Word forms—not just their lengths—are optimized for efficient communication

Stephan C. Meylan & Thomas L. Griffiths

{smeylan, tom_griffiths}@berkeley.edu

Dept. of Psychology, University of California, Berkeley

Computational Cognitive Science Lab

Zipf (1932): High frequency words are short

Piantadosi et al. (2011): More predictable words (estimated from average predicatability in sentential contexts) are shorter

Current Work: Word length is one aspect of a more general property of words: sound sequences are more or less probable in the language. More generally: less predictable words consist of less probable letter or sound sequences.

Example: "less" will appear in more predictable sentential contexts than "act"

Yesterday I went to the store

To: highly predictable Store: much less predictable

$I_L(to)$ = - log P(to | I went) = 1.12

$I_L(store)$ = - log P(store | to the) = 3.60

$I_s(to)$ = - log (P(t) × P(o|t) × P($|to)) = 3.42

$I_s(store)$ = - log (P(s) × P(s|t) × P(o|st) ...) = 5.43

$ represents word end. Surprisal values are in bans. Does this relationship hold for phonemes and letters?

Piantadosi et al. (2011)

Information content is a positive linear function of word lenth if we assume uniform character probabilities

$$IC_{WordLength}(w) = -\sum_{l \in w} log\ P\left(\frac{1}{|S|}\right)$$

Current Work: Sublexical Surprisal

Information content is computed using conditional probabilities given the type inventory

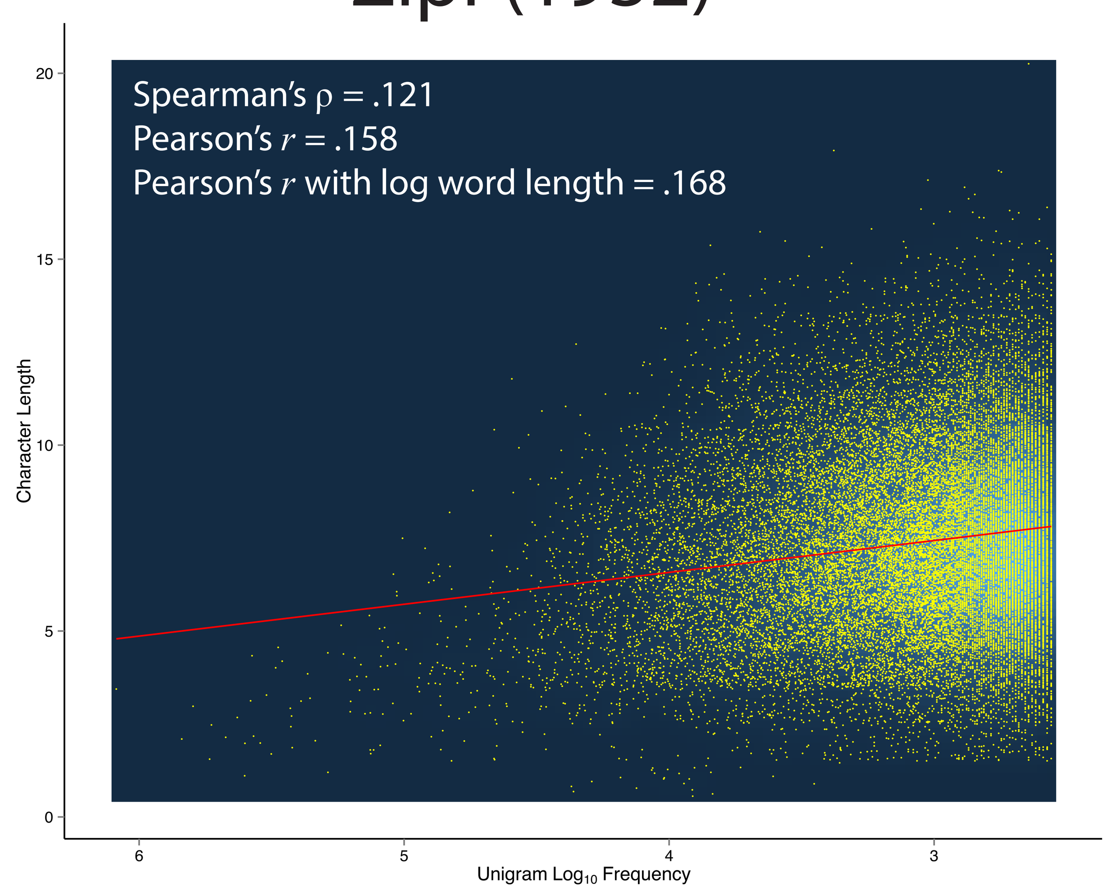
$$IC_{SublexicalSurprisal}(w) = -\sum_{l \in w} log\ P(l|c)$$

Method: Following Piantadosi et al. (2011) , we take the top 25,000 most frequent words from the BNC that also appear in the English OPUS Subtitle corpus. Phone probabilities are calculated using sequences up to length 5, with modified Kneser-Ney smoothing for phone sequences of order 3 and above.

