky, N. (1959). A review of BF Skinner's Verbal Behavior. *Language*, *35*(1), 26–58.

Cost, S., & Salzberg, S. (1993). A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, *10*(1), 57–78.

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

Dimitrijević, S. (2015). *Automatic parts of speech determination in a morphologically complex language* (Unpublished doctoral thesis). University of Novi Sad, Novi Sad, Serbia.

Divjak, D. (2015). Four challenges for usage-based linguistics. In *Change of Paradigms - New Paradoxes Recontextualizing Language and Linguistics*. Berlin, Germany: De Gruyter Mouton.

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2016). A functional theory of gender paradigms. In F. Kiefer, J. P. Blevins, & H. Bartos (Eds.),

Morphological Paradigms and Functions. Leiden, The Netherlands: Brill.

Eddington, D. (2000). Analogy and the dual-route model of morphology. *Lingua*, *110*(4), 281–298.

Ellis, N. C. (2006a). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164–194.

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

Ellis, N. C., & Larsen-Freeman, D. (2006). Language emergence: Implications for applied linguistics (Introduction to the special issue). *Applied Linguistics*, *27*(4), 558–589.

Evert, S., & Arppe, A. (2015). Some theoretical and experimental observerations on naive discriminative learning. In *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics*. Tübingen, Germany. Retrieved from: https://bibliographie.uni-tuebingen.de/xmlui/handle/10900/67200

Gallistel, C. R. (2008). Learning and representation. In J. Byrne (Ed.), *Learning and Memory: A Comprehensive Reference, Vol. IV* (pp. 227–242). New York, NY: Elsevier.

Ghirlanda, S. (2005). Retrospective revaluation as simple associative learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *31*(1), 107–111.

Gries, S. T., & Divjak, D. S. (2010). Quantitative approaches in usage-based cognitive semantics: Myths, erroneous assumptions, and a proposal. In D. Glynn & K. Fischer (Eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches* (pp. 333–354). Berlin, Germany: De Gruyter Mouton.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012), *138*(3), 415–422.

Jordanov, M., Nešković, M., & Milin, P. (2015). Feature-Label Order hypothesis and the Highlighting effect: Computational modeling (pp. 56–57). Presented at the *Empirical studies in psychology – EIP15*, Belgrade, Serbia: Faculty of Philosophy, University of Belgrade. Retrieved from: http://www.empirijskaistrazivanja.org/KnjigaRezimeaEng.aspx

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341–350.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. Campbell & R. Church (Eds.), *Punishment and aversive behavior* (pp. 279–296). New York, NY: Appleton-Century-Crofts.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633.

Keuleers, E., & Daelemans, W. (2007). Memory-Based Learning models of inflectional morphology: A methodological case-study. *Lingue E Linguaggio*, *6*(2), 151–174.

Klavan, J., & Divjak, D. (2016). The cognitive plausibility of statistical classification models: Comparing textual and behavioral evidence. *Folia*

Linguistica, to appear. Retrieved from: http://eprints.whiterose.ac.uk/95847/

Kostić, Đ. (1999). *Frekvencijski rečnik savremenog srpskog jezika [Frequency dictionary of contemporary Serbian language]*. Belgrade: Institute for Experimental Phonetics and Speech Pathology & Laboratory for Experimental Psychology, University of Belgrade.

Krott, A., Baayen, R. H., & Schreuder, R. (2001). Analogy in morphology: Modeling the choice of linking morphemes in Dutch. *Linguistics*, *39*(1), 51–94.

Lakoff, G. (1990). The Invariance Hypothesis: Is abstract reason based on image-schemas? *Cognitive Linguistics*, *1*(1), 39–74.

Lečić, D. (2016). *Morphological doublets in Croatian: A multi-methodological analysis* (Unpublished doctoral thesis). The University of Sheffield, Sheffield, UK.

MacCorquodale, K. (1970). On Chomsky's review of Skinner's Verbal behavior. *Journal of the Experimental Analysis of Behavior*, *13*(1), 83–99.

Milin, P., Divjak, D., & Baayen, R. H. (subm.). When meaningful cues remain meaningless: A learning perspective on individual differences in skilled reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (subm.). Discrimination in primed and unprimed lexical decision making. *PLOS One*.

Milin, P., Keuleers, E., & Filipović Đurđević, D. (2011). Allomorphic responses in Serbian pseudo-nouns as a result of analogical learning. *Acta Linguistica Hungarica*, *58*(1), 65–84.

Nešković, M., Jordanov, M., & Milin, P. (2015). Feature-Label Order hypothesis and the Highlighting effect: Experimental study (pp. 54–55). Presented at the *Empirical studies in psychology – EIP15*, Belgrade, Serbia: Faculty of Philosophy, University of Belgrade. Retrieved from: http://www.empirijskaistrazivanja.org/KnjigaRezimeaEng.aspx

Plaut, D. C., & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, *15*(4–5), 445–485.

Quinlan, J. R. (2014). *C 4.5 Programs for Machine Learning*. New York, NY: Elsevier.

R CoreTeam. (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria: R CoreTeam.

