An Informative Exploration of the Lexicon

Tiago Pimentel
Summary

Four parts:

- Phonotactic Complexity and Its Trade-offs
- Disambiguatory Signals are Stronger in Word-initial Positions
- (Non-)Arbitrariness of the Sign
  - Meaning to Form: Measuring Systematicity as Information
  - Finding Concept-specific Biases in Form–Meaning Associations
- Speakers Fill Lexical Semantic Gaps with Context
1. We use **bits per phoneme** as a measure of phonotactic complexity, and find it has **robust trade-offs** with **word length**.
2. We analyse **bits per phoneme** in first and second halves of words, and find a cross-linguistic **tendency** to **frontload information**.
3. We operationalise **systematicity of the sign** as a **mutual information between word forms and meanings**, and find small (but significant) values both **within and across languages**.

| Language | $H(W)$ | $MI(W;V)$ | $U(W|V)$ |
|----------|--------|------------|----------|
| English  | 3.401  | 0.11       | 3.24%    |
| German   | 3.195  | 0.168      | 5.26%    |
| Dutch    | 3.245  | 0.156      | 4.82%    |

| Test     | $H(W)$ | $MI(W;V)$ | $U(W|V)$ |
|----------|--------|------------|----------|
| Africa   | 3.773  | 0.011      | 0.28%    |
| Americas | 3.901  | 0.007      | 0.17%    |
| Eurasia  | 3.999  | 0.015‡     | 0.38%    |
| Pacific  | 3.755  | 0.016‡     | 0.42%    |
| Average  | 3.857  | 0.012‡     | 0.31%    |
4. We operationalise **lexical ambiguity** as the **conditional entropy of a meaning given a word**, and find **consistent trade-offs** with the word's contextual entropy.
Summary

Four parts:

- **Phonotactic Complexity and Its Trade-offs**
- Disambiguatory Signals are Stronger in Word-initial Positions
- (Non-)Arbitrariness of the Sign
  - Meaning to Form: Measuring Systematicity as Information
  - Finding Concept-specific Biases in Form–Meaning Associations
- Speakers Fill Lexical Semantic Gaps with Context
Phonotactic Complexity and Its Trade-offs

Tiago Pimentel, Brian Roark, and Ryan Cotterell
Phonotactics

Sequences of speech sounds allowed in a language.

For instance, in English:

- brick
- blick
- bnick
Compensation Hypothesis

Several linguists believe all languages are equally complex.

Consequence: compensatory relationships between complexity measures should exist.

- e.g. vowel inventory size in a language correlates to a language’s average word length.
Several measures of a language's phonological complexity exist:

- **Size of Phoneme Inventory:**
  - the number of vowel categories in a language;

- **Markedness in Phoneme Inventory:**
  - marked phonemes, such as clicks, make a language more complex;

- **Number of Licit Syllables:**
  - phonological constraints extend beyond individual units, so counting syllables seems like a logical next step in measuring its complexity;

- **Word Length:**
  - implicitly taken as a complexity measure when researchers examine its correlation with e.g. inventory size.
Phonotactic complexity as *bits per phoneme*
Phonotactic Complexity - Characteristics

Bits per phoneme is the **entropy** of a language’s word types.

We estimate it using a character-level language model’s **cross-entropy**:

\[
H(p_{\text{lex}}) \leq H(p_{\text{lex}}, q_{\text{lex}}) \approx -\frac{1}{N} \sum_{i=1}^{N} \log q_{\text{lex}}(\hat{x}^{(i)})
\]
Bits per phoneme is the entropy of a language’s word types.

We estimate it using a character-level language model’s cross-entropy:

$$H(p_{lex}) \leq H(p_{lex}, q_{lex}) \approx -\frac{1}{N} \sum_{i=1}^{N} \log q_{lex}(\hat{x}^{(i)})$$

Tighter, the better the language model.

- Trigram language model;
- Character-level LSTM language model.
Phonotactic Complexity - Characteristics

Linguistic rationale for bits-per-phoneme:

- Modest linguistic annotations
- Incorporates frequency of phenomena
- Captures interaction between phonemes
- Long-distance dependencies
Data: NorthEuraLex (Dellert and Jäger, 2017)

1016 concepts in 107 languages in IPA

**Concept-aligned** lexicons

Composed of “**basic**” concepts

Languages from 21 families in Europe/Asia

We omit Mandarin (no tone annotation)
Trade-Off

<table>
<thead>
<tr>
<th>Measure</th>
<th>Pearson $r$</th>
<th>Spearman $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>phonemes</td>
<td>-0.047</td>
<td>-0.054</td>
</tr>
<tr>
<td>vowels</td>
<td>-0.164</td>
<td>-0.162</td>
</tr>
<tr>
<td>consonants</td>
<td>0.030</td>
<td>0.045</td>
</tr>
<tr>
<td>Bits/phoneme:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unigram</td>
<td>-0.217</td>
<td>-0.222</td>
</tr>
<tr>
<td>trigram</td>
<td>-0.682</td>
<td>-0.672</td>
</tr>
<tr>
<td>LSTM</td>
<td>-0.762</td>
<td>-0.744</td>
</tr>
</tbody>
</table>
Trade-Off

Bits-per-phoneme versus average word length in IPA.
Control studies

Possible confound: positional effects.

- Phonemes later in a word in general have higher probability given the previous phonemes than those earlier in the string (van Son and Pols, 2003).

Truncated Words: Only consider first 3 characters in wordforms.

- Original (Full words): $\rho = -0.744$;
- Control (Truncated): $\rho = -0.469$
Inter- and Intra- Family Trade-Offs

Classic measures of phonological complexity:

- correlate with word length across a varied set of languages,
- but do not within language families.

Bits per phoneme correlates in both cases.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Pearson $r$</th>
<th>Spearman $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>phonemes</td>
<td>-0.214</td>
<td>-0.095</td>
</tr>
<tr>
<td>vowels</td>
<td>-0.383</td>
<td>-0.367</td>
</tr>
<tr>
<td>consonants</td>
<td>-0.147</td>
<td>-0.092</td>
</tr>
</tbody>
</table>

| Bits/phoneme: |             |

<table>
<thead>
<tr>
<th>Family</th>
<th>Spearman $\rho$</th>
<th>LSTM</th>
<th>Vowels</th>
<th># Langs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dravidian</td>
<td>-0.894</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Indo-European</td>
<td>-0.218</td>
<td>37</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>Nakh-Daghestanian</td>
<td>-0.530</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Turkic</td>
<td>-0.773$^1$</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Uralic</td>
<td>0.363$^1$</td>
<td>26</td>
<td>26</td>
<td></td>
</tr>
</tbody>
</table>

* Statistically significant with $p < 0.01$

$^1$ Statistically significant with $p < 0.1$
An Informative Exploration of the Lexicon

Tiago Pimentel

Part 2
Summary

Four parts:

- Phonotactic Complexity and Its Trade-offs
- **Disambiguatory Signals are Stronger in Word-initial Positions**
- (Non-)Arbitrariness of the Sign
  - Meaning to Form: Measuring Systematicity as Information
  - Finding Concept-specific Biases in Form–Meaning Associations
- Speakers Fill Lexical Semantic Gaps with Context
Disambiguatory Signals are Stronger in Word-initial Positions

Tiago Pimentel, Ryan Cotterell, Brian Roark
Research Question

Are word-initial segments more informative for disambiguation than word-final ones?
Introduction

Is it easier to guess the ending of "dino****"?
Introduction

Is it easier to guess the ending of "dino****"?
**Introduction**

Is it easier to guess the ending of "dino****"?