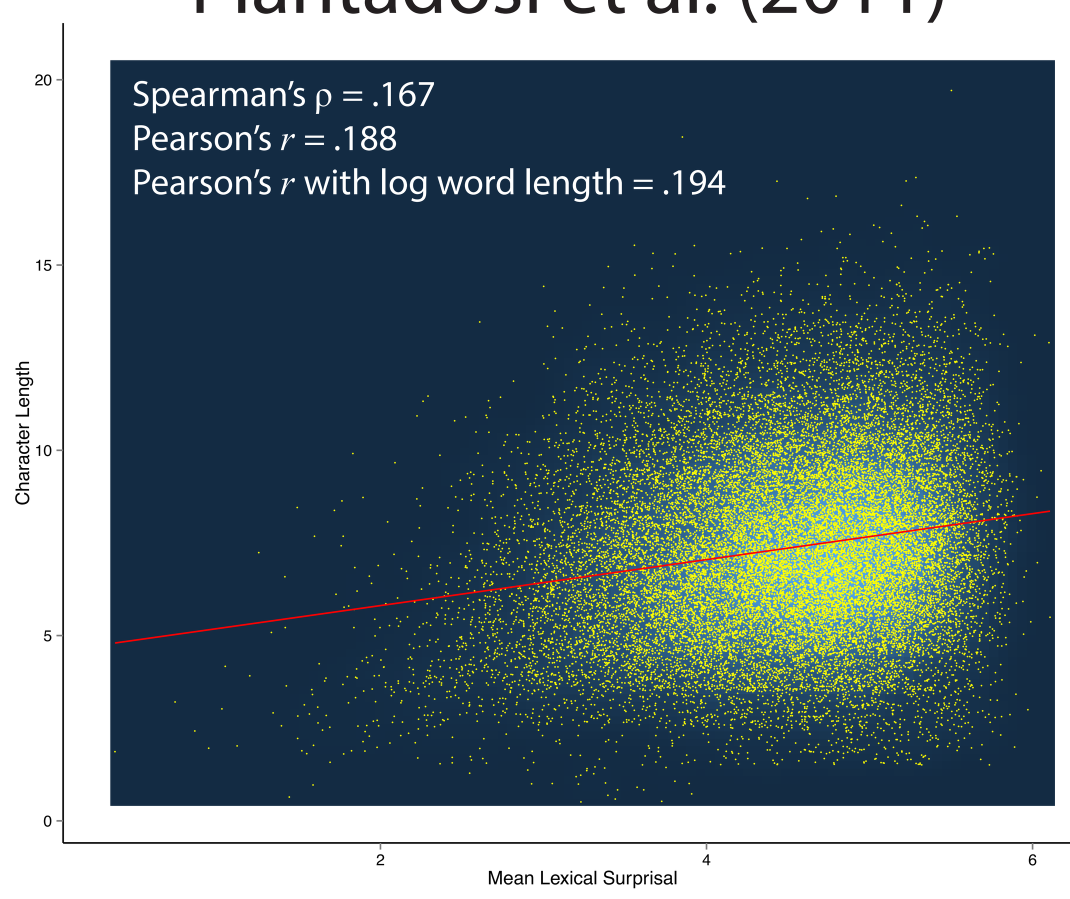
Baseline: We computed these same correlations for a set of 50 artificial languages which maintained all properties of the BNC, but permuted the order of the characters within each word. Languages of this type yielded uniformly low correlations (mean ρ= .067, SD=.006).

