Northwestern University

Low-level language statistics affect reading times independently of surprisal

Adam Goodkind
Klinton Bicknell
a.goodkind@u.northwestern.edu

1. Motivation

- Surprisal theory (Hale, 2001; Levy, 2008) states: Processing difficulty for a word is proportional to its log probability given the full prior context

  Surprisal: $\log p(w_n|w_1,w_2,\ldots,w_{n-1})$

  - Surprisal also adds a **unifying framework**: Garden paths, ambiguity, predictability, etc., affect reading times via affecting probability in context, i.e., surprisal
  - Surprisal acts as a **causal bottleneck**: Holding constant a word's log probability given its full context (surprisal), no other linguistic factors should yield additional predictive power for reading times
- But one low-level statistic, word frequency, has been shown to influence sentence processing above and beyond surprisal: when two words are equally likely in context, the more frequent one will still be read faster. Why?

  Word Frequency (unigram): $p(w_n)$ **Bigram:** $p(w_n|w_{n-1})$

  - **Option 1:** Word frequency effects have an idiosyncratic explanation, perhaps reflecting the mechanics of word retrieval independent of sentence processing
  - **Option 2:** There is special sensitivity to low-level statistics in sentence processing broadly, perhaps because comprehenders have substantial uncertainty of the more distant linguistic context, and make predictions based largely on the local (less uncertain) context, as suggested by noisy-channel surprisal (Futrell & Levy, 2017)

Goal: Determine whether there is special sensitivity to low-level statistics broadly by testing word bigram probability.

2. Prior Bigram & Surprisal Studies

Prior work

- Some studies have shown processing effects of bigrams without controlling for surprisal (McDonald & Shillcock, 2003a, 2003b; Arnon & Snider, 2010)
  - **Problem**: The reported effects could actually just reflect surprisal
- Other studies already showed bigram effects above and beyond surprisal (Demberg & Keller, 2008; Fossum & Levy, 2012; Mitchell *et al.*, 2010)
  - **Problem**: These studies computed surprisal using a probabilistic context-free grammar (PCFG)
    - A PCFG is not a great model of a word's probability in context
    - In particular, PCFGs cannot capture relationships between words that are likely to occur together (frequent phrases)
    - Bigram probabilities can fill exactly this hole!
    - Thus, simultaneous PCFG surprisal and bigram effects could have just reflected different pieces of actual surprisal

Our Improvements

- We **compute surprisal using a state-of-the-art language model** that does capture relationships between words such as frequent phrases.
- We confirm via language model analysis that **our trained language model effectively captures relevant bigram information**, and thus is not helped in predicting words by a bigram model
- Finally, we show that there are **still significant effects of word bigram probability for reading times (gaze duration) above and beyond the predictions made by this state-of-the-art surprisal model.**

3. Experiment 1: Perplexity

Goal

- Investigate whether low-level statistics improve accuracy of a surprisal-based language model

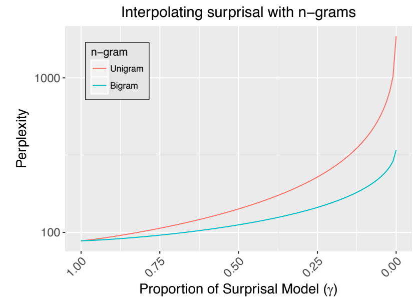
Language Model

- Our language model was trained on Google's One Billion Word Corpus
- Created by interpolating an LSTM model with a 5-gram model, where each model is proportionally weighted to create a blended probability.

$$p_{\text{interp}}(w_n|w_1^{n-1}) = \gamma p_1(w_n|w_1^{n-1}) + (1-\gamma)p_2(w_n|w_1^{n-1})$$

Methods

- We performed a grid search using our language model interpolated with different weights of a unigram/bigram model to find optimal perplexity of the Dundee Corpus (Kennedy *et al.*, 2003)
  - For each measurement we incremented γ in the equation above to take complementary weightings from the surprisal and *n*-gram models
  - Note: We also tested a balanced model (γ=0.5) to ensure generalizability. All results paralleled those reported here of the optimal blend.



