Resource requirements for neo-generative modeling in (psycho)linguistics

*R. Harald Baayen**

Current linguistic theories have in common the foundational assumption that symbolic representations, and combinatorial operations defined over these representations, are at the heart of the language engine. Connectionist theories have been proposed in which co-occurrence learning takes place in systems in which symbolic rules and representations are replaced by subsymbolic feature vectors in artificial neural networks. The connectionist approach, however, has not had much impact in linguistics, and in psychology, Bayesian models, again defined over symbolic representations, have become popular, replacing subsymbolic connectionist modeling.

In my presentation, I will introduce a simple computational model, the Naive Discriminative Reader, which is based on well-established principles of human learning as formalized in the Rescorla-Wagner equations. In this model, orthographic representations (letter unigrams and bigrams) are simply associated with word meanings through weighted links. The model, described in detail in Baayen, Milin, Filipovic Durdevic, Hendrix, and Marelli (in press), captures a wide range of phenomena in visual comprehension as gauged by the lexical decision task, including word frequency effects, constituent frequency effects, family size effects, paradigmatic entropy effects, and phrasal *n*-gram frequency effects. Crucially, the model is trained not on isolated words but on

*University of Alberta

words in small *n*-grams. Even though the model has no representations for morphemes, word forms, or *n*-gram forms, it nevertheless captures a great many factors that in current theories of lexical processing have been interpreted as diagnostics of lexical form representations. The model shares with connectionist approaches that error-driven co-occurrence learning is essential to understanding language, but it provides a simpler architecture that obviates the need for subsymbolic representations, hidden layers, and complex and biologically implausible algorithms such as backpropagation.

Two aspects of this model are relevant to the topic of this QITL4 workshop on lexical resources. First, considerable progress has been made in recent years in finding corpora capturing the linguistic register that optimizes the variance explained by the word frequency effect in chronometric tasks such as lexical decision (cf. Keuleers, Brysbaert, & New, 2011). In psychology, with the exception of the now unfashionable connectionist models, the word frequency effect is explained in terms of a property of a lexical entry, such as its supposed resting activation level (in interactive activation models) or its verification threshold (in weaver++). However, word frequency is only one of a tangled mess of highly collinear predictors, including syntactic co-occurrence frequencies. An estimate of frequency as pure repetition explains only 4% of the variance in lexical decision. Interestingly, pure repetition has no explanatory value at all according to the Naive Discriminative Reader. In this model, the frequency effect is an epiphenomenon of many co-occurrence frequencies. If this model is on the right track, current resources documenting simple isolated word frequencies are going to be insufficient for future research. This is, of course, also clear from the new wave of experimental studies documenting phrasal frequency effects.

Second, with the advent of statistical models that make it possible to model wiggly interaction surfaces, the analyst is confronted with complex fitted surfaces that resist explanation by current verbal models of lexical processing. The top panels of Figure 1.1 present an example of empirical regression surfaces for lexical de-

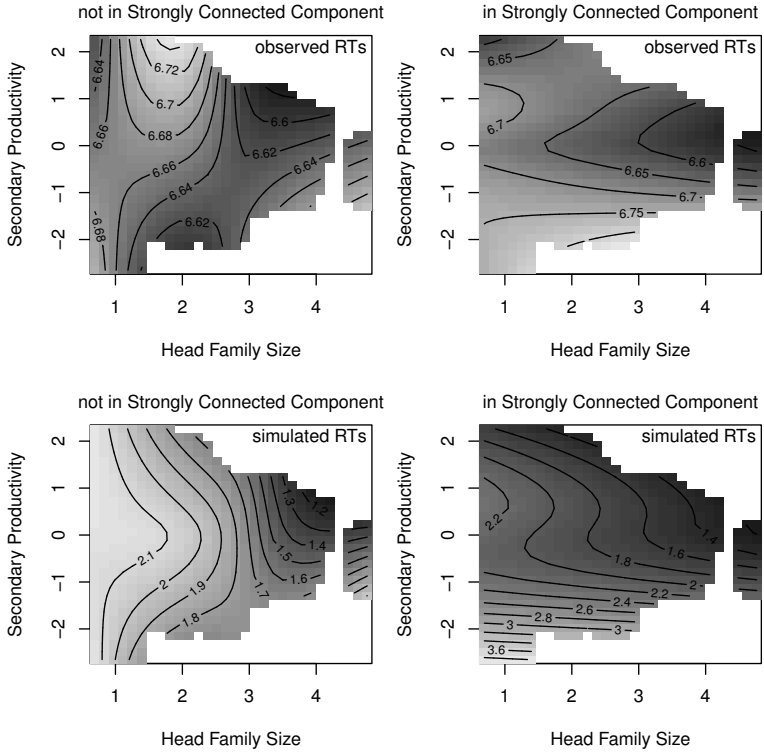

Figure 1.1

cision in English. The lower panels show the surfaces fitted by the naive discriminative reader. The fits are not perfect, but some of the patterns in the empirical data return in the predicted surfaces. As more sophisticated statistical methods are applied, the true complexity of even simple resources such as data banks of lexical decision latencies will become more visible. Such complex patterns will require 'neo-generative models', computational models that 'generate' predictions from realistic input using processing principles reflecting the forces operative in the complex dynamic system

that is language. In order to build such generative models, we will need extremely rich lexical resources providing information about the contextually appropriate lexical meanings, information status, the fine phonetic detail of the speech signal, gestures, facial expressions, and emotion. If the Naive Discriminative Reader is on the right track in emphasizing the importance of discriminative learning of rich arrays of co-occurring information, current resources are only a first step in the direction of the much more comprehensive resources that we will need to understand language and language processing.

References

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, *5*(3), 436-461.

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

Keuleers, E., Brysbaert, M., & New, B. (2011). An evaluation of the google books ngrams for psycholinguistic research. In J. Heister, E. Pohl, & K.-M. Würzner (Eds.), *Lexical Resources in Psycholinguistic Research.* vol. 3 of *Potsdam Cognitive Science Series.* Universitätsverlag Potsdam.