Additional Notes: Character lengths are jittered in left two plots. Right boundary in first graph is from taking 25,000 most frequent items. Word probabilities were computed using an end character. Spearman's ρ with sublexical surprisal calculated over phone transitions is .253

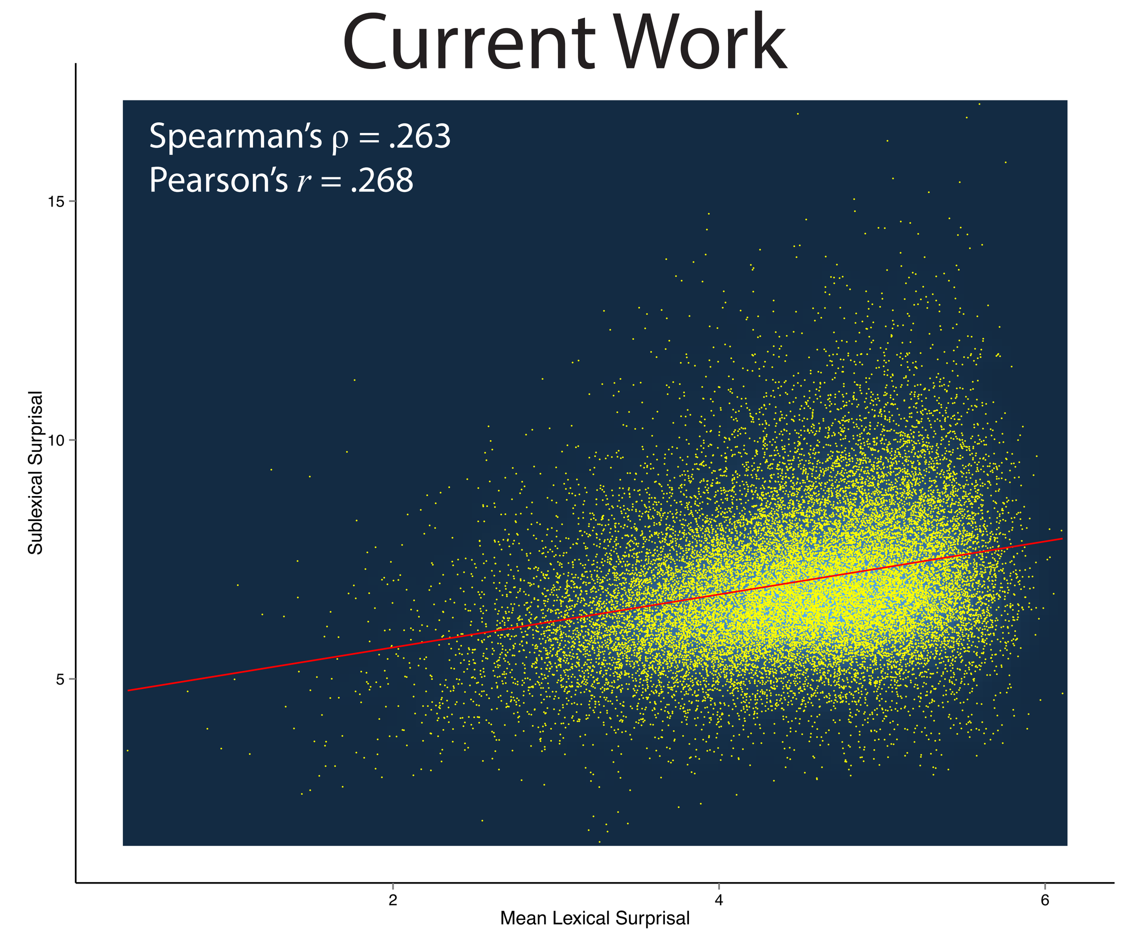
Zipf (1932)



Piantadosi et al. (2011)



Current Work



As Neighborhood Density Metric

Sublexical surprisal is strongly related to existing measures of phonological and orthographic neighborhood density. It shares many formal similarities with frequency-weighted neighborhood probability (Luce and Pisoni, 1998). Differences are as follows:

1) **Phone-by-phone estimation**: information content is the sum of surprisal values computed after receiving each successive phoneme or letter, rather than the entire string

2) **Bayesian treatment of likelihood:** though in practice confusion matrices are largely symmetric

