Word forms are optimized for efficient communication

Stephan Meylan
Computational Cognitive Science Lab (Tom Griffiths)

13 February 2017

Berkeley NLP Seminar

Cross-Linguistic Regularities

- Languages display remarkable regularities
  - Greenberg's universals (1963), e.g. gender categories in nouns re-used in pronouns
  - Argument structure regularities (Fedzechkina et al., 2012)
  - Minimized dependency lengths (Futrell et al., 2015)
- Potential explanations (not mutually-exclusive!):
  - Innate biases ("Universal Grammar")
  - Shared origin
  - Communicative efficiency / robustness
  - Learnability
- Language regularities are important for linguistics, psychology, neuroscience, machine learning

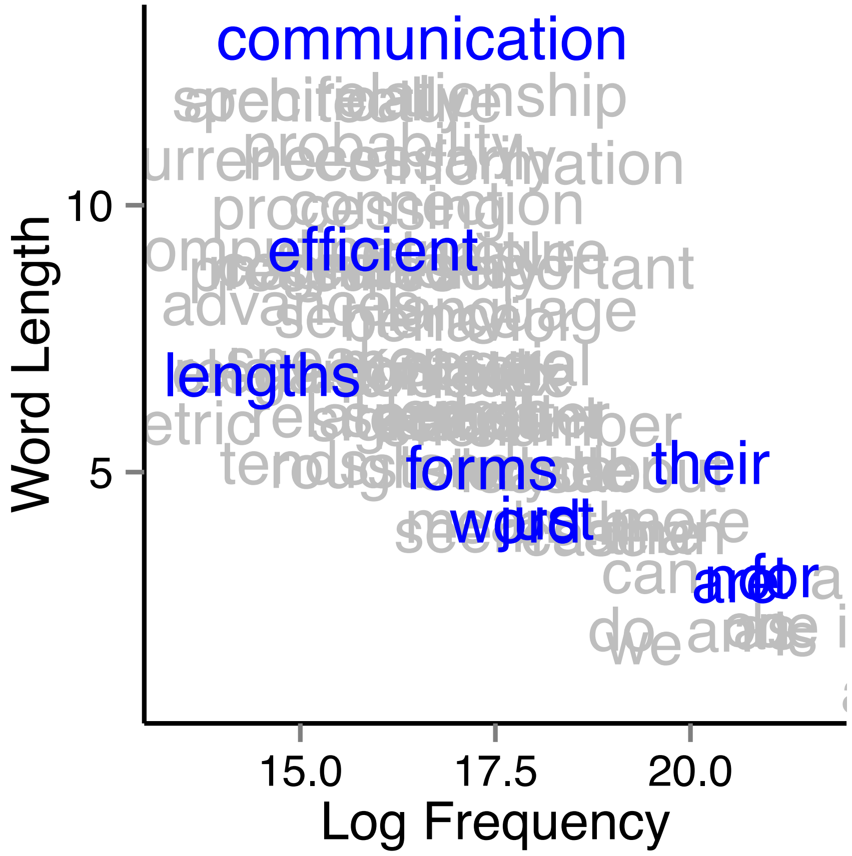
"Stable engineering solutions" (Evans and Levinson, 2009)

Today’s Talk

- Today: revisit a long-known (80+ y.o.) relationship in language: inverse relationship between a word’s length and the frequency of its use (Zipf, 1935)
- Show some evidence why this regularity might emerge as a “stable engineering solution”
- Statistical language modeling using formalisms from NLP
  - Inference is at the heart of language processing!
  - NLP = a set of hypotheses about inference in language

Zipf (1935) and Word Length

- “the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences”
- Motivated by “Principle of Least Effort”
- Most frequent code should be the shortest (variable length encoding)



Generalizing Zipf's Second Law

- Not length, per se, but *distinctiveness* of the audio signal, which varies with frequency
- Successful recognition depends on the number and strength of competitors
- Distinctive = low probability under the prior distribution on phone transitions; string is diagnostic of a particular word
- Basic logic: A low frequency word needs a distinctive word form, otherwise it loses to higher-frequency competitors which are partially consistent with the observed phone sequence

Outline

- Rational Model of Distinctiveness
- Relationship to Spoken Word Recognition
- Frequency vs. In-Context Predictability
- Analytical Methods
- Empirical Findings
- Implications (for NLP and more broadly)
- Simplifying Assumptions
- Future Work

Outline

- Rational Model of Distinctiveness
- Relationship to Spoken Word Recognition
- Frequency vs. In-Context Predictability
- Analytical Methods
- Empirical Findings
- Implications (for NLP and more broadly)
- Simplifying Assumptions
- Future Work

Bayesian Speech Recognition

- Upon hearing a string of sounds *s* a listener has to infer what word *w* was intended by the speaker. Should use Bayesian inference!

$$P(w|s) = \frac{P(s|w)P(w)}{P(s)}$$

- Assuming sounds are produced faithfully, then $P(s_w|w) = 1$ (and $P(s_w|w')$ decreases as w' is less similar to w)

$$P(w|s_w) \approx \frac{P(w)}{P(s_w)}$$

Prior Probability
(Frequency or Predictability)

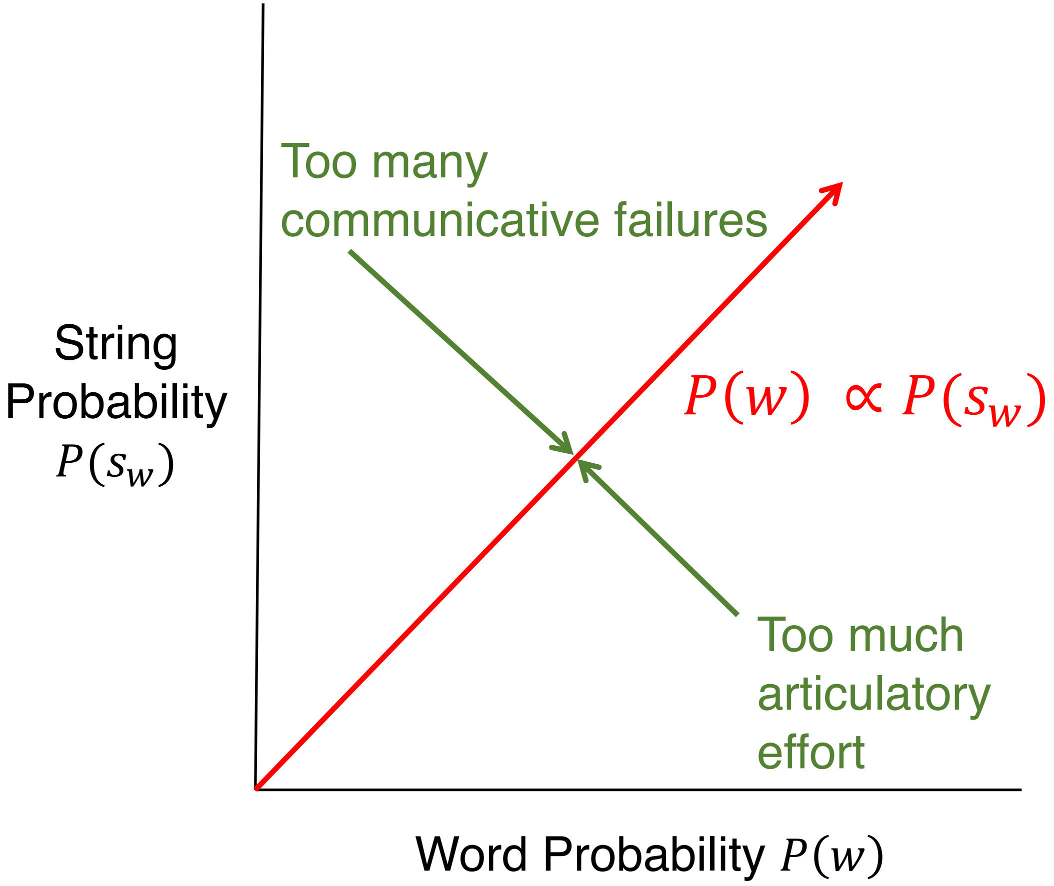
Word Form Probability
(distinctiveness = - log $P(s_w)$)

