An Informative Exploration of the Lexicon

Tiago Pimentel

UNIVERSITY OF CAMBRIDGE

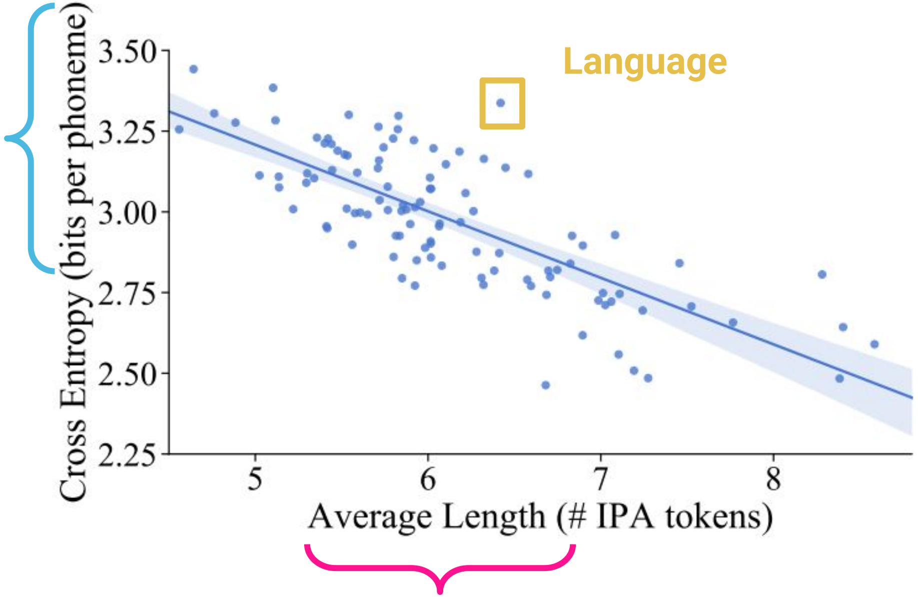
Summary

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

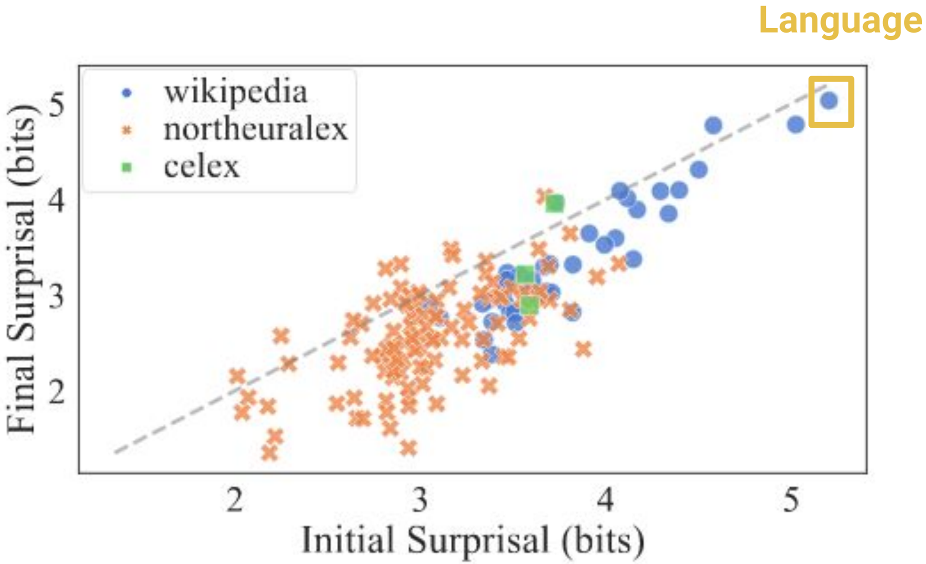
Spoiler

1. We use **bits per phoneme** as a measure of phonotactic complexity, and find it has **robust trade-offs** with **word length**.