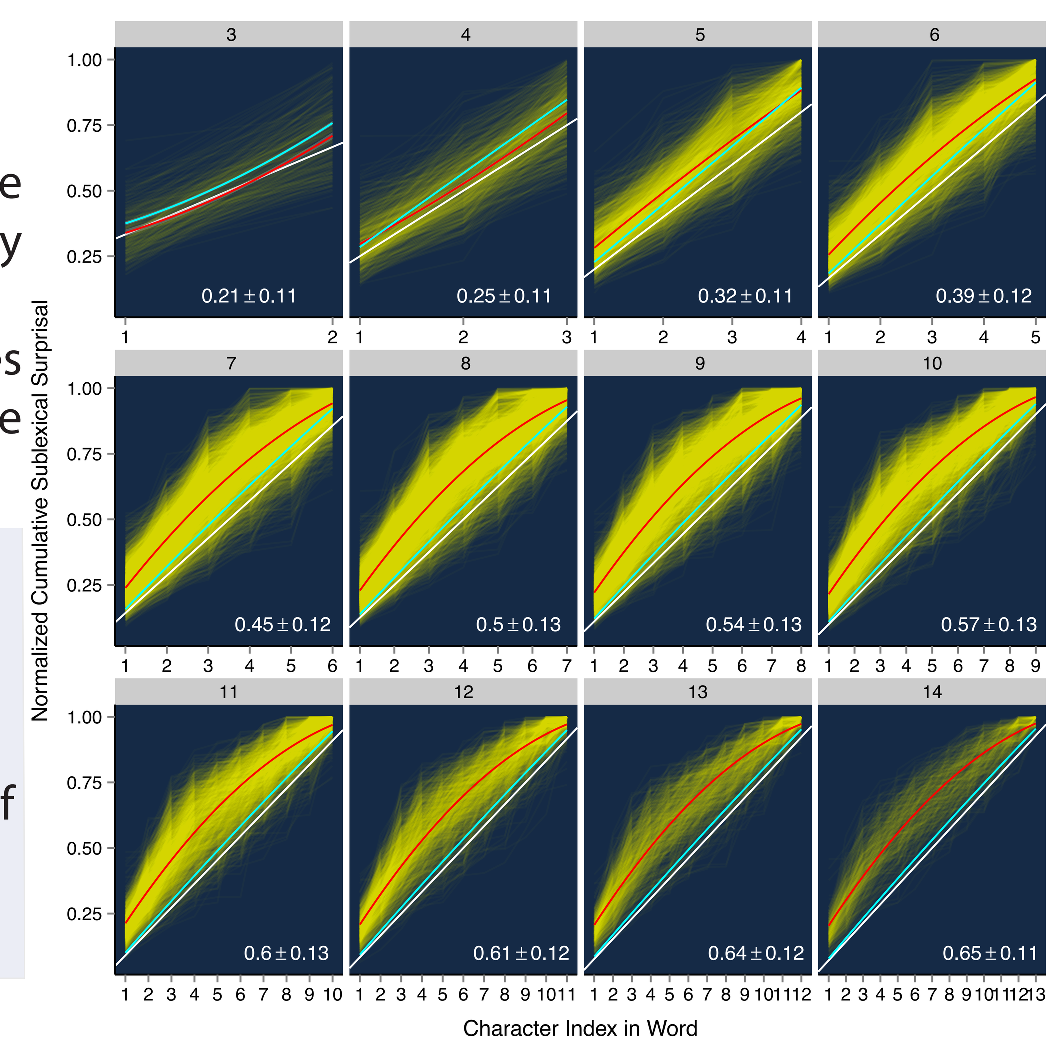
3) **Uniform confusion probabilities:** Can perceptual confusability be isolated?

4) **Context-based estimation**: probability of a phoneme or letter is estimated as the probability of that particular continuation, using the type counts in a large language sample