Distinctiveness and Frequency

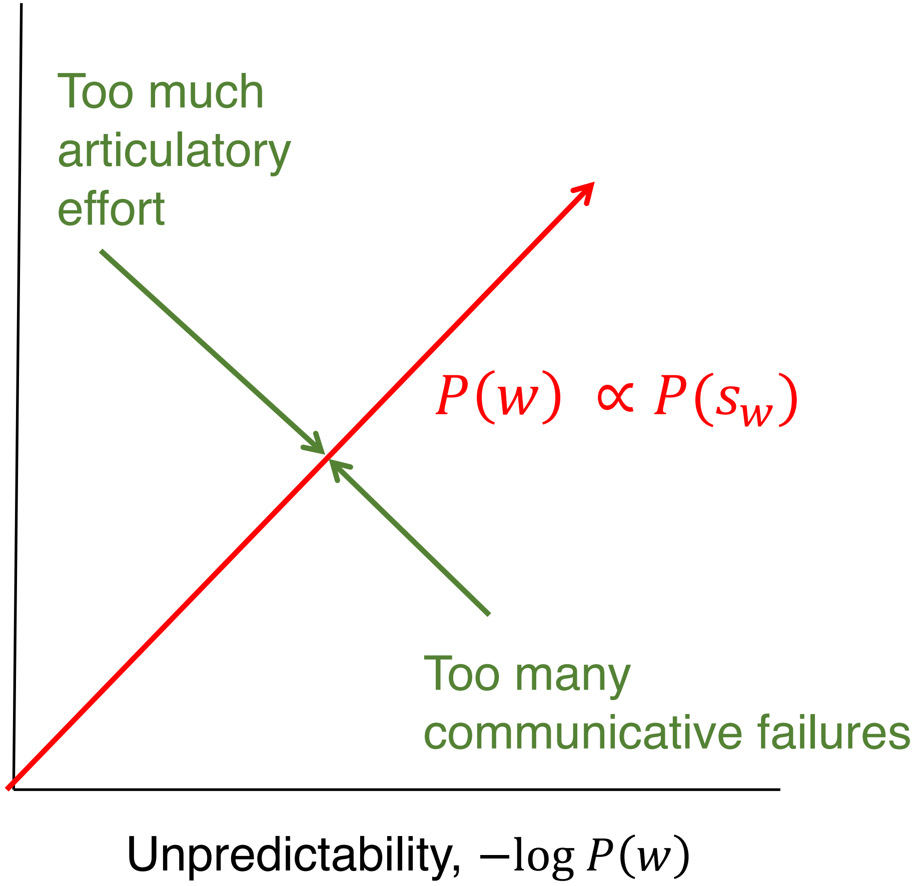
- Two reasons to think that $P(w) \propto P(s_w)$ (or $P(w)$ is inversely related to distinctiveness, $-\log P(s_w)$)

- 1) *Measure of listener effort*: Total effort is minimized when the most frequent words are the least distinctive
- 2) *Uniform Recognizability*: probability that any word is successfully recognized is approximately the same.

Communicative Pressures



Distinctiveness
Ease of distinguishing from other words,
$-\log P(s_w)$

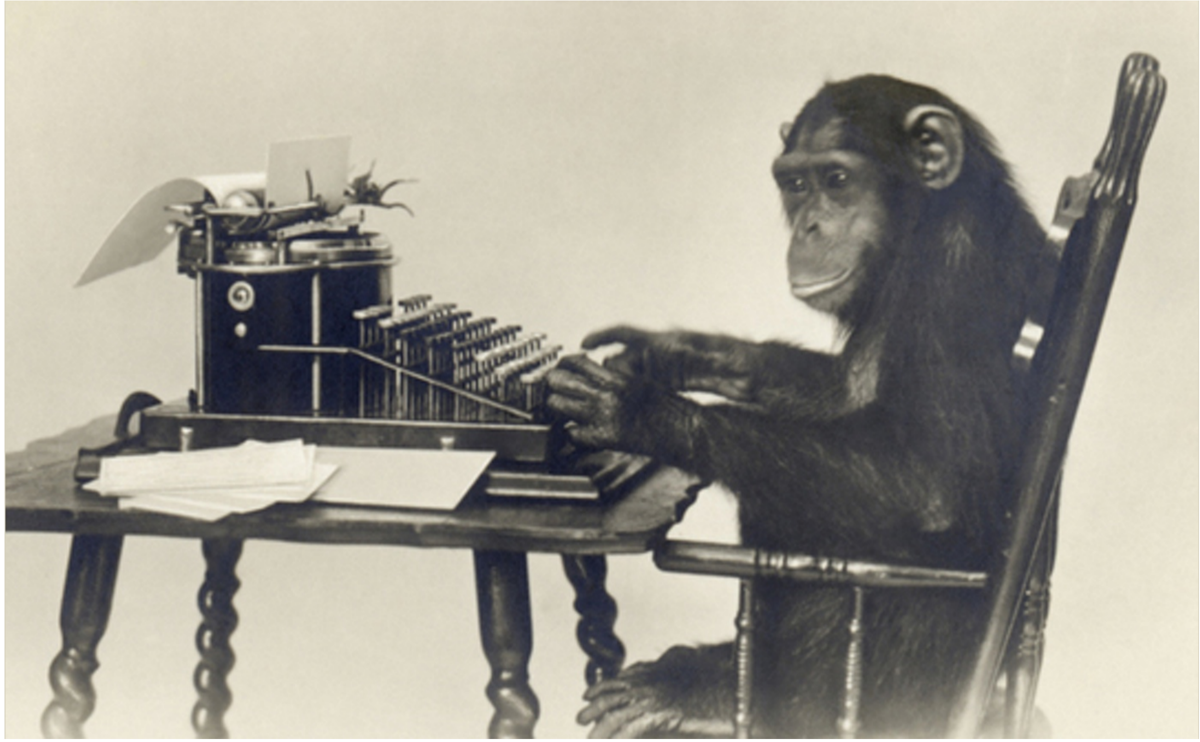


Marginal String Probability

- Computing $p(s_w)$ is hard. Equivalent to asking “how likely will any word be confused for this word”
- $p(s_w) = \sum_{w' \in V} p(s_w|w')p(w')$
- Trivial observation: high frequency words contain more common sequences
- Stronger test: compute distinctiveness under a type-weighted model $p(s_w) = \sum_{w' \in V} p(s_w|w')$ x 1
- Still need a way to compute $p(s_w|w')$
- Treat s_w as a sequence of phones