Spoiler

2. We analyse **bits per phoneme** in first and second halves of words, and find a cross-linguistic **tendency** to **frontload information**.



Spoiler

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%

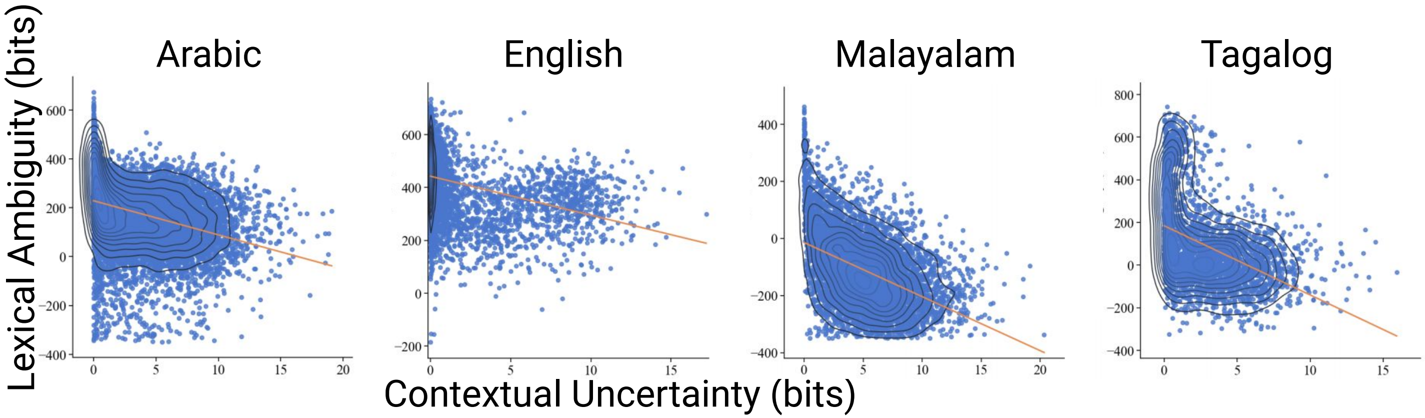
Within Languages

Across Languages

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%

Spoiler

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.

Phonological Complexity

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**:

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

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**:

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

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

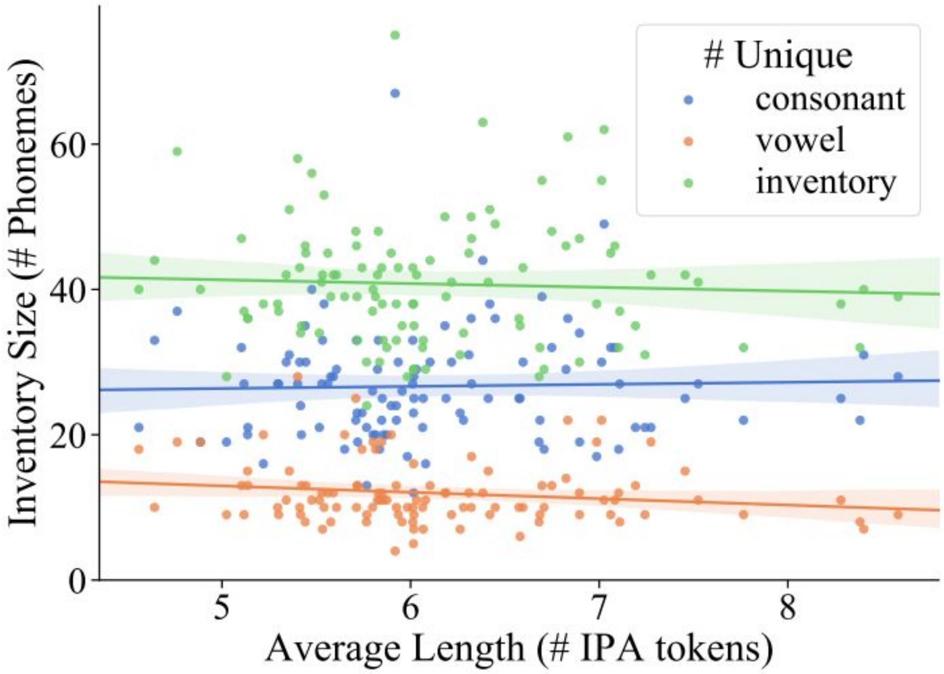
Composed of **“basic” concepts**

Languages from 21 families in Europe/Asia

We omit Mandarin (no tone annotation)