Or the prefix of "****saur"?

****saur!

Did he just say dinosaur or venusaur?
How can we measure this?

Psycholinguistic experiments with human subjects:
How can we measure this?

Psycholinguistic experiments with human subjects:

- Listeners find word-initial consonant deletions more disruptive than word-final (Bagley, 1900).
- Mispronunciations are more likely in word endings (Fay and Cutler, 1977).
- Recognizing written words with flipped initial characters is harder than with final ones (Bruner and O’Dowd, 1958).
How can we measure this?

Psycholinguistic experiments with human subjects:

- listeners find word-initial consonant deletions more disruptive than word-final (Bagley, 1900)
- mispronunciations are more likely in word endings (Fay and Cutler, 1977)
- recognizing written words with flipped initial characters is harder than with final ones (Bruner and O'Dowd, 1958)

This does not measure how informative segments are. Only how useful they are for humans.
How can we measure this?

Information-theoretic measurements on natural corpora:

- Estimate a segment’s information as its contextual entropy:

\[
H(W_i | W_{<i}) = - \sum_{w \in \Sigma^*} p(w_i | w_{<i}) \log p(w_i | w_{<i})
\]
How can we measure this?

Information-theoretic measurements on natural corpora:

- Estimate a segment's information as its contextual entropy:

\[
H(W_i \mid W_{<i}) = - \sum_{w \in \Sigma^*} p(w_i \mid w_{<i}) \log p(w_i \mid w_{<i})
\]
How can we measure this?

Information-theoretic measurements on natural corpora:

- Estimate a segment's information as its contextual entropy:

\[
H(W_i \mid W_{<i}) = - \sum_{w \in \Sigma^*} p(w_i \mid w_{<i}) \log p(w_i \mid w_{<i})
\]
How can we measure this?

Information-theoretic measurements on natural corpora:

- Estimate a segment’s information as its contextual entropy:

\[
H(W_i \mid W_{<i}) = - \sum_{w \in \Sigma^*} p(w_i \mid w_{<i}) \log p(w_i \mid w_{<i})
\]
How can we measure this?

Information-theoretic measurements on natural corpora:

- Estimate a segment’s information as its contextual entropy:

$$H(W_i | W_{<i}) = - \sum_{w \in \Sigma^*} p(w_i | w_{<i}) \log p(w_i | w_{<i})$$

This inherently confounds context size and word position.
How can we measure this?

Information-theoretic measurements on natural corpora:

- Estimate a segment’s information as its contextual entropy:

\[ H(W_i \mid W_{<i}) = - \sum_{w \in \Sigma^*} p(w_i \mid w_{<i}) \log p(w_i \mid w_{<i}) \]

This inherently confounds context size and word position.
Left-to-right Conditional Entropy

Conditioning information can only reduce entropy!

\[ H(W_t | W_{<t}) \leq H(W_t | W_{t-1}) \leq H(W_t) \]
Left-to-right Conditional Entropy

Conditioning information can only reduce entropy!

$$H(W_t | W_{<t}) \leq H(W_t | W_{t-1}) \leq H(W_t)$$

Left-to-right conditional entropies, thus:

- confound the amount of conditional information with word position.

After hearing *dino-*", saur is not very informative...
Why not Left-to-right?

Consider an artificial language where every word contains a copy of its first half:

- e.g., foofoo, barbar, foobarfoobar, etc.
- initial and final halves have identical disambiguatory strength; they are the same!
- conditional surprisal would be nearly zero for final halves.
Does left-to-right conditional entropy measure a property of the lexicon or simply the fact that conditioning reduces entropy?
Results - Forward Surprisal

All but one language in the three analysed datasets had larger word-initial surprisal

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Languages</th>
<th>Forward Unigram</th>
<th>Forward Position-Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELEX</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>NorthEuraLex</td>
<td>107</td>
<td>106</td>
<td>0</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>41</td>
<td>41</td>
<td>0</td>
</tr>
</tbody>
</table>
Results - Forward Surprisal

All but one language in the three analysed datasets had larger word-initial surprisal

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Languages</th>
<th>Forward</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELEX</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>NorthEuraLex</td>
<td>107</td>
<td>106</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>41</td>
<td>41</td>
</tr>
</tbody>
</table>
Many languages have higher word-final surprisals.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Languages</th>
<th>Forward</th>
<th>Backward</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELEX</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>NorthEuraLex</td>
<td>107</td>
<td>106</td>
<td>11</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>41</td>
<td>41</td>
<td>0</td>
</tr>
</tbody>
</table>
Results

There seems to be both:

- a large effect of the amount of conditional information
- a lexical effect of front-loading disambiguatory signals
Controlling for Context Size

We propose the use of three measures to control for context size:
Controlling for Context Size

We propose the use of three measures to control for context size:

● Unigram surprisal: $\mathbb{H}(W_t)$
Controlling for Context Size

We propose the use of three measures to control for context size:

- Unigram surprisal: $H(W_t)$
Controlling for Context Size

We propose the use of three measures to control for context size:

- Unigram surprisal: $\mathcal{H}(W_t)$

Without knowing anything else, how relevant is $s$?
Controlling for Context Size

We propose the use of three measures to control for context size:

- Unigram surprisal: $\mathbb{H}(W_t)$
- Cloze Surprisal: $\mathbb{H}(W_t \mid W_{\neq t})$
Controlling for Context Size

We propose the use of three measures to control for context size:

- Unigram surprisal: $H(W_t)$
- Cloze Surprisal: $H(W_t \mid W \neq t)$

After hearing *dino*aur, how relevant is s?
Controlling for Context Size

We propose the use of three measures to control for context size:

- Unigram surprisal: \( u(W_t) \)
- Cloze Surprisal: \( H(W_t | W \neq t) \)

After hearing \textit{dino*aur}, how relevant is \textit{s}?
Controlling for Context Size

We propose the use of three measures to control for context size:

- Unigram surprisal: $H(W_t)$
- Cloze Surprisal: $H(W_t \mid W_{\neq t})$
- Position-specific Surprisal: $H(W_t \mid T = t, |W|)$
Controlling for Context Size

We propose the use of three measures to control for context size:

- Unigram surprisal: $H(W_t)$
- Cloze Surprisal: $H(W_t | W \neq t)$
- Position-specific Surprisal: $H(W_t | T = t, |W|)$

****s***!

Knowing the position, how relevant is $s$?
Controlling for Context Size

We propose the use of three measures to control for context size:

- **Unigram surprisal:** $H(W_t)$
- **Cloze Surprisal:** $H(W_t \mid W \neq t)$
- **Position-specific Surprisal:** $H(W_t \mid T = t, |W|)$
Controlling for Context Size

We propose the use of three measures to control for context size:

- Unigram surprisal: $H(W_t)$
- Cloze Surprisal:
- Position-specific Surprisal:

Knowing the position, how relevant is $s$?