Naïve Model of Distinctiveness

- Naïve model: equally probable, independent phones or characters (Mandelbrot 1954; Miller, 1957)
- Longer strings are less probable / more informative
- $p(l_i) = \frac{1}{v}$, where v is the size of the phone inventory
- $D(s_w) = -\sum_i \log p(l_i)$
- $D(s_w) = |s_w| \log v$
- $D(s_w) \propto |s_w|$



People Are Many Things, But Not Naïve…

- People have rich knowledge of language structure. *t* is usually followed by *i* or *e*, but much less commonly by *b* or *g*
- “Sequential hangman” game from Shannon ‘51: guess each consecutive letter. Receive confirmation when right, and move to the next letter.

```
(1) T H E R E   I S   N O   R E V E R S E   O N   A   M O T O R C Y C L E   A
(2) 1 1 1 5 1 1 2 1 1 2 1 1 15 1 17 1 1 1 2 1 3 2 1 2 2 7 1 1 1 1 4 1 1 1 1 1 3 1
(1) F R I E N D   O F   M I N E   F O U N D   T H I S   O U T
(2) 8 6 1 3 1 1 1 1 1 1 1 1 1 1 1 6 2 1 1 1 1 1 1 2 1 1 1 1 1 1
(1) R A T H E R   D R A M A T I C A L L Y   T H E   O T H E R   D A Y
(2) 4 1 1 1 1 1 1 1 1 5 1 1 1 1 1 1 1 1 1 1 1 1 6 1 1 1 1 1 1 1 1 1 1 1 1 1    (9)
```

- Baseline of uniform character probabilities: 13*n* guesses (excluding spaces)
- Expected for first phrase: 29x13= 377 guesses
- Human performance: 80 guesses
- People use this knowledge in spoken word recognition (Vitevitch & Luce, 1998; Tanenhaus, Marslen-Wilson & Welsh, 1978, *inter alia*)

Improving Distinctiveness Estimate

- We can improve distinctiveness with a probabilistic model of phone-to-phone or character-to-character transitions from the language
- Approximate probabilistic phonotactic knowledge with an *n*-gram model built over phones or characters (~phones; high correlation)
- Type-weighted to avoid circularities
- Order = 5, modified Kneser-Ney smoothing with interpolation on the higher orders
- Distinctive = many unlikely transitions when inferring the identity of each sequential phone

Phonological Information Content

- Once we have a generative model for strings, we can then get some measure of how predictable each word form is in a given language: phonological information content
- PIC = Phonological Surprisal = relative entropy of the observed distribution with respect to the expected distribution

$$
\begin{aligned}
PIC(w) &= -\log P(s_w) \\
&= -\log P(l_1, \ldots, l_{|s_w|}) \text{ for } l \in s_w \\
&= -\sum_{i=1}^{|s_w|} \log P(l_i | l_{i-(n-1)}, \ldots, l_{i-1}).
\end{aligned}
$$

PIC Estimate vs. Length

"M"	"O"	"T"	"O"	…	Σ Bits
Uniform Character Probabilities					
$P(M)$	$P(O)$	$P(T)$	$P(O)$		
4.700	4.700	4.700	4.700	…	47.004
5-Character Phonological Information Content Model					
$P(M)$	$P(O\|M)$	$P(T\|MO)$	$P(O\|MOT)$		
5.358	3.223	3.451	6.110	…	26.327

"There is no reverse on a motorcycle" - 115 guesses

Draws from the English Model

breasonry blard remeditormer zengemsis rists craver bonderely mitting embard instaming triminations sabotatic patrian evictors clouddle funge drinths campful troke exprecapting aposed steride condits aremisting commony acture tallfidity renades idiousness disson furtly thuming twelving thema unmatizing hions trously pillates subcialists hoarsmic psychotter eths revollectors shoveliner scamen bloker contrained warblement clumsines telegating mumb clefolions admoning

Outline

- Rational Model of Distinctiveness
- Relationship to Spoken Word Recognition
- Frequency vs. In-Context Predictability
- Analytical Methods
- Empirical Findings
- Implications (for NLP and more broadly)
- Simplifying Assumptions
- Future Work