Concept	**Language**	**Word**	**IPA**
eye	portuguese	olho	/oʎu/
ear	finnish	korva	/kɔrʋɑ/
give	north karelian	antua	/ɑntʊɑ/
tooth	veps	hambaz	/hɑmbɑz/
black	northern sami	čáhppes	/t͡ʃaahppes/
immediately	hill mari	töpök	/tørøk/

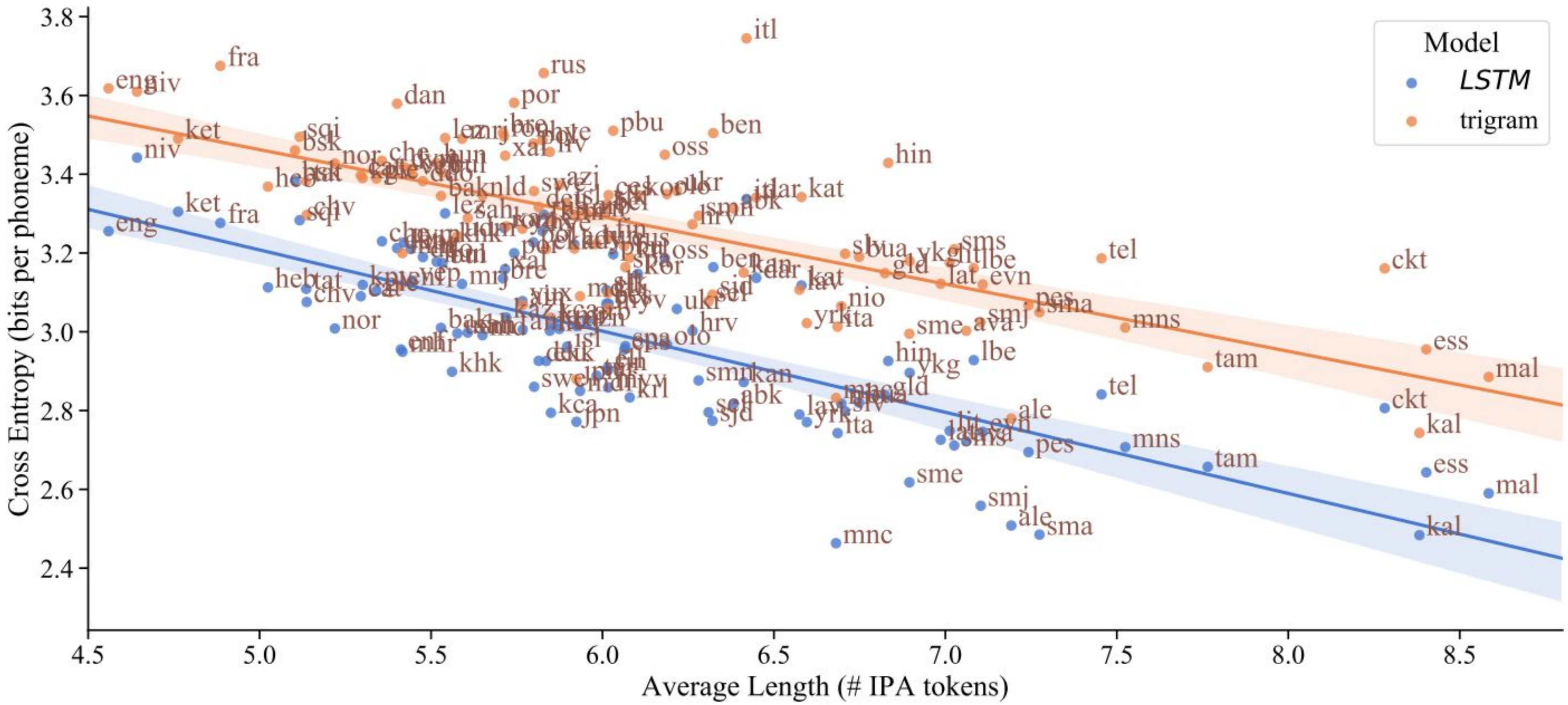
Trade-Off



Measure	Correlation	
	Pearson r	Spearman ρ
Number of:		
phonemes	-0.047	-0.054
vowels	-0.164	-0.162
consonants	0.030	0.045
Bits/phoneme:		
unigram	-0.217	-0.222
trigram	-0.682	-0.672
LSTM	-0.762	-0.744

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.

	Correlation	
Measure	Pearson r	Spearman ρ
Number of:		
phonemes	-0.214	-0.095
vowels	-0.383	-0.367
consonants	-0.147	-0.092
Bits/phoneme:		

	Spearman ρ		
Family	**LSTM**	**Vowels**	**# Langs**
Dravidian		-0.894	4
Indo-European		-0.218	37
Nakh-Daghestanian		-0.530	6
Turkic		$-0.773^{\dagger}$	8
Uralic		$0.363^{\dagger}$	26

