Statistical language models reveal cross-linguistic differences in processing difficulty

Stephan C. Meylan & Thomas L. Griffiths (University of California, Berkeley)
smeylan@berkeley.edu

The long-standing *compensation hypothesis* suggests that while languages vary in the complexity of subsystems (syntax, morphology, phonology, etc.), they are approximately equally matched in overall complexity [1]. While this hypothesis is highly controversial and difficult to evaluate empirically, we suggest that there is a substantive, testable cognitive correlate: all languages are of equal *processing difficulty*. Following recent work, we use surprisal—in-context predictability—as an estimate of processing difficulty [2]. A cognitive correlate of the compensation hypothesis would thus be that all languages are equally predictable, as measured by average surprisal.

Experimental elicitation of surprisal estimates at scale is challenging in a single language [3]; measuring it for a large selection of languages is harder still. Following the approach of [4] , we use a model-based method for estimating surprisal across languages using a Long Short-Term Memory (LSTM) recurrent neural network—a state-of-the-art neural language model [5]—and a large near-parallel corpus. Like probabilistic context-free grammars (PCFGs), LSTMs can learn long-distance dependencies [6], but can be easily trained from novel corpora without phrase structure annotation. Consequently, LSTMs may be able to learn structural regularities in typologically-diverse languages that *N*-gram models overlook, reducing model-based bias.

To match discursive and semantic content to the degree possible across languages, we construct a corpus using the tokenized OPUS 2016 Subtitle Corpus [7]. While movies may deviate from naturalistic language use, their statistical properties are strong predictors of psycholinguistic behaviors [8]. First, we find the subset of movies with transcripts in the 18 languages with >50m tokens (4.4k movies, 17-22m tokens per language). From this matched corpus we then generate 45 smaller cross-linguistic datasets with non-overlapping training, validation, and test sets (150, 30, and 30 movies respectively; ~1m, ~200k, and ~200k tokens).

For each combination of language and subset of movies, we estimate the parameters of several models on the training set, then measure the average surprisal in the test set under the estimated models. The LSTM, implemented using Google's TensorFlow library [9], has two layers with 650 hidden nodes each. We also estimate average surprisal under *N*-gram models (*N*=1-5) with modified Kneser-Ney smoothing [10] obtained with the SRILM toolkit [11].

We excluded Turkish from the analysis because of a large number of out-of-dictionary tokens (>30%). Among the remaining 17 languages, average surprisal is clustered across languages at 7.59 bits (SD=.38) for the best-performing *N*-gram model (order=3, Figure 1) and 5.99 for the LSTMs (SD=.27; Figure 2). The lower surprisal values for the LSTMs indicate that they are better at predicting material in these texts than *N*-gram models. LSTMs exhibit decreased cross-linguistic variance in average surprisal compared to the *N*-gram models. However, against the compensation hypothesis, within-language variance is substantially smaller than between-language variance ($F(16,765)=798.4$, $p < .0001$), suggesting numerically-small but significant differences in predictability across languages (which may be accounted for in phonology or morphology). Finally, pairwise differences between mean surprisal estimates is small for phylogenetically-similar languages (Figure 4), when phylogenetic distance is computed from weighted sequence alignment over the lexicon [12].

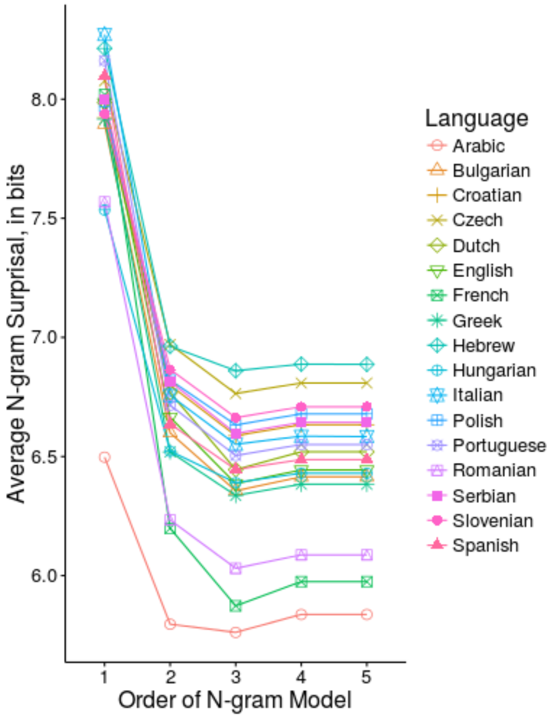

Figure 1. Average surprisal in 17 languages across Kneser-Ney smoothed *N*-gram models of orders 1-5. The lowest average surprisal is found for corpora of this scale at *N*=3. The same legend is used for Figures 3 and 4.