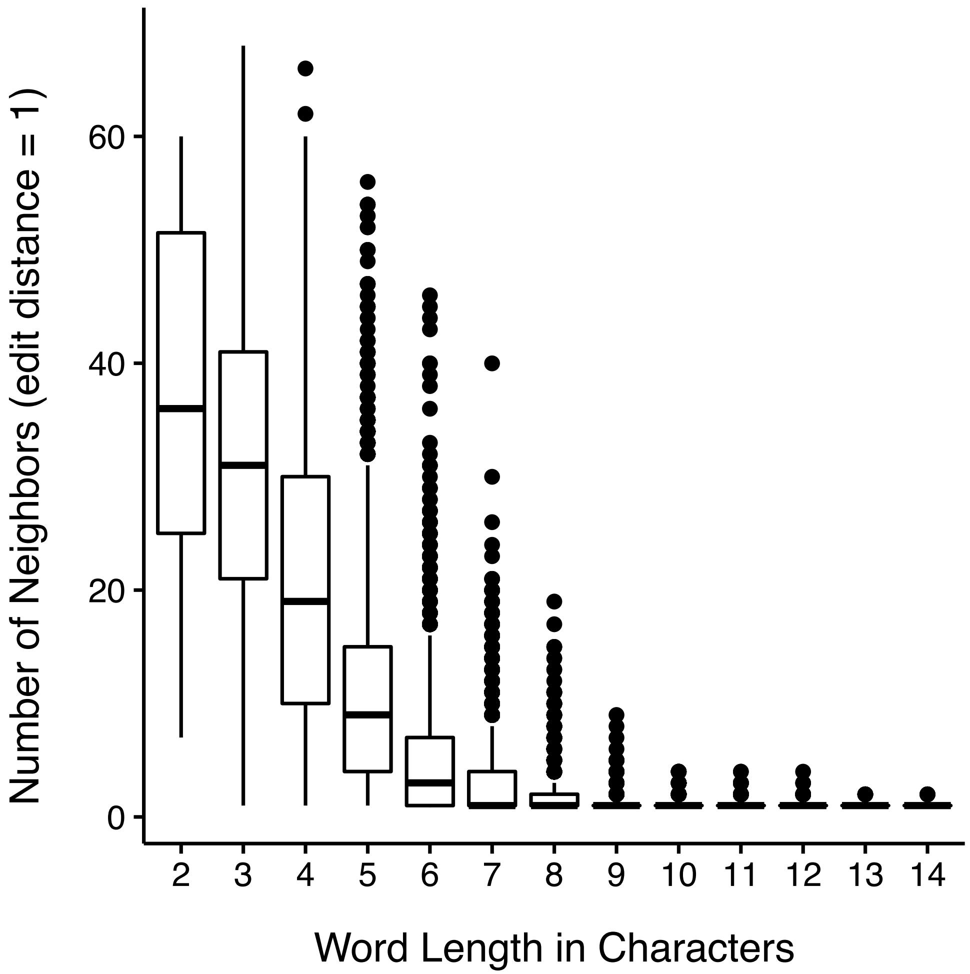
Prior Probability

$$P(w|s_w) \approx \frac{P(w)}{P(s_w)}$$

Word Form Probability
(distinctiveness = - log $P(s_w)$)

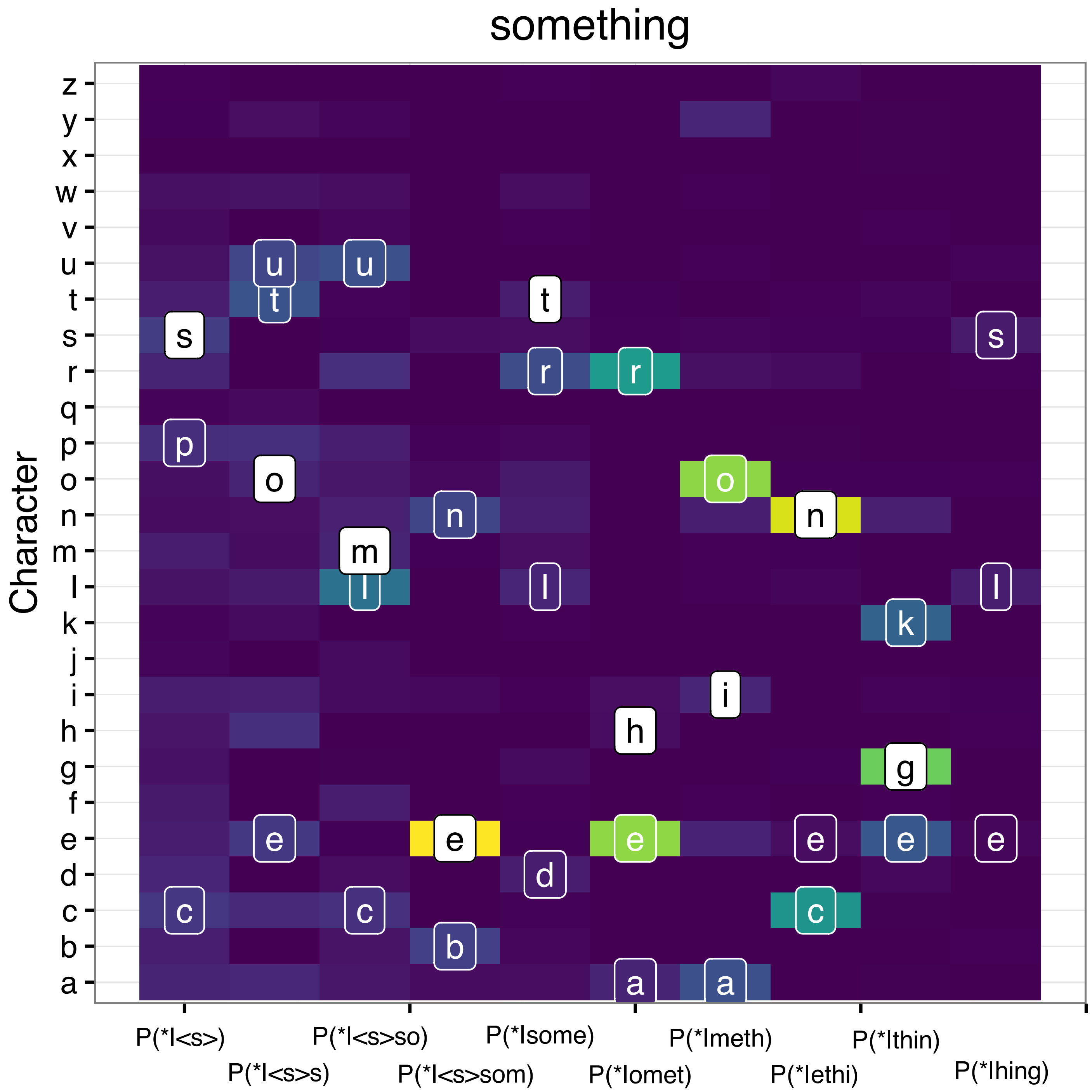
Length and Neighborhood Density

- Shorter words have more similar “neighbors” (competitors); long words have more sparse neighborhoods
- Neighbor: approximated by edit distance of one phone or character (Coltheart’s *N*)
- Complex effects of neighborhood density