Inspired by Nooteboom and van der Vlugt’s (1988) experiments.
Data

CELEX (Baayen et al., 2015):

NorthEuraLex (Dellert et al., 2019)

Wikipedia
Data

CELEX (Baayen et al., 2015):
● English, German and Dutch;

NorthEuraLex (Dellert et al., 2019)
● 107 languages from 21 language families;

Wikipedia
● 41 typologically diverse languages;
Data

CELEX (Baayen et al., 2015):
- English, German and Dutch;
- monomorphemic words.

NorthEuraLex (Dellert et al., 2019)
- 107 languages from 21 language families;
- concept aligned word lists for these languages.

Wikipedia
- 41 typologically diverse languages;
- no phonetic information (only graphemes)
Results

From the controlled metrics we see:

- a cross-linguistic tendency to front-load disambiguatory information
- not a universal phenomena—some languages have more informative word-final segments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Languages</th>
<th>Forward</th>
<th>Backward</th>
<th>Unigram</th>
<th>Position-Specific</th>
<th>Cloze</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELEX</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>NorthEuraLex</td>
<td>107</td>
<td>106</td>
<td>11</td>
<td>71</td>
<td>24</td>
<td>45</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>41</td>
<td>41</td>
<td>0</td>
<td>39</td>
<td>31</td>
<td>35</td>
</tr>
</tbody>
</table>
Results

From the controlled metrics we see:

- a cross-linguistic tendency to front-load disambiguatory information
- not a universal phenomena—some languages have more informative word-final segments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Languages</th>
<th>Forward</th>
<th>Backward</th>
<th>Unigram</th>
<th>Position-Specific</th>
<th>Cloze</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELEX</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>NorthEuraLex</td>
<td>107</td>
<td>106</td>
<td>0</td>
<td>11</td>
<td>31</td>
<td>71</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>41</td>
<td>41</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>39</td>
</tr>
</tbody>
</table>
Results

From the controlled metrics we see:

- a cross-linguistic tendency to front-load disambiguatory information
- not a universal phenomena—some languages have more informative word-final segments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Languages</th>
<th>Forward</th>
<th>Backward</th>
<th>Unigram</th>
<th>Position-Specific</th>
<th>Cloze</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELEX</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>NorthEuraLex</td>
<td>107</td>
<td>106</td>
<td>0</td>
<td>11</td>
<td>31</td>
<td>71</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>41</td>
<td>41</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>39</td>
</tr>
</tbody>
</table>
Results

From the controlled metrics we see:

- a cross-linguistic tendency to front-load disambiguatory information
- not a universal phenomena—some languages have more informative word-final segments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Languages</th>
<th>Forward</th>
<th>Backward</th>
<th>Unigram</th>
<th>Position-Specific</th>
<th>Cloze</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELEX</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>NorthEuraLex</td>
<td>107</td>
<td>106</td>
<td>0</td>
<td>11</td>
<td>31</td>
<td>71</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>41</td>
<td>41</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>39</td>
</tr>
</tbody>
</table>
Results

From the controlled metrics we see:

- a cross-linguistic tendency to front-load disambiguatory information
- not a universal phenomena—some languages have more informative word-final segments

Prior work investigated the ease of recovering lexical items from its beginnings or endings:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Languages</th>
<th>Forward</th>
<th>Backward</th>
<th>Unigram</th>
<th>Position-Specific</th>
<th>Cloze</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELEX</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>NorthEuraLex</td>
<td>107</td>
<td>106</td>
<td>0</td>
<td>11</td>
<td>31</td>
<td>45</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>41</td>
<td>41</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>35</td>
</tr>
</tbody>
</table>
Results

From the controlled metrics we see:

- a cross-linguistic tendency to front-load disambiguatory information
- not a universal phenomena—some languages have more informative word-final segments

Prior work investigated the ease of recovering lexical items from its beginnings or endings:

- Nooteboom (1981) found it is easier to recover words from their beginnings!

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Languages</th>
<th>Forward</th>
<th>Backward</th>
<th>Unigram</th>
<th>Position-Specific</th>
<th>Cloze</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELEX</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>NorthEuraLex</td>
<td>107</td>
<td>106</td>
<td>0</td>
<td>11</td>
<td>31</td>
<td>45</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>41</td>
<td>41</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>35</td>
</tr>
</tbody>
</table>
Prior work investigated the ease of recovering lexical items from its beginnings or endings:

- Nooteboom (1981) found it is easier to recover words from their beginnings!
- Nooteboom and van der Vlugt's (1988) found difference vanishes when you prime people with the length of the word.
An Informative Exploration of the Lexicon

Tiago Pimentel

Part 3
Summary

Four parts:

● Phonotactic Complexity and Its Trade-offs
● Disambiguatory Signals are Stronger in Word-initial Positions
● (Non-)Arbitrariness of the Sign
  ○ Meaning to Form: Measuring Systematicity as Information
  ○ Finding Concept-specific Biases in Form–Meaning Associations
● Speakers Fill Lexical Semantic Gaps with Context
(Non-)Arbitrariness of the Sign

+ Arya D. McCarthy; Brian Roark; Søren Wichmann; Damián Blasi; Ryan Cotterell
Saussure claimed the association between wordforms and meanings is arbitrary.

- The arbitrariness of the sign;
- Example: Why is dog called cachorro in Portuguese?
- Signifiers are arbitrarily attached to signifieds.
Introduction

There are small but systematic patterns in these connections:

- Iconicity: Word forms that “resemble” their meanings, e.g. *meow*
- Systematicity of the sign: Similar meanings are more likely to have similar forms.
- Phonesthemes: Sub-morphemic units which are associated with some small semantic domain.
Meaning to Form: Measuring Systematicity as Information

Tiago Pimentel, Arya D. McCarthy, Damián Blasi, Brian Roark, Ryan Cotterell
Can we quantify a language's systematicity of the sign?
Prior Work

Pearson correlation between word-pair distances:

- Phonological distance: raw word form edit distance.
- Semantic distance: word2vec cosine distance.

Problems:

- Hand defined distance metrics;
- Only linear relations between distances;
- No control for other factors (e.g. part-of-speech)
Our work

We define systematicity as mutual information:

$$\text{MI}(\text{meanings}; \text{forms}) = H(\text{forms}) - H(\text{forms} | \text{meanings})$$

What does meaning tell us about forms?