Ramscar, M., & Dye, M. (2011). Learning language from the input: Why innate constraints can't explain noun compounding. *Cognitive Psychology*, *62*(1), 1–40.

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*(6), 909–957.

Rescorla, R. A., & Wagner, A. 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). New York, NY: Appleton-Century-Crofts.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1. Cambridge, MA: MIT Press.

Seidenberg, M. S., & Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences*, *4*(9), 353–361.

Skinner, B. F. (1957). *Verbal behavior*. Acton, MA: Copley Publishing.

Skousen, R. (1989). *Analogical modeling of language*. Berlin, Germany: Springer Science & Business Media.

Theijssen, D., ten Bosch, L., Boves, L., Cranen, B., & van Halteren, H. (2013). Choosing alternatives: Using Bayesian Networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory*, *9*(2), 227–262.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *LREC* (pp. 2214–2218). Retrieved from: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Trimmer, P. C., McNamara, J. M., Houston, A. I., & Marshall, J. A. (2012). Does natural selection favour the Rescorla–Wagner rule? *Journal of Theoretical Biology*, *302*, 39–52.

Virués-Ortega, J. (2005). The case against B. F. Skinner 45 years later: An encounter with N. Chomsky. *The Behavior Analyst*, *29*(2), 243–251.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. In *1960 WESCON Convention Record*, Part IV (pp. 96–104).

Wood, S. (2006). *Generalized additive models: An introduction with R*. New York, NY: Chapman & Hall / CRC Press.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36.

Zeschel, A. (2008). Introduction: Usage-based approaches to language processing and representation. *Cognitive Linguistics*, *19*(3), 349–355.

APPENDIX A: A worked TiMBL example

For the example of “polagan”, TiMBL’s task would be to produce the form “polaganih” given “polagan”, relying on the set of closest neighbours as follows:

=,	**=,**	**=,**	**p,**	**o,**	**=,**	**l,**	**a,**	**=,**	**g,**	**a,**	**n,**	**201221,**	**?**
=,	=,	=,	p,	o,	=,	s,	e,	=,	b,	a,	n,	201221,	987654320ih
=,	=,	=,	p,	o,	=,	m,	e,	=,	š,	a,	n,	201221,	9876543210ih
=,	=,	=,	=,	o,	=,	t,	e,	=,	r,	a,	n,	201221,	9876543210ih
=,	=,	=,	i,	=,	z,	l,	a,	=,	g,	a,	n,	201221,	9876543210ih
=,	=,	p,	r,	i,	=,	k,	a,	=,	z,	a,	n,	201221,	9876543210ih
=,	=,	=,	=,	o,	=,	p,	a,	=,	s,	a,	n,	201221,	987654320ih
=,	=,	=,	n,	e,	=,	d,	a,	=,	v,	a,	n,	201221,	987654320ih

In this particular example, the most probable *inflectional class* is 9876543210ih, containing 4 out of 7 exemplars (p = 0.57). The novel form will thus be “polaganih”. Conversely, would the other class be the selected candidate (i.e. 987654320ih, with support of p = 0.43), TiMBL would produce the erroneous form “polagnih”.

APPENDIX B: Generalized additive mixed model specifications

B.1. *Generalized additive mixed model fitted to the lexical decision latencies for Serbian nominative masculine plural nouns, using lexical-distributional predictors. Reported are parametric coefficients (Part A) and non-linear terms (Part B) with effective degrees of freedom (edf), reference degrees of freedom (Ref.df), F and p values. (AIC = 347.04, -ML = 289.2, Adjusted R-sq. = 0.43)*

A. Parametric coefficients	**Estimate**	**Std. Error**	**t value**	**p-value**
Intercept	-1.405	0.031	-45.326	< 0.001
Word difficulty: easy	-0.055	0.021	-2.584	0.010
B. Smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(LemmaFreq, W.diff: hard)	1.000	1.000	27.276	< 0.001
s(LemmaFreq, W.diff: easy)	1.000	1.000	0.819	0.366
s(WordFreq, W.diff: hard)	1.000	1.000	0.758	0.384
s(WordFreq, W.diff: easy)	1.000	1.000	4.565	0.033
s(Participant)	37.210	38.000	48.763	< 0.001
s(Item)	88.670	103.000	6.719	< 0.001

B.2. *Generalized additive mixed model fitted to the lexical decision latencies for Serbian nominative masculine plural nouns, using discrimination learning predictors. Reported are parametric coefficients (Part A) and non-linear terms (Part B) with effective degrees of freedom (edf), reference degrees of freedom (Ref.df), F and p values. (AIC = 337.32, -ML = 278.4, Adjusted R-sq. = 0.43)*

A. Parametric coefficients	**Estimate**	**Std. Error**	**t value**	**p-value**
Intercept	-1.406	0.031	-45.408	< 0.001
Word difficulty: easy	-0.056	0.021	-2.616	0.009
B. Smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(G2L-Diversity)	1.000	1.000	20.356	< 0.001
s(G2L-Prior)	1.000	1.000	51.221	< 0.001
s(Participant)	37.160	38.000	45.908	< 0.001
s(Item)	85.350	100.000	6.077	< 0.001