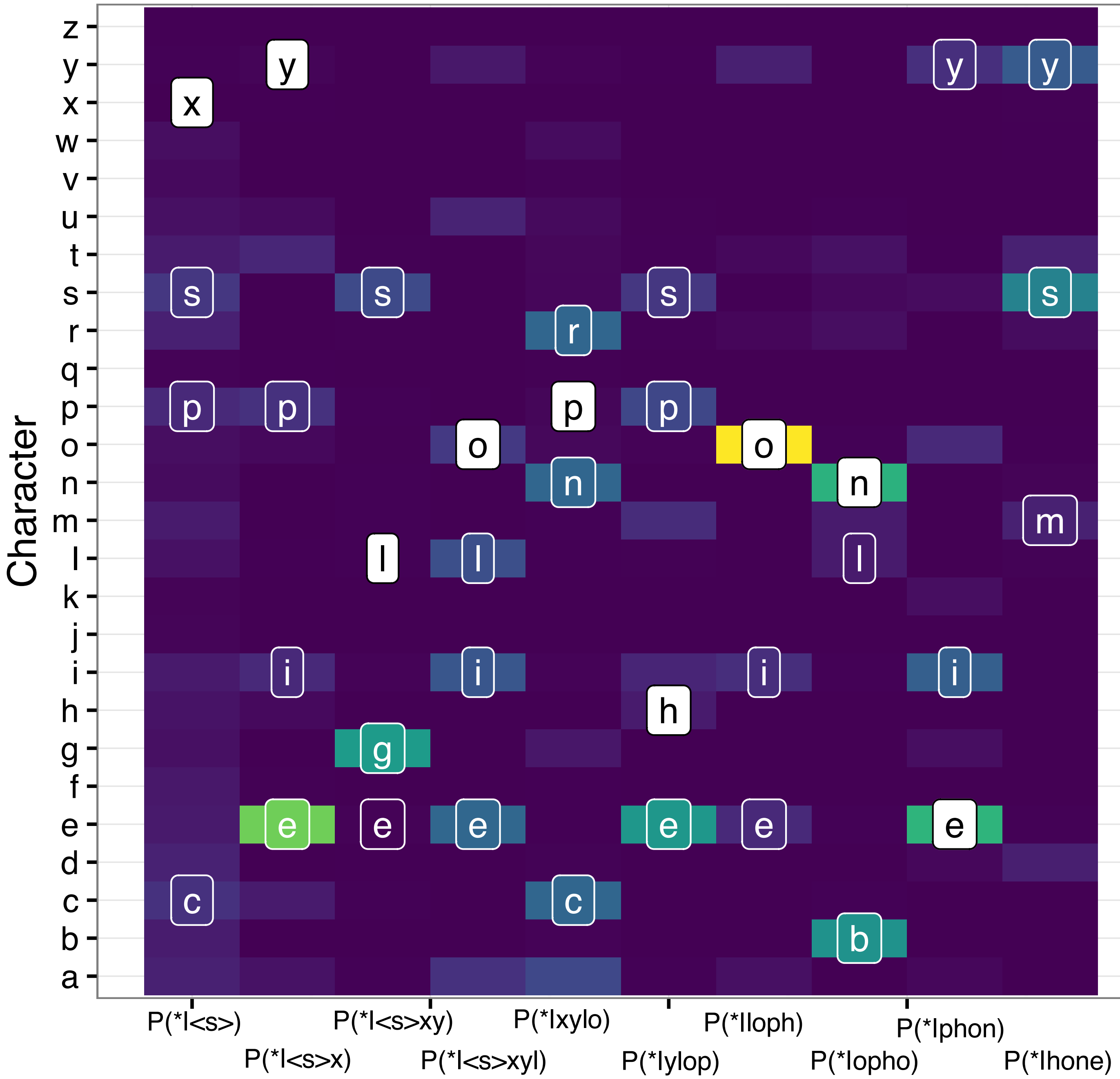


PIC and Neighborhood Density

- PIC: surprisal from incremental phone recognition
  - consistent with the cohort model of Marslen Wilson, 1978
- PIC captures competition effects from early in the word for longer words
  - 40 possible candidates for *thesis* after *the*; compared with only 2 words within Levenshtein distance of 1: *theses* and *Theseus*
- Significantly better than length in predicting lexical decision times for nonwords
- Estimate of processing difficulty, by analogy with lexical surprisal (Levy, 2008)


something
Character
z
y
x
w
v
u
t
s
r
q
p
o
n
m
l
k
j
i
h
g
f
e
d
c
b
a
P(*|<s>)
P(*|<s>s)
P(*|<s>so)
P(*|<s>som)
P(*|some)
P(*|omet)
P(*|meth)
P(*|ethi)
P(*|thin)
P(*|hing)

xylophone


Character
z
y
x
w
v
u
t
s
r
q
p
o
n
m
l
k
j
i
h
g
f
e
d
c
b
a
P(*l<s>)
P(*l<s>x)
P(*l<s>xy)
P(*l<s>xyl)
P(*lxylo)
P(*lylop)
P(*lloph)
P(*lopho)
P(*lphon)
P(*lhone)

Outline

- Rational Model of Distinctiveness
- Relationship to Spoken Word Recognition
- Frequency vs. In-Context Predictability
- Analytical Methods
- Empirical Findings
- Implications (for NLP and more broadly)
- Simplifying Assumptions
- Future Work

Prior Probability

$$P(w|s_w) \approx \frac{P(w)}{P(s_w)}$$

Word Form Probability
(distinctiveness = - log $P(s_w)$)

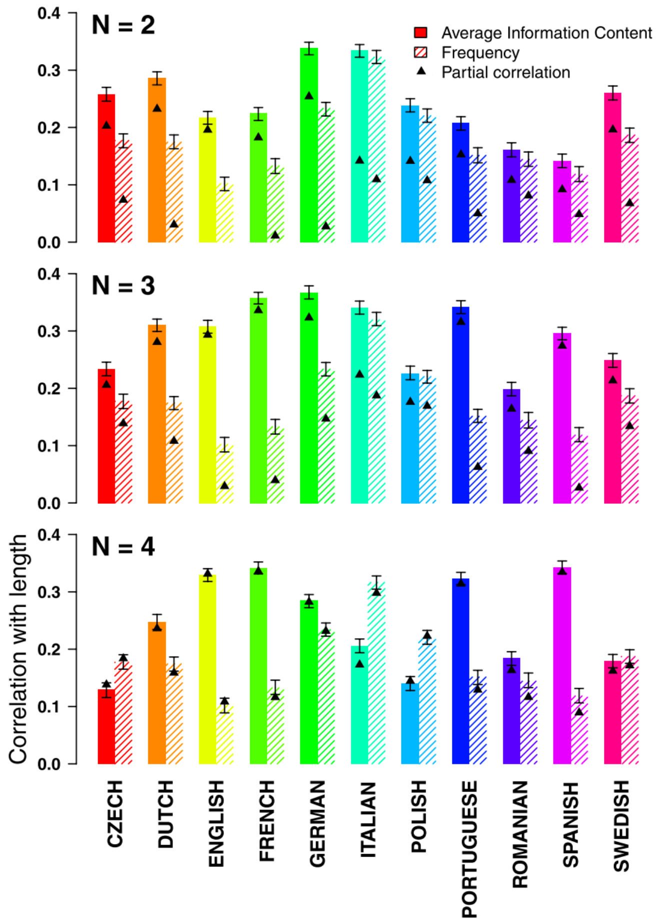
In-Context Lexical Surprisal

- Piantadosi et al. 2011: What if length is driven by *mean information content across contexts* for a word rather than frequency
- Use lexical surprisal in place of frequency

$$-\frac{1}{N}\sum_{i=1}^{N}\log P(W = w|C = c_i).$$

- Where *N* is the number of tokens, *w* is the word, c_i is the context
- The frequency and lexical surpisal are highly correlated, so they partial out the former
- Test with *n*-gram models on Google 1T corpora (100B – 1T words, nominally)

In-Context Lexical Surprisal, Continued



- Significantly stronger correlation
- Elegant connection to advances in sentence processing (Levy, 2008); relates word length primarily to processing pressures
- Here: also look if in-context mean predictability correlates with PIC

Outline

- Rational Model of Distinctiveness
- Relationship to Word Recognition
- Frequency vs. In-Context Predictability
- Analytical Methods
- Empirical Findings
- Implications (for NLP and more broadly)
- Simplifying Assumptions
- Future Work

Corpora and Preprocessing

- Google 1T (web), Google Books 2012 (books, limited to after 1800), OPUS 2013 (movie subtitles)
- 13 languages in total (11, 7, 13 respectively), 43M to 266B words
- UTF-8, lowercase (specific to locale), punctuation removed, POS merged, stopwords retained