Overall form uncertainty

Form uncertainty given meaning
Our work

Advantages:

- No need to define distance metrics;
- Capture non-linear interactions;
- Straightforward to control for other factors;

\[ \text{MI}\left(\text{meanings}; \text{forms} \mid \text{POS}\right) = \text{H}\left(\text{forms} \mid \text{POS}\right) - \text{H}\left(\text{forms} \mid \text{meanings}, \text{POS}\right) \]

But, how can we measure \(\text{H}\left(\text{forms}\right)\) and \(\text{H}\left(\text{forms} \mid \text{meanings}\right)\)?
Our work

We use two LSTMs to get the language’s entropy

1. $H(\text{forms})$:
   - Predict phone given previous ones;
   - $p_\theta(\text{form}) = \prod p_\theta(w_t | w_{t-1})$
   - $H(\text{forms}) \leq H_\theta(\text{forms}) \approx - \sum \log p_\theta(\text{form}) / N$
Our work

We use two LSTMs to get the language’s entropy

2. $H(\text{forms} \mid \text{meanings})$

- Condition LSTM on meaning (word2vec embedding);
- $p_\theta(\text{form} \mid \text{meaning}) = \prod p_\theta(w_t \mid w_{t-1}, m)$
- $H_\theta(\text{forms} \mid \text{meanings}) \approx - \sum \log p_\theta(\text{form} \mid \text{meaning}) / N$
Our work

We now estimate the MI with the cross-entropies:

\[ \text{MI}(\text{meanings}; \text{forms}) \approx H_\theta(\text{forms}) - H_\theta(\text{forms} | \text{meanings}) \]

We also compute the uncertainty coefficient:

\[ \text{Unc}(\text{forms} | \text{meanings}) = \frac{\text{MI}(\text{meanings}; \text{forms})}{H(\text{forms})} \]
Results - CELEX

Used only monomorphemic words.

Results:

- Statistically significant systematicity in all three languages.
- Systematicity effect is reduced when we condition on POS.
Results - NorthEuraLex

Lexicon consists of “basic” concepts;

- We assume words are not multi-morphemic.

Use word2vec trained in English for all languages;

- Hard to train vectors for some languages.
Results - NorthEuraLex

Lexicon consists of “basic” concepts;

- We assume words are not multi-morphemic.

Use word2vec trained in English for all languages;

- Hard to train vectors for some languages.

Results:

- Significant systematicity in 87 of 106 languages;
- When we condition on POS tags, only 17 are statistically significant;
- Important to consider grammatical class on analysis.
Phonesthemes

Submorphemic affixal units

Usually flag a relatively small semantic domain

Classic example (Bergen, 2004):

- \textit{gl-}
- related to light or vision;
- glimmer, glisten, glitter, gleam, glow and glint.

Should have higher mutual information values when compared to other k-grams.
Results:

- We can find lists of known phonesthemes:
- All but two of our English phonesthemes are attested in prior work.
- Also find affixes which are pieces of fossilized morphology.

<table>
<thead>
<tr>
<th>Language</th>
<th>Phonestheme</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch</td>
<td>/sx/-</td>
<td>schelp, schild, schot, schacht, schaar</td>
</tr>
<tr>
<td></td>
<td>-/sI/</td>
<td>kegel, nevel, beitel, vleugel, zetel</td>
</tr>
<tr>
<td></td>
<td>-/xt/</td>
<td>beicht, nacht, vocht, plicht, licht</td>
</tr>
<tr>
<td></td>
<td>-/xp/</td>
<td>stop, shop, drop, top, bob</td>
</tr>
<tr>
<td>English</td>
<td>/m/-</td>
<td>infidel, intellect, institute, enigma, interim</td>
</tr>
<tr>
<td></td>
<td>/sI/</td>
<td>slop, slough, sluice, slim, slush</td>
</tr>
<tr>
<td></td>
<td>-/kt/</td>
<td>aspect, object, fact, viaduct, tact</td>
</tr>
<tr>
<td></td>
<td>-/mx/</td>
<td>panorama, asthma, trachoma, eczema, magma</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Phonestheme</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>/pa/-</td>
<td>geschehen, Gebiet, gering, Geruecht, gesinnt</td>
</tr>
<tr>
<td></td>
<td>-/an/</td>
<td>rascheln, rumpeln, tummeln, torkeln, mogeln</td>
</tr>
<tr>
<td></td>
<td>-/en/</td>
<td>rascheln, rumpeln, tummeln, torkeln, mogeln</td>
</tr>
<tr>
<td></td>
<td>-/en/</td>
<td>goennen, saeen, besuchen, giessen, streiten</td>
</tr>
</tbody>
</table>
Finding Concept-specific Biases in Form–Meaning Associations

Tiago Pimentel, Brian Roark, Søren Wichmann, Ryan Cotterell, Damián Blasi
Research Question

Are there cross-linguistic associations between the forms and meanings of words? And how do we find them?
Example

The word for "tongue" is more likely than chance to contain the phone [l]

Very biased map with hand-picked languages!
Data - ASJP

- Basic vocabulary wordlists
- Almost $\frac{3}{4}$ of world’s languages (5189)!
- 100 basic concepts
  - body parts, colour terms, lower numerals, general properties (big, round), and some common flora and fauna (e.g. trees and dogs)
Method

We use the same definition as before:

$$\text{MI}(\text{meanings}; \text{forms}) = H(\text{forms}) - H(\text{forms} | \text{meanings})$$

How do we compute this?

$$H(\text{forms}) \leq H_\theta(\text{forms}) \approx - \sum \log p_\theta(\text{form}) / N$$

$$H(\text{forms} | \text{meanings}) \leq - \sum \log p_\theta(\text{form} | \text{meaning}) / N$$

Cross-entropy

Issue: Cross-entropy needs to be computed on data independent from training!

Languages have been in contact (not i.i.d.)!
Method

To maximize independence, we split our data per macro-area.

- 2 areas for training, 1 development, 1 test;
- Cross-validation.
Method

Some language families cross macro-areas:

- Group them in the macro-area with more of the family’s languages
Results

A very small average contribution of meaning into form.

- Approximately 0.3% (it was 3~5% within languages).
- Significant in half the macro-areas.

There are very small, but significant biases. But we draw no strong conclusions at this level of the analysis.

| Region          | H(W) | MI(W;V) | U(W|V) |
|-----------------|------|---------|--------|
| Pacific, Americas | 3.773 | 0.011 | 0.28%  |
| Americas, Eurasia | 3.901 | 0.007 | 0.17%  |
| Americas, Pacific | 3.999 | 0.015‡ | 0.38%  |
| Europe, Africa   | 3.755 | 0.016‡ | 0.42%  |
| Average          | 3.857 | 0.012‡ | 0.31%  |
Results - Per concept

- Out of 100 concepts, 26 have significantly positive MI;
  - Pronouns present the highest values.
  - Most colours show statistically positive MI.
  - Some concepts related to body parts and several concepts related to the environment have statistically positive results.