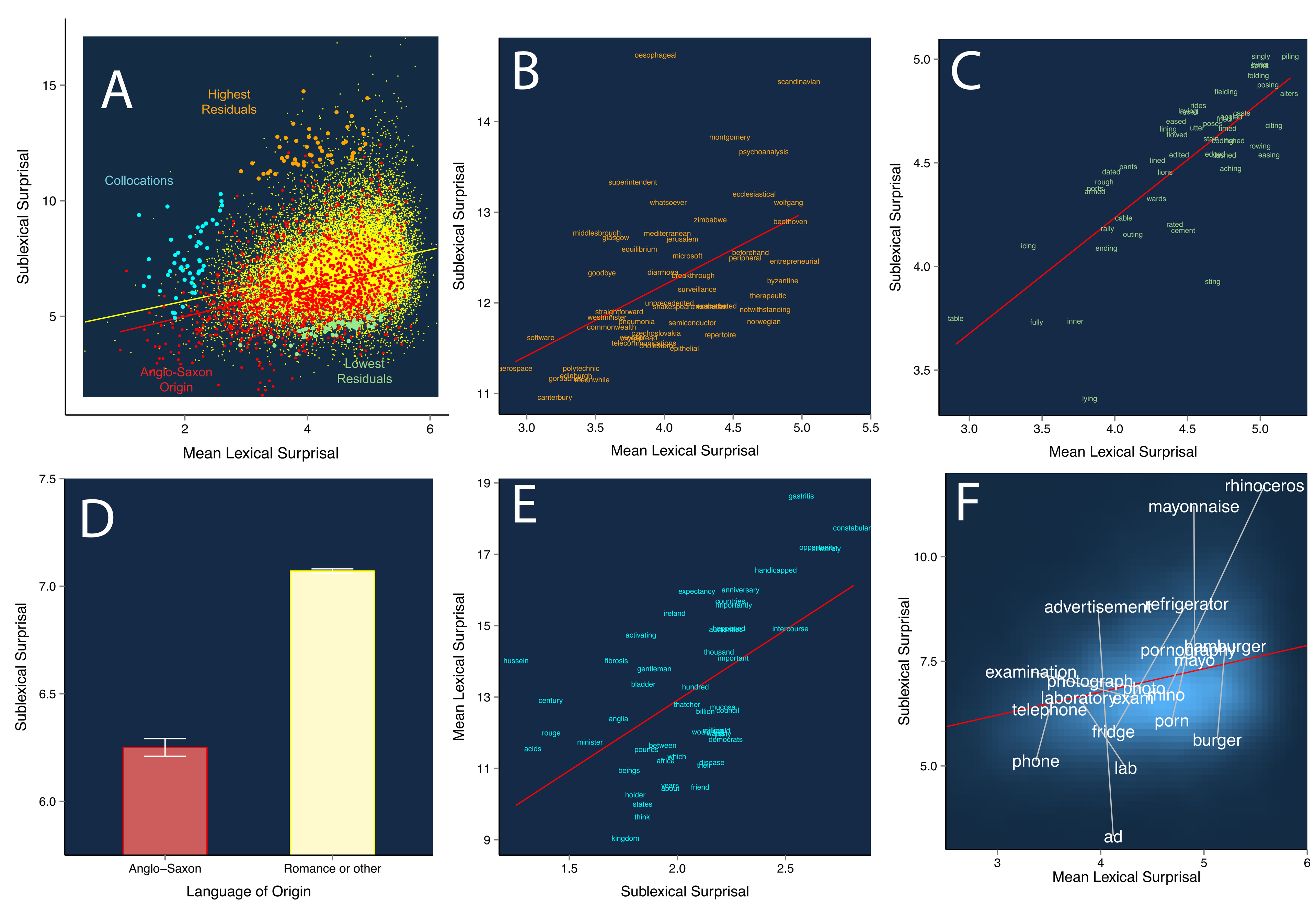
Sublexical UID

Following Maurits et al. (2010) we calculate divergence from uniform information density for letter sequences in the language.

Right: yellow lines represent words, white lines the uniform information rate, and red the best-fit quadratic equation.



Historical Change



A: Four groups of interest: Words with highest and lowest sublexical surprisal given their predictability, words with highest predictability given their sublexical surprisal, and terms of Anglo-Saxon origin.

B: Highest residual words have rich morphology and tend to have Latin roots. Many have shorter variants

C: Words of Anglo-Saxon origin have lower sublexical surprisal than Romance or other origin words

D: Highest negative residual words

E: Highest residuals for lexical suprisal predicted from sublexical surprisal include second words in two-word collocations

F: Pairs of words from Mahowald et al. (2012)

Notes: Filtered to omit names, toponyms, words of length < 4, and log unigram probability < -5

Next Steps

- Replicate with Google datasets and extend to other languages
- Rational use of context: use preceding context less when estimate is less peaked
- Revisit token-based model: highly informative, but circularity problem
- Relate to historical lexical predictability

References

Zipf, G.K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge: Harvard University Press.

Piantadosi, S.T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. *PNAS* 108.

Chen, S.F. and Goodman, . (1999) An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13 (4), 359–394.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing*, 19(1), 1–36.

Maurits, L., Perfors, A., Navarro, D. (2010) Why are some word orders more common than others? A uniform information density account. *NIPS*, 23, 1585-1593.

Mahowald, K., Fedorenko, E., Piantadosi, S.T., and Gibson, E. (2013). Info/information theory: speakers choose shorter words in predictive contexts. *Cognition*, 126, 313-318.