- For each (language * corpus), compute frequency, bigram and trigram surprisal; type inventory for phone and character models
- Type inventory for analysis: restrict to 25k highest frequency words in OPUS ∩ Aspell
- ZS: parallel decompression of *n*-gram tries (English trigrams 20TB uncompressed text)

Generating Phonemic Transcriptions

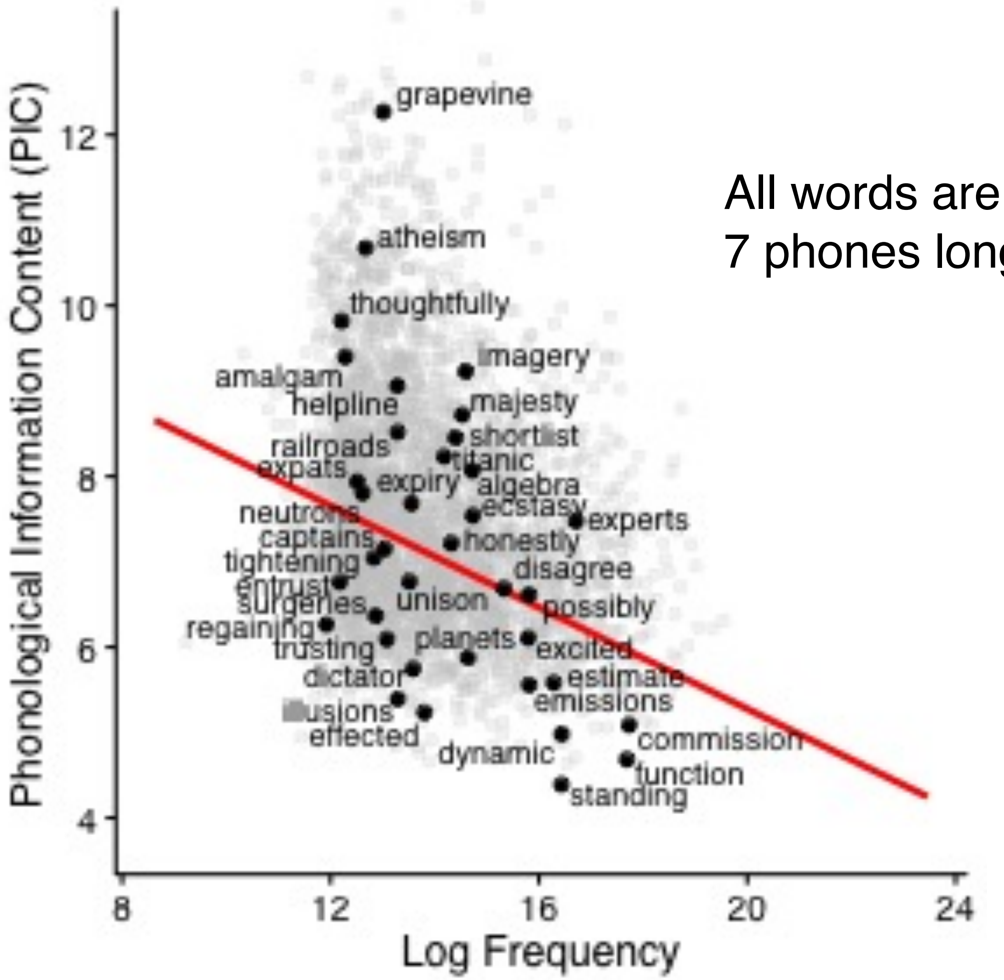
- Orthography itself is influenced by efficient communication. “Don’t signal what can be reasonably assumed”
- Example: Spanish accents included only when they violate the usual pattern (end in consonant other than *n* or *s*: penultimate syllable)
- Phone transcriptions of the type list with a cross-linguistic speech synthesizer (espeak) that can output IPA
- Variable quality. Confirmed with L1 or proficient L2 speakers a sample from English, German, Spanish, Hebrew, and French

Outline

- Rational Model of Distinctiveness
- Relationship to Word Recognition
- Frequency vs. In-Context Predictability
- Analytical Methods
- Empirical Findings
- Implications (for NLP and more broadly)
- Simplifying Assumptions
- Future Work

Empirical Findings

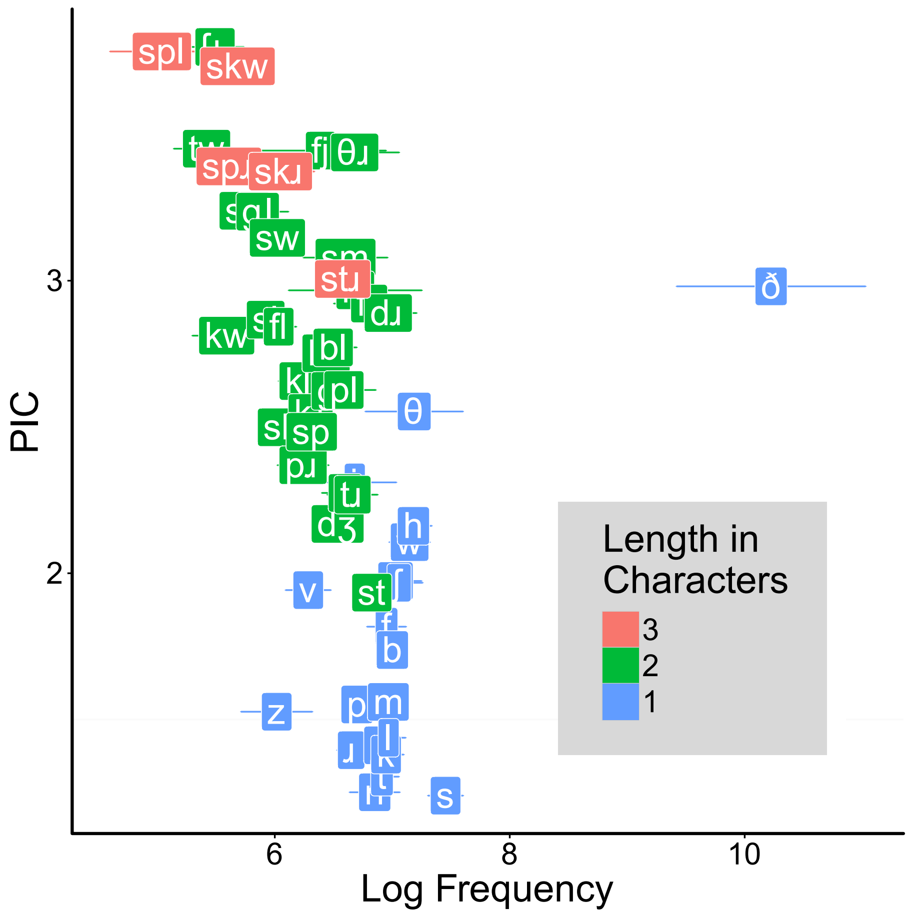
- PIC accounts for significantly more variance in frequency than does length
- Example: strong relationship between PIC and frequency among words of the same length