Results

- Adding any proportion of low-level statistics did not improve perplexity, unlike prior studies using a PCFG.

4. Experiment 2: Gaze Duration

Goal

- Since n-grams do not improve a surprisal-based language model, determine if unigrams and bigrams improve predictions of gaze duration when controlling for surprisal

Methods

- Processing data came from the English portion of the Dundee Corpus.
- Used a mixed effects regression model that included unigram and bigram statistics along with surprisal, whether the previous word was fixated (π_n) and the word sequence number (ν_n)

$$\text{gaze duration} \sim \text{surprisal}_n + \text{surprisal}_{n\text{-}1} + \text{freq}_n + \text{freq}_{n\text{-}1} + \text{bigram}_n + \text{bigram}_{n\text{-}1} + \text{freq}_n : \text{length}_n + \text{freq}_{n\text{-}1} : \text{length}_{n\text{-}1} + (\text{freq}_n + \text{freq}_{n\text{-}1} + \text{bigram}_n + \text{bigram}_{n\text{-}1} \,||\, \text{subject}) + \pi_n + \nu_n$$

- The predictors of interest for the model were the *n*-grams
- We also tested a generalized additive mixed-effect model (GAMM) to see if non-linear surprisal changed results.
  - GAMMs used non-linear smoothing splines for all controlling predictors
  - Only the predictors of interest were kept linear

Predictor	LME $\hat{\beta}$ (ms)		GAMM $\hat{\beta}$ (ms)	
log frequency w_n	-11.58	$p < 0.01$	-10.42	$p < 0.01$
log bigram w_n	-1.49	$p < 0.05$	-1.13	$p = 0.09$
log frequency w_{n-1}	-2.16	$p < 0.01$	-2.83	$p < 0.001$
log bigram w_{n-1}	-0.74	$p = 0.08$	-1.09	$p < 0.02$

Results

- The current and prior word frequency have a significant effect on gaze duration, even when surprisal is taken into account.
- Bigrams are significant or approaching significance, although questions remain.
- These results are consistent when allowing for non-linear surprisal in GAMMs.

5. Conclusion

- Frequency is not special: another low-level statistic (word bigram probability) also affects reading time in a way not explained by classic surprisal
- So what's the explanation?
  - Could be noisy-channel surprisal (Futrell & Levy, 2017)
  - Could also be low-level perceptual learning
    - Perhaps the visual system has adapted to recognizing words in particular local environments (e.g., in frequent bigrams)
- Motivates further work on how comprehenders form predictions from context

Acknowledgements

This research was supported by NSF Award 1734217 (Bicknell).

References

Inbal Arnon and Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. Journal of memory and language 62(1):67–82.
Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition 109(2):193–210.
Victoria Fossum and Roger Levy. 2012. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In Proceedings of the CMCL. ACL, pages 61–69.
Richard Futrell and Roger Levy. 2017. Noisy-context surprisal as a human sentence processing cost model. In Proceedings of the EACL. pages 688–698.
John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In NAACL-HLT. ACL, pages 1–8.
Alan Kennedy, Robin Hill, and Joel Pynte. 2003. The Dundee corpus. In Proceedings of the 12th European conference on eye movement.
Roger Levy. 2008. Expectation-based syntactic comprehension. Cognition 106(3):1126–1177
Scott A McDonald and Richard C Shillcock. 2003a. Eye movements reveal the on-line computation of lexical probabilities during reading. Psychological science 14(6):648–652.
---. 2003b. Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. Vision research 43(16):1735–1751.
Jeff Mitchell, Mirella Lapata, Vera Demberg, and Frank Keller. 2010. Syntactic and semantic factors in processing difficulty: An integrated measure. In Proceedings of the ACL