Effect seems to be driven mostly by a subset of the concepts.
Results - Per language

● Out of 5189 languages, 85 have significantly positive MI;
  ○ At most 100 data points per language, hard statistical test after the corrections for multiple testing.
  ○ Cross-linguistic form–meaning biases are *potentially* not as rare or weak as believed. We can get significant language-level results with at most 100 concepts.
  ○ Further studies needed for stronger conclusions.
Results - Per concept–token pair

- Concept–token pairs with particularly large MI across all four macro-areas.
  - # is the end-of-string token.
  - Associations between [l] and “tongue” and between [p] and “full” (Blasi et al., 2016)
  - Associations between [m] and [u] and “breast” (Jakobson, 1960; Traunmüller, 1994)
  - Pronouns—e.g. I, we, you—and end-of-string [#]

<table>
<thead>
<tr>
<th>Concept</th>
<th>Tokens</th>
<th>Concept</th>
<th>Tokens</th>
<th>Concept</th>
<th>Tokens</th>
<th>Concept</th>
<th>Tokens</th>
<th>Concept</th>
<th>Tokens</th>
<th>Concept</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>blood</td>
<td>s</td>
<td>eye</td>
<td>i</td>
<td>liver</td>
<td>k l r t</td>
<td>path</td>
<td>d t</td>
<td>tree</td>
<td>#</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bone</td>
<td>s u</td>
<td>fire</td>
<td># t</td>
<td>louse</td>
<td>m n</td>
<td>say</td>
<td>#</td>
<td>two</td>
<td>r</td>
<td></td>
<td></td>
</tr>
<tr>
<td>breast</td>
<td>m u</td>
<td>fish</td>
<td>a s</td>
<td>mountain</td>
<td>b d g l o r</td>
<td>see</td>
<td># e</td>
<td>water</td>
<td>#</td>
<td></td>
<td></td>
</tr>
<tr>
<td>come</td>
<td># e</td>
<td>full</td>
<td>l o p t</td>
<td>name</td>
<td># i</td>
<td>skin</td>
<td>k l p r t</td>
<td>we</td>
<td># e i n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>die</td>
<td>t</td>
<td>give</td>
<td>#</td>
<td>neck</td>
<td>o</td>
<td>star</td>
<td>k l o r s t u w</td>
<td>who</td>
<td>#</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td>k</td>
<td>horn</td>
<td>a r</td>
<td>new</td>
<td>a</td>
<td>stone</td>
<td>k t</td>
<td>you</td>
<td># a i n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>drink</td>
<td># u</td>
<td>I</td>
<td># a n</td>
<td>night</td>
<td>N d i l m p r t u</td>
<td>tongue</td>
<td>d e l n r</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ear</td>
<td>e l t</td>
<td>knee</td>
<td>N b g k m o r t u</td>
<td>nose</td>
<td>N i u</td>
<td>tooth</td>
<td>e i</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eat</td>
<td>#</td>
<td>leaf</td>
<td>a l p t</td>
<td>one</td>
<td>k t</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
An Informative Exploration of the Lexicon

Tiago Pimentel

Part 4
Summary

Four parts:

- Phonotactic Complexity and Its Trade-offs
- Disambiguatory Signals are Stronger in Word-initial Positions
- (Non-)Arbitrariness of the Sign
  - Meaning to Form: Measuring Systematicity as Information
  - Finding Concept-specific Biases in Form–Meaning Associations
- Speakers Fill Lexical Semantic Gaps with Context
Speakers Fill Lexical Semantic Gaps with Context

Tiago Pimentel, Rowan Hall Maudslay, Damián Blasi, Ryan Cotterell
Lexical Ambiguity

Words can mean more than one thing 😲

Consider the English word *buffalo*:

- You can pet a large buffalo (animal);
- You can visit Buffalo (US city);
- You can buffalo (intimidate) a person;
Lexical Ambiguity

Words can mean more than one thing 😐

Consider the English word *buffalo*:

- You can pet a large buffalo (animal);
- You can visit Buffalo (US city);
- You can buffalo (intimidate) a person;

Buffalo buffalo buffalo Buffalo buffalo!

- Paraphrased as NY bison intimidate other NY bison
The Good Linguistic Question

Do speakers compensate for **lexical ambiguity** by making **words more predictable** (i.e. less uncertain) given their **context** in order to accommodate the listeners?

Don't want to make the listener's job too hard!
The Good Linguistic Question

Do speakers compensate for **lexical ambiguity** by making **words more predictable** (i.e. less uncertain) given their **context** in order to accommodate the listeners?

→ **Put differently**: Is there a **negative correlation** between **contextual uncertainty** and **lexical ambiguity**?
The Good Linguistic Question

Do speakers compensate for *lexical ambiguity* by making *words more predictable* (i.e. less uncertain) given their *context* in order to accommodate the listeners?

→ Put differently: Is there a *negative* correlation between *contextual uncertainty* and *lexical ambiguity*?

Side note: We do not test for causality in any form. Only correlation!
Our Operationalisations
A Measure of Lexical Ambiguity

We operationalise **lexical ambiguity** as the half-pointwise entropy:

\[ H(M \mid W=w) \]
A Measure of Lexical Ambiguity

We operationalise *lexical ambiguity* as the half-pointwise entropy:

\[ H(M \mid W=w) \]
We operationalise lexical ambiguity as the half-pointwise entropy:

$$H(M \mid W=w)$$

Equivalent, up to an additive constant, to a mutual information (MI)

$$I(M; W = w) = H(M) - H(M \mid W = w)$$

**A Measure of Lexical Ambiguity**
We operationalise lexical ambiguity as the half-pointwise entropy:

$$H(M | W=w)$$

Equivalent, up to an additive constant, to a mutual information (MI):

$$I(M; W = w) = H(M) - H(M | W = w)$$
How to Measure Lexical Ambiguity?

WordNet

BERT
How to Measure Lexical Ambiguity?

WordNet
- Discrete senses

BERT
- Continuous-meaning space
How to Measure Lexical Ambiguity?

WordNet
- Discrete senses
- Hand-annotated

BERT
- Continuous-meaning space
- No hand annotation required!
How to Measure Lexical Ambiguity?

WordNet
- Discrete senses
- Hand-annotated
- Only available in high-resource languages

BERT
- Continuous-meaning space
- No hand annotation required!
- Easily obtainable for new languages
How to Measure Lexical Ambiguity?

**WordNet**
- Discrete senses
- Hand-annotated
- Only available in high-resource languages
- We assume an uniform distribution over senses

\[ H(M \mid W=w) \approx \log_2(\#senses[w]) \]

**BERT**
- Continuous-meaning space
- No hand annotation required!
- Easily obtainable for new languages
How to Measure Lexical Ambiguity?

**WordNet**
- Discrete senses
- Hand-annotated
- Only available in high-resource languages
- We assume an uniform distribution over senses

\[ H(M \mid W=w) \approx \log_2(#\text{senses}[w]) \]

**BERT**
- Continuous-meaning space
- No hand annotation required!
- Easily obtainable for new languages
- Assume embeddings are the word meaning: \( m \approx \text{BERT}(p \circ w \circ s) \)
How to Measure Lexical Ambiguity?

