

Naive discriminative learning and usage-based models of grammar

R. Harald Baayen

University of Alberta, Edmonton

In usage-based models, it is often assumed that linguistic units (simplex words, complex words, phrases, constructions) have their own representations in memory. The strength of entrenchment of these representations is viewed as proportional to frequency of use. Generalizations are typically understood as arising from connections between representations based on similarity or identity. Furthermore, higher-order cognitive schemas are posited to emerge from lower-order networked representations (see, e.g., Bybee, 2001; Langacker, 1991)).

In experimental research, frequency effects for complex units are often interpreted as evidence supporting the existence of representations in memory for these units. Recent years have seen the emergence of experimental evidence supporting frequency effects not only for complex words, but also for multi-word units (Bannard and Matthews, 2008; Arnon and Snider, 2010; Tremblay et al., 2011). However, as there are hundreds of millions of multiword units, these findings challenge the full-entry models assuming that all these n-grams would have their own representations in memory. First, such storage would lead to exponential amounts of redundancy. Second, the efficiency of the retrieval of individual n-grams from memory would decrease substantially, questioning the extent to which such representations would actually be a useful alternative processing route side by side with compositional processing. What is required is an inductive learning theory that captures the n-gram frequency effect without committing itself to a full-entry model.

What are the representations and links between representations that are to be targeted by such an inductive learning theory? Along with simple words, one might think of complex words (as morphological constructions, see Booij (2010)), and n-grams. However, frequency effects have been observed for n-grams that are not themselves phrases (Tremblay and Baayen, 2010). This suggests that frequency effects exist that are not tied to typical units in usage-based or construction-based models. Furthermore, evidence is accumulating that the word frequency effect is to a very large extent determined by contextual distributional information. This was first pointed out by McDonald and Shillcock (2001). In a recent study (Baayen, 2010), I showed that frequency in the sense of pure repetition in use explains only a small proportion of the variance in a chronometric measures of lexical processing. As a consequence, attributing degrees of entrenchment to frequency of use probably severely underestimates the extent to which entrenchment is driven by contextual co-occurrence.

In my presentation, I will introduce a computational model for comprehension in reading (Baayen et al., 2011) that may have potential as a computational implementation of inductive learning in usage-based approaches to grammar and language processing. This model, that we refer to as a ‘naive discriminative learner’, is based on well-established principles of discriminative learning as formalized by Wagner and Rescorla (1972), see Ramscar et al. (2010) for discriminative learning and child language acquisition. The model links (symbolic) orthographic units (letter unigrams and bigrams) to (symbolic) semantic units. The weights from the orthographic representations to the semantic representations are estimated using the equilibrium equations for the Rescorla-Wagner equations of Danks (2003), on the basis of the co-occurrence matrices of orthographic and semantic information in 11,172,554 two and three-word phrases from the British National Corpus, comprising 26,441,155 word tokens of 24710 monomorphemic words and compounds, derived and inflected words containing these monomorphemic words. The model generates simulated processing latencies that correlate significantly at the item level with observed latencies

as available in the English Lexicon Project for lexical decision.

Crucially, the model is extremely sparse in the representations it assumes, which are restricted at the orthographic level to letter unigrams and bigrams, and at the semantic level to the 6767 meanings of the monomorphemic (base) words. None of the 17943 complex words, and none of the 11,172,554 phrases to which the model was exposed during training received their own representations. Nevertheless, the model correctly predicts frequency effects for simple words, complex words, and n-grams. Furthermore, the model also captured correctly a lexical construction effect, namely, the inhibitory effect of relative entropy for case paradigms first reported by Milin et al. (2009) for Serbian. Milin and colleagues showed that words which make atypical use of nominal case endings, compared to the prototype of their inflectional class, incur a processing cost. The same processing cost emerges in the naive discriminative reader, even though the model is withheld any information about inflectional classes and case paradigms. Importantly, in this model, frequency in the sense of pure repetition, devoid of co-occurrence information, is not predictive at all.

Naive discriminative learning is similar to subsymbolic connectionist approaches with respect to the importance of context-sensitive learning. However, the model achieves generalization and prediction without requiring subsymbolic representations or hidden units. In fact, naive discriminative learning has been found to perform well as a statistical classifier (Baayen, 2011), which shows that subsymbolic multilayer networks are only a subclass of network models allowing generalization and prediction.