* Statistically significant with $p < 0.01$

† Statistically significant with $p < 0.1$

An Informative Exploration of the Lexicon

Tiago Pimentel

Part 2

UNIVERSITY OF CAMBRIDGE

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

UNIVERSITY OF CAMBRIDGE

ETH zürich

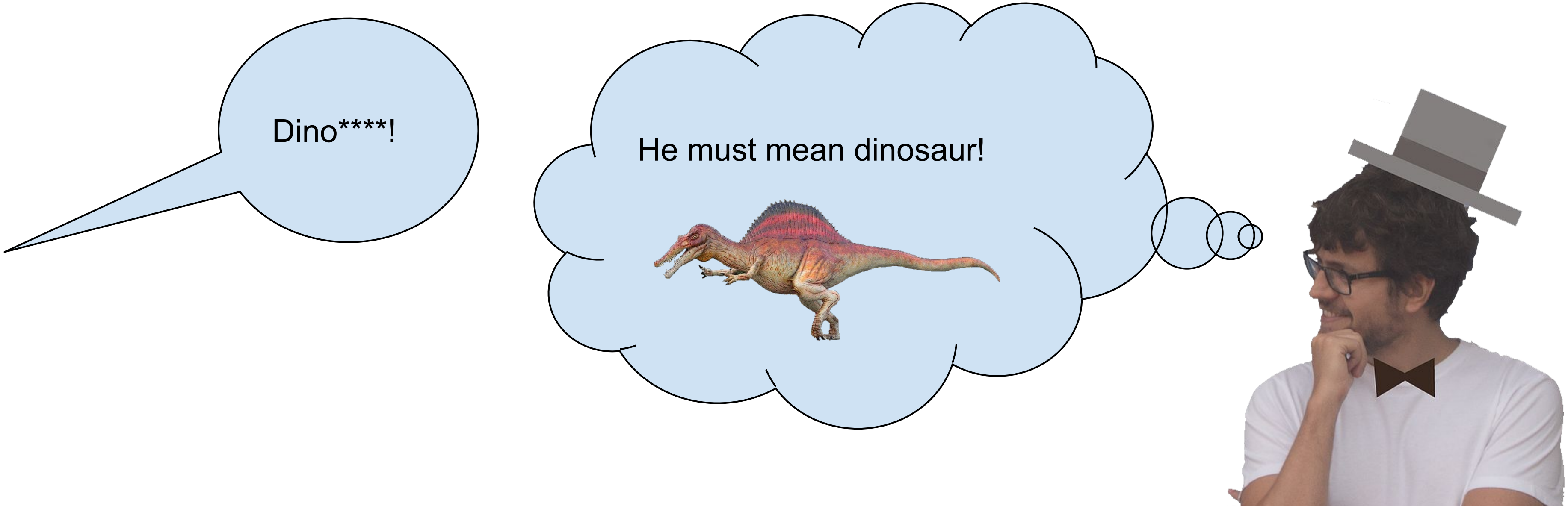
Google

Research Question

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

Introduction

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



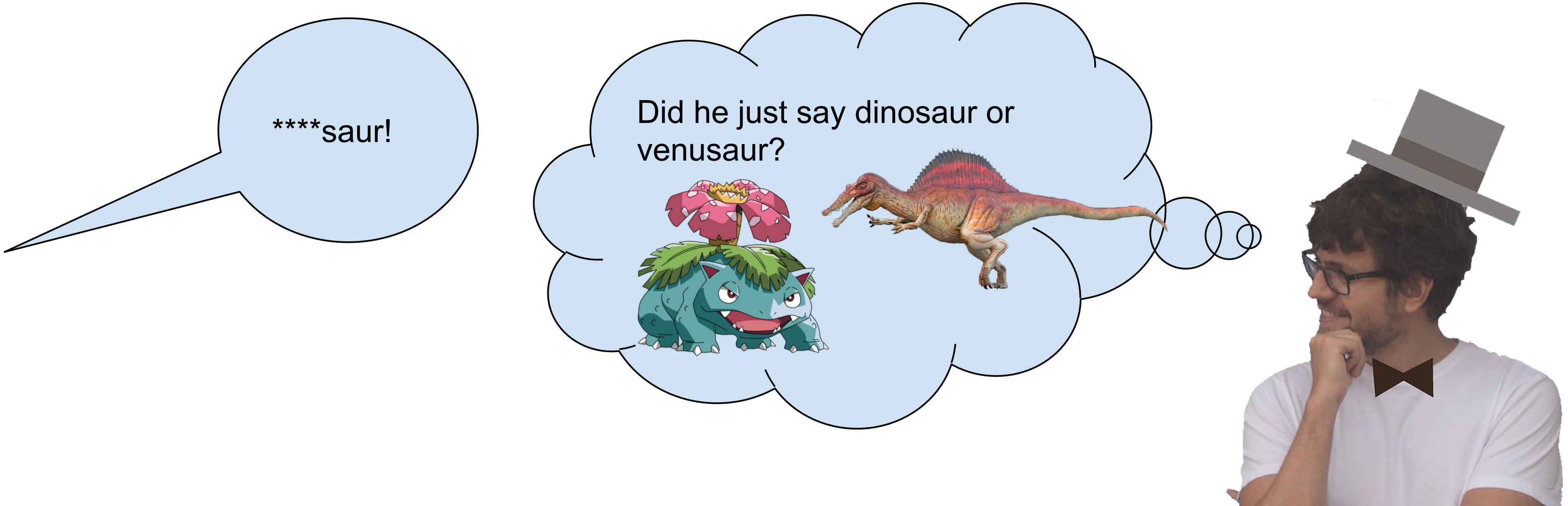
Introduction

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

Introduction

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

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



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 s[illegible]cts:

- listeners find word-initial consonant del[illegible] than word-final (Bagley, 1900)
- mispronunciations are m[illegible]y and Cutler, 1977)
- recognizin[illegible]itial characters is harder than [illegible]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:

$$\mathrm{H}(W_i \mid W_{<i}) = - \sum_{\mathbf{w} \in \Sigma^*} p(w_i \mid \mathbf{w}_{<i}) \log p(w_i \mid \mathbf{w}_{<i})$$

How can we measure this?

Information-theoretic measurements on natural corpora:

Character

- Estimate a segme… …ormation as its contextual entropy:

$$\mathrm{H}(W_i \mid W_{<i}) = - \sum_{\mathbf{w} \in \Sigma^*} p(w_i \mid \mathbf{w}_{<i}) \log p(w_i \mid \mathbf{w}_{<i})$$

How can we measure this?