**WordNet**
- Discrete senses
- Hand-annotated
- Only available in high-resource languages
- We assume an uniform distribution over senses

\[ H(M \mid W=w) \approx \log_2 (#\text{senses}[w]) \]

**BERT**
- Continuous-meaning space
- No hand annotation required!
- Easily obtainable for new languages
- Assume embeddings are the word meaning: \( m \approx \text{BERT}(p \circ w \circ s) \)
- We use a Gaussian approximation (max-entropy upper bound)

\[ H(M \mid W=w) \approx H(N(\mu_w, \Sigma_w)) \]
\[ = \frac{1}{2} \log_2 \det (2\pi e\Sigma_w) \]
Validating our Measures of Lexical Ambiguity

How well do the BERT and WordNet measures of lexical ambiguity correlate \textit{with each other}?
Validating our Measures of Lexical Ambiguity

How well do the BERT and WordNet measures of lexical ambiguity correlate with each other?

→ Relatively well!

<table>
<thead>
<tr>
<th>Language</th>
<th># Types</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>836</td>
<td>0.25**</td>
<td>0.30**</td>
</tr>
<tr>
<td>English</td>
<td>6995</td>
<td>0.40**</td>
<td>0.40**</td>
</tr>
<tr>
<td>Finnish</td>
<td>1247</td>
<td>0.06*</td>
<td>0.07*</td>
</tr>
<tr>
<td>Indonesian</td>
<td>3308</td>
<td>0.12**</td>
<td>0.13**</td>
</tr>
<tr>
<td>Persian</td>
<td>2648</td>
<td>0.14**</td>
<td>0.13**</td>
</tr>
<tr>
<td>Portuguese</td>
<td>3285</td>
<td>0.13**</td>
<td>0.13**</td>
</tr>
</tbody>
</table>
Validating our Measures of Lexical Ambiguity

How well do the BERT and WordNet measures of lexical ambiguity correlate *with each other*?

→ Relatively well!

<table>
<thead>
<tr>
<th>Language</th>
<th># Types</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>836</td>
<td>0.25**</td>
<td>0.30**</td>
</tr>
<tr>
<td>English</td>
<td>6995</td>
<td>0.40**</td>
<td>0.40**</td>
</tr>
<tr>
<td>Finnish</td>
<td>1247</td>
<td>0.06*</td>
<td>0.07*</td>
</tr>
<tr>
<td>Indonesian</td>
<td>3308</td>
<td>0.12**</td>
<td>0.13**</td>
</tr>
<tr>
<td>Persian</td>
<td>2648</td>
<td>0.14**</td>
<td>0.13**</td>
</tr>
<tr>
<td>Portuguese</td>
<td>3285</td>
<td>0.13**</td>
<td>0.13**</td>
</tr>
</tbody>
</table>
Validating our Measures of Lexical Ambiguity

How well do the BERT and WordNet measures of lexical ambiguity correlate with each other?

→ Relatively well!

<table>
<thead>
<tr>
<th>Language</th>
<th># Types</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>836</td>
<td>0.25**</td>
<td>0.30**</td>
</tr>
<tr>
<td>English</td>
<td>6995</td>
<td>0.40**</td>
<td>0.40**</td>
</tr>
<tr>
<td>Finnish</td>
<td>1247</td>
<td>0.06*</td>
<td>0.07*</td>
</tr>
<tr>
<td>Indonesian</td>
<td>3308</td>
<td>0.12**</td>
<td></td>
</tr>
<tr>
<td>Persian</td>
<td>2648</td>
<td>0.14**</td>
<td></td>
</tr>
<tr>
<td>Portuguese</td>
<td>3285</td>
<td>0.13**</td>
<td>0.13**</td>
</tr>
</tbody>
</table>

Side note: Correlations seem stronger when either BERT or WordNet are better.
A Measure of Contextual Uncertainty

We operationalise *contextual uncertainty* as the half-pointwise entropy:

\[ H(W=w \mid C) \]
A Measure of Contextual Uncertainty

We operationalise contextual uncertainty as the half-pointwise entropy:

$$H(W=w \mid C)$$

Average uncertainty of a word in all its contexts
A Measure of Contextual Uncertainty

We operationalise contextual uncertainty as the half-pointwise entropy:

$$H(W=w | C)$$

Average uncertainty of a word in all its contexts

May be approximated with a cloze language model

- This uses bidirectional context, which is different than most previous work
Our Empirical Findings
Empirical Evidence for the Trade-off

In WordNet
Empirical Evidence for the Trade-off

In WordNet

- Yes, in 5 of the 6 analysed languages!

<table>
<thead>
<tr>
<th>Language</th>
<th># Types</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>(ar)</td>
<td>836</td>
<td>-0.14**</td>
</tr>
<tr>
<td>English</td>
<td>(en)</td>
<td>6995</td>
<td>-0.07**</td>
</tr>
<tr>
<td>Finnish</td>
<td>(fi)</td>
<td>1247</td>
<td>0.01</td>
</tr>
<tr>
<td>Indonesian</td>
<td>(id)</td>
<td>3308</td>
<td>-0.09**</td>
</tr>
<tr>
<td>Persian</td>
<td>(fa)</td>
<td>2648</td>
<td>-0.11**</td>
</tr>
<tr>
<td>Portuguese</td>
<td>(pt)</td>
<td>3285</td>
<td>-0.10**</td>
</tr>
</tbody>
</table>
Empirical Evidence for the Trade-off

In WordNet

- Yes, in 5 of the 6 analysed languages!

<table>
<thead>
<tr>
<th>Language</th>
<th># Types</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>836</td>
<td>-0.14**</td>
<td>-0.15**</td>
</tr>
<tr>
<td>English</td>
<td>6995</td>
<td>-0.07**</td>
<td>-0.11**</td>
</tr>
<tr>
<td>Finnish</td>
<td>1247</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>Indonesian</td>
<td>3308</td>
<td>-0.09**</td>
<td>-0.14**</td>
</tr>
<tr>
<td>Persian</td>
<td>2648</td>
<td>-0.11**</td>
<td>-0.12**</td>
</tr>
<tr>
<td>Portuguese</td>
<td>3285</td>
<td>-0.10**</td>
<td>-0.11**</td>
</tr>
</tbody>
</table>
Empirical Evidence for the Trade-off

In BERT
Empirical Evidence for the Trade-off

In BERT

- Yes, in all the 18 analysed languages!