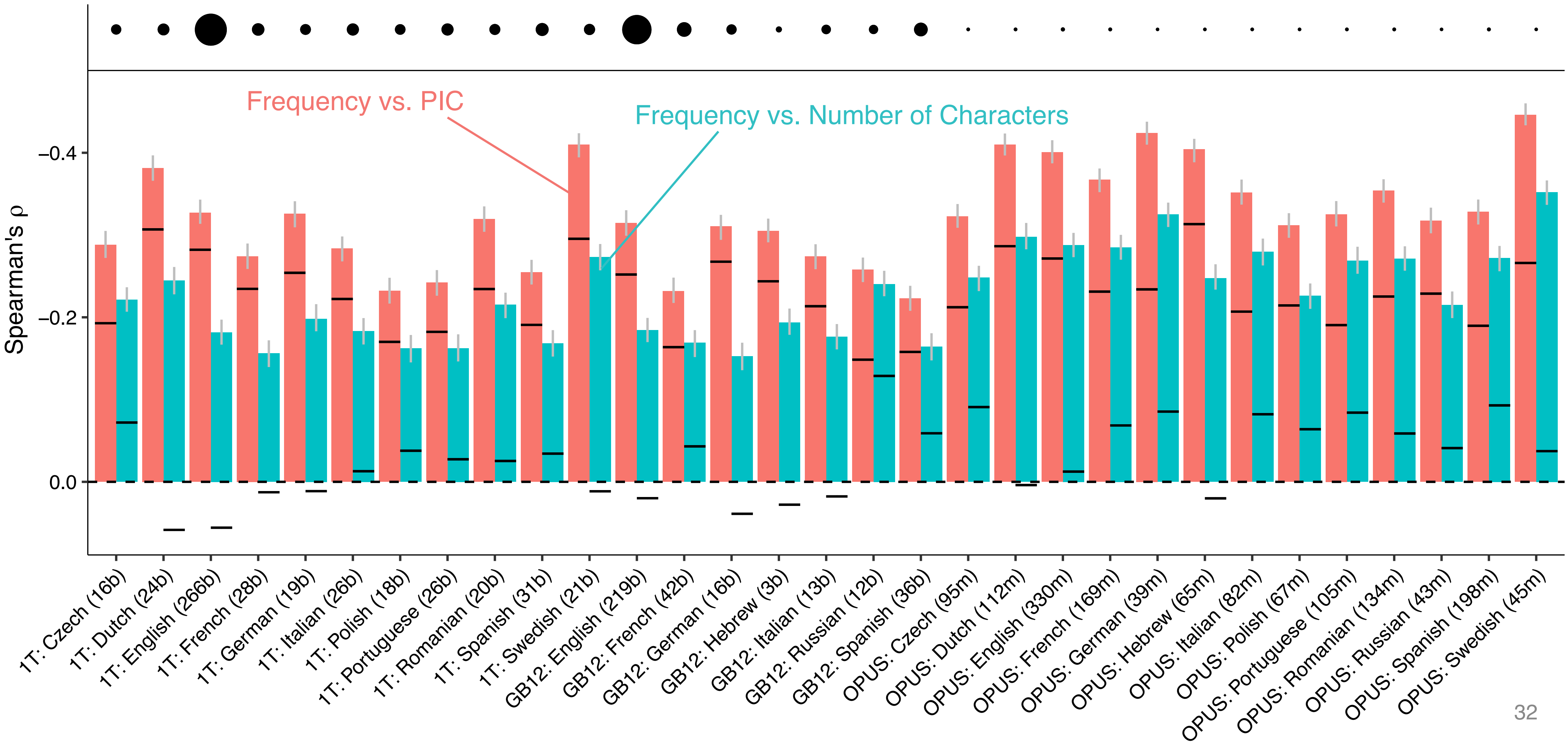
All words are
7 phones long

Improvements In Fit From PIC

- Differentiation
  - *something* (27.13 bits) vs. *xylophone* (34.48 bits)
- Inversion
  - *depth* (5 letters, 17.86 bits) vs. *ground* (6 letters, 12.85 bits)
- PIC of consonant clusters generally corresponds with length in characters

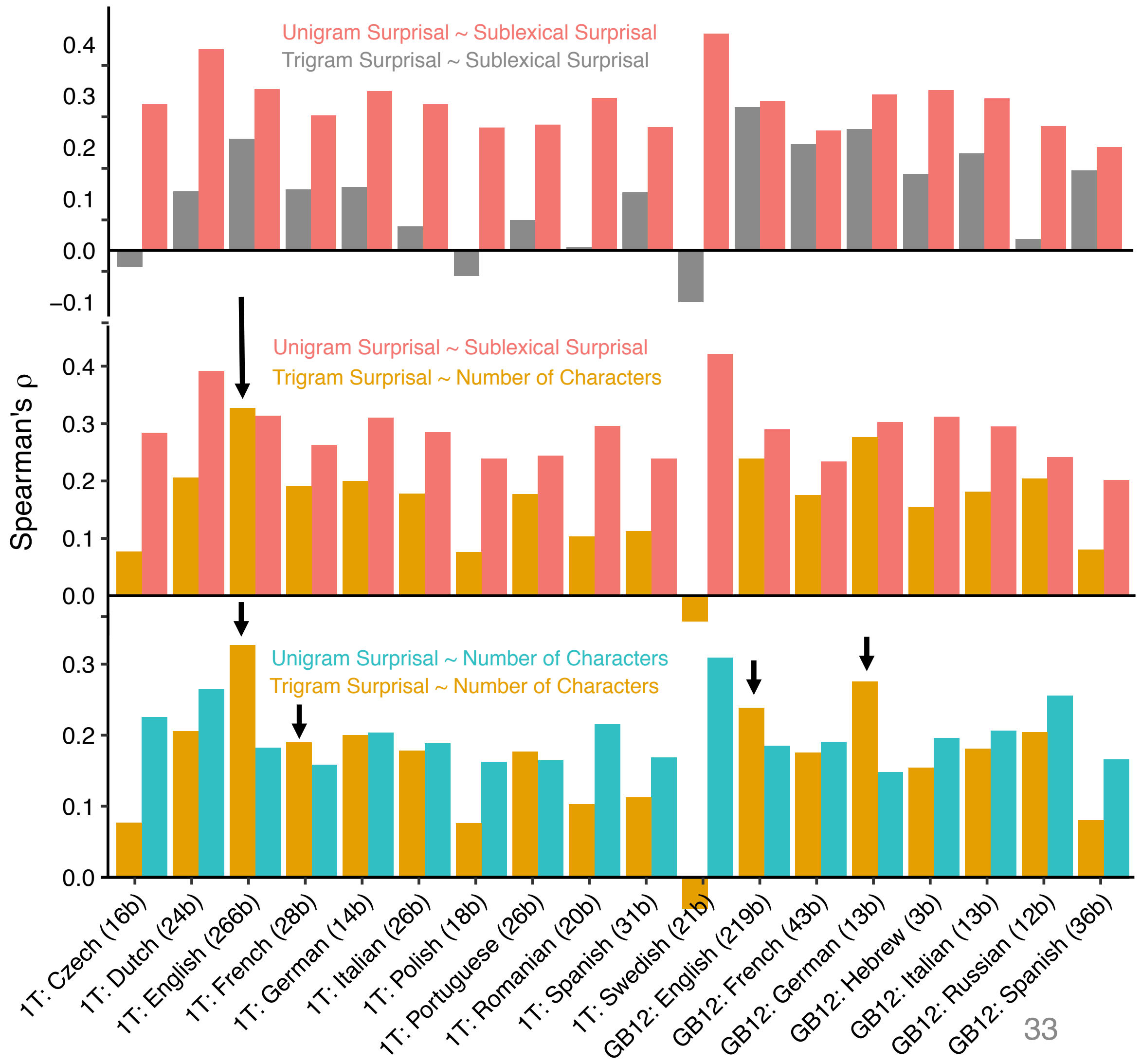


Cross-Linguistic Correlation



Vs. Lexical Surprisal?