Information-theoretic measurements on natural corpora:

- Estimate a segme[...]ormation [...]ntextual entropy:

Character

Previous characters

$$\mathrm{H}(W_i \mid W_{<i}) = -\sum_{\mathbf{w} \in \Sigma^*} p(w_i \mid \mathbf{w}_{<i}) \log p(w_i \mid \mathbf{w}_{<i})$$

How can we measure this?

Information-theoretic measurements on natural corpora:

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

$$\mathrm{H}(W_i \mid W_{<i}) = - \sum_{\mathbf{w} \in \Sigma^*} p(w_i \mid \mathbf{w}_{<i}) \log p(w_i \mid \mathbf{w}_{<i})$$

How can we measure this?

Information-theoretic measurements on natural corpora:

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

$$\mathrm{H}(W_{\boxed{i}} \mid \boxed{W_{<i}}) = - \sum_{\mathbf{w} \in \Sigma^*} p(w_i \mid \mathbf{w}_{<i}) \log p(w_i \mid \mathbf{w}_{<i})$$

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:

$$\mathrm{H}(W_{\boxed{i}} \mid \boxed{W_{<i}}) = -\sum_{\mathbf{w} \in \Sigma^*} p(w_i \mid \mathbf{w}_{<i}) \log p(w_i \mid \mathbf{w}_{<i})$$

This inherently confounds **context size** and **word position**.

Left-to-right Conditional Entropy

Conditioning information can only reduce entropy!

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

dino-saur!

After hearing dino-, saur
is not very informative...

Left-to-right Conditional Entropy

Conditioning information can only reduce entropy!

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

Left-to-right conditional entropies, thus:

- confound the amount of conditional information with word position.