<table>
<thead>
<tr>
<th>Language</th>
<th># Types</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afrikaans (af)</td>
<td>4505</td>
<td>-0.41**</td>
<td>-0.52**</td>
</tr>
<tr>
<td>Arabic (ar)</td>
<td>10181</td>
<td>-0.33**</td>
<td>-0.41**</td>
</tr>
<tr>
<td>Bengali (bn)</td>
<td>8128</td>
<td>-0.43**</td>
<td>-0.44**</td>
</tr>
<tr>
<td>English (en)</td>
<td>7097</td>
<td>-0.33**</td>
<td>-0.35**</td>
</tr>
<tr>
<td>Estonian (et)</td>
<td>4482</td>
<td>-0.40**</td>
<td>-0.44**</td>
</tr>
<tr>
<td>Finnish (fi)</td>
<td>3928</td>
<td>-0.38**</td>
<td>-0.45**</td>
</tr>
<tr>
<td>Hebrew (he)</td>
<td>13819</td>
<td>-0.34**</td>
<td>-0.37**</td>
</tr>
<tr>
<td>Indonesian (id)</td>
<td>4524</td>
<td>-0.45**</td>
<td>-0.57**</td>
</tr>
<tr>
<td>Icelandic (is)</td>
<td>3578</td>
<td>-0.44**</td>
<td>-0.46**</td>
</tr>
<tr>
<td>Kannada (kn)</td>
<td>9695</td>
<td>-0.42**</td>
<td>-0.41**</td>
</tr>
<tr>
<td>Malayalam (ml)</td>
<td>6203</td>
<td>-0.47**</td>
<td>-0.46**</td>
</tr>
<tr>
<td>Marathi (mr)</td>
<td>5821</td>
<td>-0.39**</td>
<td>-0.40**</td>
</tr>
<tr>
<td>Persian (fa)</td>
<td>6788</td>
<td>-0.39**</td>
<td>-0.49**</td>
</tr>
<tr>
<td>Portuguese (pt)</td>
<td>5685</td>
<td>-0.31**</td>
<td>-0.45**</td>
</tr>
<tr>
<td>Tagalog (tl)</td>
<td>3332</td>
<td>-0.45**</td>
<td>-0.50**</td>
</tr>
<tr>
<td>Turkish (tr)</td>
<td>4386</td>
<td>-0.40**</td>
<td>-0.46**</td>
</tr>
<tr>
<td>Tatar (tt)</td>
<td>2997</td>
<td>-0.34**</td>
<td>-0.39**</td>
</tr>
<tr>
<td>Yoruba (yo)</td>
<td>417</td>
<td>-0.55**</td>
<td>-0.64**</td>
</tr>
</tbody>
</table>
Empirical Evidence for the Trade-off

In BERT

- Yes, in all the 18 analysed languages!

<table>
<thead>
<tr>
<th>Language</th>
<th># Types</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afrikaans</td>
<td>(af)</td>
<td>-0.41**</td>
<td>-0.52**</td>
</tr>
<tr>
<td>Arabic</td>
<td>(ar)</td>
<td>-0.33**</td>
<td>-0.41**</td>
</tr>
<tr>
<td>Bengali</td>
<td>(bn)</td>
<td>-0.43**</td>
<td>-0.44**</td>
</tr>
<tr>
<td>English</td>
<td>(en)</td>
<td>-0.33**</td>
<td>-0.35**</td>
</tr>
<tr>
<td>Estonian</td>
<td>(et)</td>
<td>-0.40**</td>
<td>-0.44**</td>
</tr>
<tr>
<td>Finnish</td>
<td>(fi)</td>
<td>-0.38**</td>
<td>-0.45**</td>
</tr>
<tr>
<td>Hebrew</td>
<td>(he)</td>
<td>-0.34**</td>
<td>-0.37**</td>
</tr>
<tr>
<td>Indonesian</td>
<td>(id)</td>
<td>-0.45**</td>
<td>-0.46**</td>
</tr>
<tr>
<td>Icelandic</td>
<td>(is)</td>
<td>-0.44**</td>
<td>-0.46**</td>
</tr>
<tr>
<td>Kannada</td>
<td>(kn)</td>
<td>-0.42**</td>
<td>-0.41**</td>
</tr>
<tr>
<td>Malayalam</td>
<td>(ml)</td>
<td>-0.47**</td>
<td>-0.46**</td>
</tr>
<tr>
<td>Marathi</td>
<td>(mr)</td>
<td>-0.39**</td>
<td>-0.40**</td>
</tr>
<tr>
<td>Persian</td>
<td>(fa)</td>
<td>-0.39**</td>
<td>-0.49**</td>
</tr>
<tr>
<td>Portuguese</td>
<td>(pt)</td>
<td>-0.42**</td>
<td>-0.41**</td>
</tr>
<tr>
<td>Tagalog</td>
<td>(tl)</td>
<td>-0.47**</td>
<td>-0.46**</td>
</tr>
<tr>
<td>Yoruba</td>
<td>(yo)</td>
<td>-0.34**</td>
<td>-0.39**</td>
</tr>
</tbody>
</table>

Lexical Ambiguity (bits)

Contextual Uncertainty (bits)
Empirical Evidence for the Trade-off

In BERT

- Yes, in all the 18 analysed languages!

### Table: Language Analysis

<table>
<thead>
<tr>
<th>Language</th>
<th># Types</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afrikaans (af)</td>
<td>4505</td>
<td>-0.41**</td>
<td>-0.52**</td>
</tr>
<tr>
<td>Arabic (ar)</td>
<td>10181</td>
<td>-0.33**</td>
<td>-0.41**</td>
</tr>
<tr>
<td>Bengali (bn)</td>
<td>8128</td>
<td>-0.43**</td>
<td>-0.44**</td>
</tr>
<tr>
<td>English (en)</td>
<td>7097</td>
<td>-0.33**</td>
<td>-0.35**</td>
</tr>
<tr>
<td>Estonian (et)</td>
<td>4482</td>
<td>-0.40**</td>
<td>-0.44**</td>
</tr>
<tr>
<td>Finnish (fi)</td>
<td>3928</td>
<td>-0.38**</td>
<td>-0.45**</td>
</tr>
<tr>
<td>Hebrew (he)</td>
<td>13910</td>
<td>-0.34**</td>
<td>-0.37**</td>
</tr>
<tr>
<td>Persian (fa)</td>
<td>6788</td>
<td>-0.39**</td>
<td>-0.49**</td>
</tr>
<tr>
<td>Portuguese (pt)</td>
<td>5685</td>
<td>-0.31**</td>
<td>-0.45**</td>
</tr>
<tr>
<td>Tagalog (tl)</td>
<td>3332</td>
<td>-0.45**</td>
<td>-0.50**</td>
</tr>
<tr>
<td>Tatar (tt)</td>
<td>2997</td>
<td>-0.34**</td>
<td>-0.39**</td>
</tr>
<tr>
<td>Yoruba (yo)</td>
<td>417</td>
<td>-0.55**</td>
<td>-0.64**</td>
</tr>
</tbody>
</table>

---

![Graphs showing lexical ambiguity vs contextual uncertainty for Arabic, English, Malayalam, and Tagalog languages.]
A Functionalist Derivation of the Trade-Off
Clarity

- **Clarity** is the functionalist principle that a listener be able to reconstruct the speaker’s intended meaning.
Clarity

- **Clarity** is the functionalist principle that a listener be able to reconstruct the speaker’s intended meaning.

- Information-theoretically, we operationalise *clarity* as:

\[ H(M | W, C) \]
Clarity

- **Clarity** is the functionalist principle that a listener be able to reconstruct the speaker’s intended meaning.

- Information-theoretically, we operationalise **clarity** as:

  \[ H(M \mid W, C) \]

  which is the uncertainty of the meaning, given the context and the word.
Robustness

- Robustness is the functionalist principle that a speaker’s utterance should be resilient to noise.
Robustness

- **Robustness** is the functionalist principle that a speaker’s utterance should be resilient to noise

- We operationalise *robustness* as a tripartite MI:

  $$I(M; C; W = w)$$
Robustness

- Robustness is the functionalist principle that a speaker’s utterance should be resilient to noise.