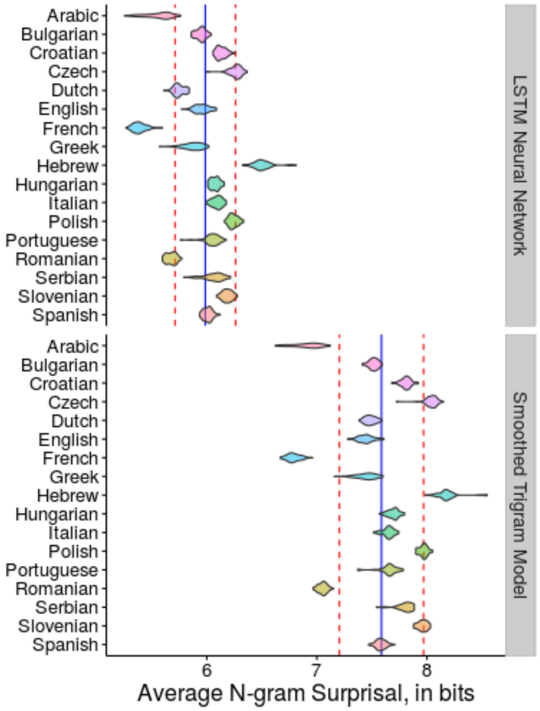

Figure 2. LSTM-based and trigram-based estimates of average surprisal show significant differences between languages. Mean is marked in blue and standard deviation in red. Each violin represents 45 matched (train / validation / test) sets.

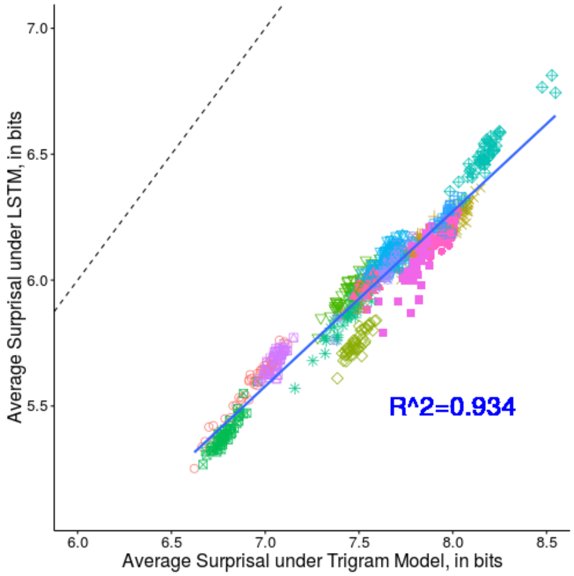

Figure 3: LSTM and *N*-gram estimates are strongly correlated. Each point is a matched corpus and language.

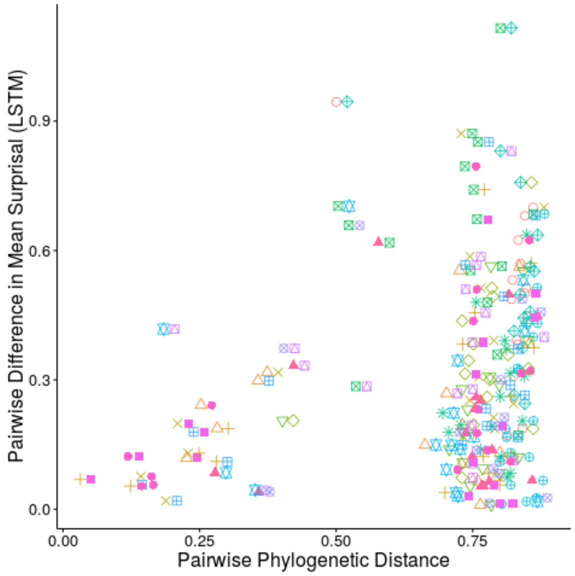

Figure 4. Differences in mean surprisal estimates are low for phylogenetically-related pairs of languages per Jäger (2015).

[1] Siewierska (1998); Kroch (2001); Gil (2008) [2] Levy (2008); Piantadosi et al. (2011) [3] Smith & Levy (2013) [4] Hahn and Keller (2016) [5] Zaremba et al. (2015) [6] Hochreiter & Schmidhuber (1997) [7] Lison & Tiedemann (2016) [8] Brysbaert et al. (2011) [9] Abadi et al. (2015) [10] Chen and Goodman (1998) [11] Stolcke (2002) [12] Jäger (2015)