- Trigram surprisal not a good predictor of PIC (top)
- Trigram surprisal not good at predicting word length (middle)
- Advantage of trigram surprisal w.r.t. frequency from Piantadosi et al. (2011) largely disappears when we use UTF-8 and limit words in the analysis to those in the dictionary (aspell)

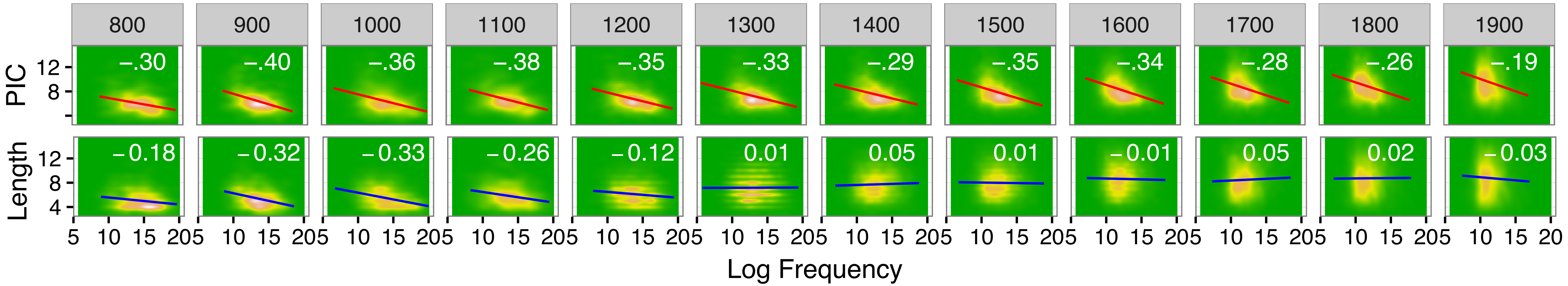


Estimating mean trigram surprisal in morphologically complex languages

- Sparsity problems for when we restrict to in-dictionary words?
- Example with vender:
  - Handful of forms in English: sell, sells, sold, selling.
  - Spanish: 140 forms in the top 50k (verb conjugation + clitic object markers)
- Zipfian distribution: higher frequency lemma can have many lemmas in the analysis before one form of a lower frequency lemma

Historical Fit

- Changes in length captures deletion and epenthesis (e.g., *luncheon* to *lunch*)
- Change in length is the “nuclear option” in language change
- PIC captures other forms of weathering: assimilation, dissimilation, metathesis. (*aks* -> *ask*). How to test?



- Some words that are relatively uncommon now still have very high string probability (ewe)
- Is there such a thing as reverse weathering?

Outline

- Rational Model of Distinctiveness
- Relationship to Spoken Word Recognition
- Frequency vs. In-Context Predictability
- Analytical Methods
- Empirical Findings
- Implications (for NLP and more broadly)
- Simplifying Assumptions
- Future Work

Back to cross-linguistic regularities

- Language structure shaped to support efficient inference
- Recipe for language:
    - general learning mechanisms: sequence learning
    - rich, learned representations: n-gram model
    - *highly evolved substrate:* $P(w) \propto P(s_w)$
- Compare with computer vision: natural scene statistics are *not* subject to pressures from inference
- Languages are themselves part of the solution—it's not just the processing architecture which is special!

Implications for NLP / Machine Learning

- Primarily, this work adapts tools from NLP to answer questions about regularities in natural language
- “Don’t need to know how feathers work to build an airplane”
  - This is a trade-off at the level of aerodynamics…
- Emphasizes the prevalence of reduction: strong listener expectations license reduction / deviation from the target form
  - (probably -> prolly -> pry)
- Stresses the importance of priors: world-knowledge, discourse information, syntactic information
- Not noise!

Simplifying Assumptions

- Treat words as discrete symbols
  - Gradient effects in English (farmer’s market; black bird)
  - Words in agglutinative languages (e.g., Turkish)
- Estimate distinctiveness with phones, not sub-phone features (see Futrell et al. in press)
  - Word = an ordered collection of phone symbols
  - No long-distance dependencies
- Compute probabilities while observing word boundaries

Simplifying Assumptions, Continued

- Markov assumption… not marginalizing over possible preceding phones in the prediction
- More ambiguity in short words?

Caveat Corpora

- Language use approximated by web pages, books, or subtitles
- Most common trigram ending in ‘Romanian’ from Google 1T is “Polish Portuguese Romanian” (~22% trigrams ending in Romanian)
- English 1T is kill-it-with-fire / remove from LDC bad
- ~20% of the unigram probability mass is taken by symbols corresponding to bound morphemes in the Hebrew Google Books dataset
- Looking for other (cross-linguistic) corpora (new work with OPUS 2016)

Google 1T English Common Trigrams

Trigram	Count
© Yahoo !	110,465,581
\ \ \	109,239,351
[date]	22,792,481
[thread]	22,743,710
Your Account I	19,472,801
[PubMed -	17,073,885
Write a review	13,288,438
Your Buddy List	12,195,201

Future Directions

- Extend to a larger cross-linguistic sample
- Explore token-weighted phonological model
  - token-weighted model displays human performance on the Shannon Game
- Investigate interaction with speech rate
- Look at historical changes in wordforms
  - Change in predictability precedes a change in the wordform?
- Develop better information content estimates (e.g. LSTMs)
  - How do PCFGs / n-gram models / LSTMs perform across languages?

Conclusion

- The relationship between word length and frequency observed across natural languages supports efficient inference
- A relatively simple probabilistic language model (from NLP) gives us a significantly improved estimate of string distinctiveness
- The mapping between word form and use is not arbitrary: cognitive pressures shape word forms

Thanks!

- Tom Griffiths, Roger Levy, Terry Regier, Steve Piantadosi
- Nathaniel Smith: pySRILM and ZS
- Teeranan Pokaprakarn: Ngrawk
- NSF Big Data Grant SMA-1228541
- NSF Graduate Research Fellowship DGE-1106400

Questions?