- We operationalise robustness as a tripartite MI: 
  \[ I(M; C; W = w) \]

  which is the redundant information shared by meaning, context and word.
Why Should There Be a Trade-off?

- Assume language is clear, i.e. $H(M | W, C) = 0$
Why Should There Be a Trade-off?

- Assume language is **clear**, i.e. $H(M \mid W, C) = 0$
- Assume language is **robust**, i.e. $I(M; C; W = w) \geq k$
Why Should There Be a Trade-off?

- Assume language is **clear**, i.e. $H(M \mid W, C) = 0$
- Assume language is **robust**, i.e. $I(M; C; W = w) \geq k$

constant
Why Should There Be a Trade-off?

- Assume language is **clear**, i.e. $H(M \mid W, C) = 0$
- Assume language is **robust**, i.e. $I(M; C; W = w) \geq k$
- We show:

$$I(M; C; W = w) = H(M) - H(M \mid W = w) - H(W = w \mid C)$$
Why Should There Be a Trade-off?

- Assume language is clear, i.e. $H(M \mid W, C) = 0$
- Assume language is robust, i.e. $I(M; C; W = w) \geq k$

We show:

$$I(M; C; W = w) = H(M) - H(M \mid W = w) - H(W = w \mid C)$$

constant
Why Should There Be a Trade-off?

- Assume language is clear, i.e. $H(M \mid W, C) = 0$
- Assume language is robust, i.e. $I(M; C; W = w) \geq k$

- We show:
  
  \[ I(M; C; W = w) = H(M) - H(M \mid W = w) - H(W = w \mid C) \]

- Thus, $H(M \mid W = w) + H(W = w \mid C) \leq H(M) - k$
Why Should There Be a Trade-off?

- Assume language is **clear**, i.e. $H(M | W, C) = 0$
- Assume language is **robust**, i.e. $I(M; C; W = w) \geq k$

- We show:
  
  $$I(M; C; W = w) = H(M) - H(M | W = w) - H(W = w | C)$$

- Thus, $H(M | W = w) + H(W = w | C) \leq H(M) - k$

There is a trade-off!!!

- Lexical ambiguity
- Contextual uncertainty
Takeaways
| Language  | H(W)   | MI(W;V) | U(W|V)  |
|----------|--------|---------|---------|
| English  | 3.401  | 0.11    | 3.24%   |
| German   | 3.195  | 0.168   | 5.26%   |
| Dutch    | 3.245  | 0.156   | 4.82%   |

| Test     | H(W)   | MI(W;V) | U(W|V)  |
|----------|--------|---------|---------|
| Africa   | 3.773  | 0.011   | 0.28%   |
| Americas | 3.901  | 0.007   | 0.17%   |
| Eurasia  | 3.999  | 0.015   | 0.38%   |
| Pacific  | 3.755  | 0.016   | 0.42%   |
| Average  | 3.857  | 0.012   | 0.31%   |
Phonotactic complexity correlates with word length
- Pressure towards a roughly equal sized phonotactic space?
### Contextual Uncertainty (bits)

| Language | $H(W)$ | $MI(W;V)$ | $U(W|V)$ |
|----------|-------|---------|--------|
| English  | 3.401 | 0.11   | 3.24%  |
| German   | 3.195 | 0.168  | 5.26%  |
| Dutch    | 3.245 | 0.156  | 4.82%  |

### Lexical Ambiguity (bits)

| Test     | $H(W)$ | $MI(W;V)$ | $U(W|V)$ |
|----------|-------|---------|--------|
| Africa   | 3.773 | 0.011   | 0.28%  |
| Americas | 3.901 | 0.007   | 0.17%  |
| Eurasia  | 3.999 | 0.015* | 0.38%  |
| Pacific  | 3.755 | 0.016* | 0.42%  |
| Average  | 3.857 | 0.012* | 0.31%  |
Cross-linguistic tendency to frontload information

- Pressure towards faster lexical access?

| Language | H(W)  | MI(W;V) | U(W|V) |
|----------|-------|---------|--------|
| English  | 3.401 | 0.11    | 3.24%  |
| German   | 3.195 | 0.168   | 5.26%  |
| Dutch    | 3.245 | 0.156   | 4.82%  |

| Test     | H(W)   | MI(W;V) | U(W|V) |
|----------|--------|---------|--------|
| Africa   | 3.773  | 0.011   | 0.28%  |
| Americas | 3.901  | 0.007   | 0.17%  |
| Eurasia  | 3.999  | 0.015   | 0.38%  |
| Pacific  | 3.755  | 0.016   | 0.42%  |
| Average  | 3.857  | 0.012   | 0.31%  |
### Language

| Language | H(W)  | MI(W;V) | U(W|V)  |
|----------|-------|---------|---------|
| English  | 3.401 | 0.11    | 3.24%   |
| German   | 3.195 | 0.168   | 5.26%   |
| Dutch    | 3.245 | 0.156   | 4.82%   |

### Test

| Test    | H(W)  | MI(W;V) | U(W|V)  |
|---------|-------|---------|---------|
| Africa  | 3.773 | 0.011   | 0.28%   |
| Americas| 3.901 | 0.007   | 0.17%   |
| Eurasia | 3.999 | 0.015   | 0.38%   |
| Pacific | 3.755 | 0.016   | 0.42%   |
| Average | 3.857 | 0.012   | 0.31%   |
Speakers compensate lexical ambiguity by reducing contextual uncertainty.

- Pressure towards fewer communication errors?
Language | H(W) | MI(W;V) | U(W|V)  
--- | --- | --- | ---  
English | 3.401 | 0.11 | 3.24%  
German | 3.195 | 0.168 | 5.26%  
Dutch | 3.245 | 0.156 | 4.82%  

**Test** | H(W) | MI(W;V) | U(W|V)  
--- | --- | --- | ---  
Africa | 3.773 | 0.011 | 0.28%  
Americas | 3.901 | 0.007 | 0.17%  
Eurasia | 3.999 | 0.015 | 0.38%  
Pacific | 3.755 | 0.016 | 0.42%  
Average | 3.857 | 0.012 | 0.31%  

Contextual Uncertainty (bits) | Arabic | English | Malayalam | Tagalog  
--- | --- | --- | --- | ---  
Lexical Ambiguity (bits) | Contextual Uncertainty (bits)  

Final Surprisal (bits) | 2 | 3 | 4 | 5  
--- | --- | --- | --- | ---  
Initial Surprisal (bits)  

wikipedia | northeuralex | celex  
--- | --- | ---  

Cross Entropy (bits per phoneme) | Average Length (# IPA tokens)  
--- | ---  

Semantic similarities:

- Semantically similar words have more similar forms
- Pressure towards learnability?
Thanks :-) 

And thanks to all co-authors:
Rowan Hall Maudslay, Brian Roark, Damián Blasi, Ryan Cotterell, Arya D. McCarthy, Søren Wichmann