Naive discriminative learning also differs from interactive activation models. In interactive activation models, representations are typically assumed to have resting activation levels proportional to their frequency, mirroring the concept of entrenchment in usage based grammar. In these models, the interactive activation mechanism actually performs statistical maximum likelihood estimation ‘on-line’. Each time a token of a particular linguistic unit is processed, exactly the same statistical estimation is repeated. Supposedly the same holds for linguistic generalizations computed over representations and their links in usage-based network models such as sketched by Bybee (1988). In our discriminative learning model, by contrast, only a single forward pass of activation is required, as the statistically optimal solution is already implicit in the model’s connection weights. Processing costs are not assessed in how long it takes for the model to settle into a stable solution state, but in terms of how well a particular solution has been learned given past experience.

The model as currently implemented for reading approximates only one layer of complexity that is part of a much richer hierarchy of layers (see, e.g., Hawkins and Blakeslee, 2004; Hawkins and George, 2006; Numenta, 2010). Nevertheless, it may provide a useful framework for thinking about how usage affects grammar. Many of the insights of usage-based grammar that are traditionally framed in terms of interconnected representations can be re-expressed in the discriminative learning approach. However, naive discriminative learning as a computational theory of inductive learning may make certain kinds of representations superfluous. In morphology, for instance, separate representations for the *forms* of bound morphemes, simple words, complex words, and n-grams are redundant. This has as advantage that processing effects for ‘intermediate’ categories such as phonaesthemes, as documented by Bergen (2004), can be straightforwardly accommodated.

Assuming that there is a grain of truth to the naive discriminative learning approach, it will be crucial for modeling more complex aspects of language structure and processing to clarify how discriminative learning networks might work together in hierarchically organized layers.

References

- Arnon, I. and Snider, N. (2010). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Journal of Memory and Language*, 62:67–82.
- Baayen, R. (2011). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*, submitted.
- Baayen, R. H. (2010). The directed compound graph of english. an exploration of lexical connectivity and its processing consequences. In Olson, S., editor, *New impulses in word-formation (Linguistische Berichte Sonderheft 17)*, pages 383–402. Buske, Hamburg.
- Baayen, R. H., Milin, P., Filipović Durđević, D., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* (in press).
- Bannard, C. and Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science*, 19(3):241.
- Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, 80:290–311.
- Booij, G. E. (2010). *Construction Morphology*. Oxford University Press, Oxford.
- Bybee, J. L. (1988). Morphology as lexical organization. In Hammond, M. and Noonan, M., editors, *Theoretical Morphology: Approaches in Modern Linguistics*, pages 119–141. Academic Press, London.
- Bybee, J. L. (2001). *Phonology and language use*. Cambridge University Press, Cambridge.
- Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, 47(2):109–121.
- Hawkins, J. and Blakeslee, S. (2004). *On intelligence*. Henry Holt and Company, New York.
- Hawkins, J. and George, D. (2006). Hierarchical temporal memory. Concepts, theory and terminology. Numenta Technology, <http://www.numenta.com/technology.php>.
- Langacker, R. W. (1991). *Foundations of Cognitive Grammar. Vol. 2, Descriptive Application*. Stanford University Press, Stanford.
- McDonald, S. and Shillcock, R. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, 44:295–323.
- Milin, P., Filipović Durđević, D., and Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, 60(1):50–64.
- Numenta (2010). Hierarchical temporal memory including HTM cortical learning algorithms. Version 0.1.1, November 23, 2010.
- Ramscar, M., Yarlett, D., Dye, M., Denny, K., and Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, 34(6):909–957.

- Tremblay, A. and 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 Wood, D., editor, *Perspectives on formulaic language. Acquisition and communication*, pages 151–173.
- Tremblay, A., Derwing, B., Libben, G., and Westbury, C. (2011). Processing advantages of lexical bundles: evidence from self-paced reading and sentence recall tasks. *Language Learning*, in press.
- Wagner, A. and Rescorla, R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical Conditioning II*, pages 64–99. Appleton-Century-Crofts.