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

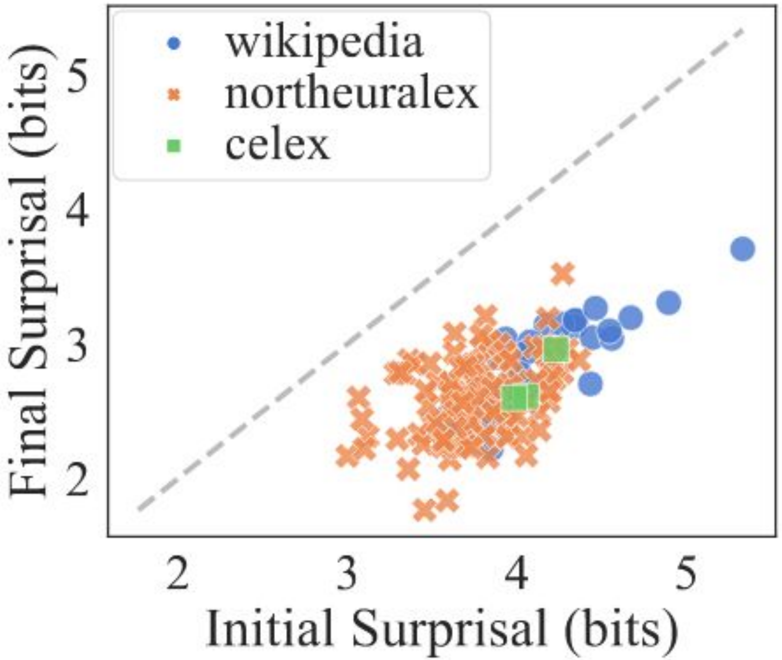
Significant word-initial

Significant word-final

Dataset	Languages	Forward	
CELEX	3	3	0
NorthEuraLex	107	106	0
Wikipedia	41	41	0

Results - Forward Surprisal

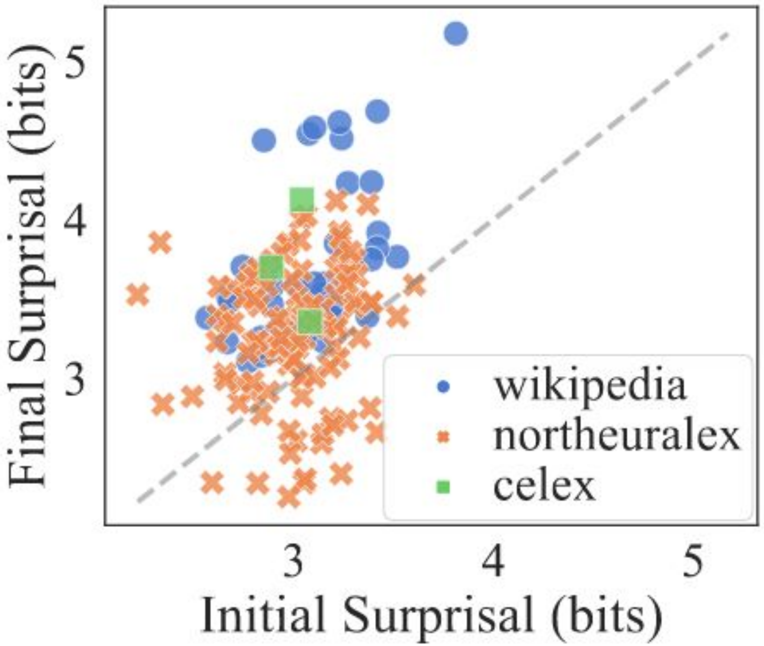
All but one language in the three analysed datasets had larger word-initial surprisal



Dataset	Languages	Forward	
CELEX	3	3	0
NorthEuraLex	107	106	0
Wikipedia	41	41	0

Results - Backward Surprisal

Many languages have higher word-final surprisals.



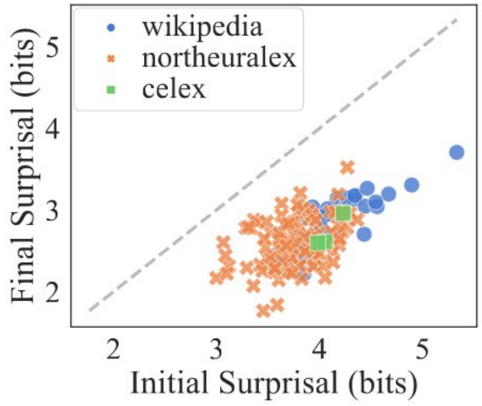
Dataset	Languages	Forward		Backward	
CELEX	3	3	0	0	3
NorthEuraLex	107	106	0	11	31
Wikipedia	41	41	0	0	39

Results

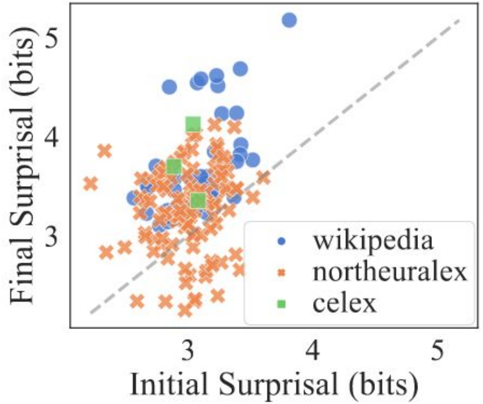
There seems to be both:

- a large effect of the amount of conditional information
- a lexical effect of front-loading disambiguatory signals

Forward Surprisal



Backward Surprisal



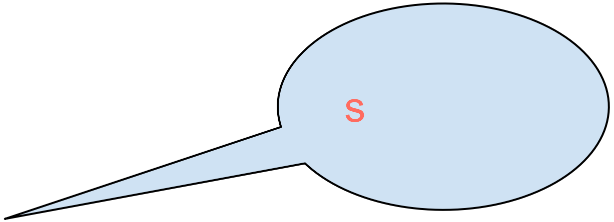
Controlling for Context Size

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:

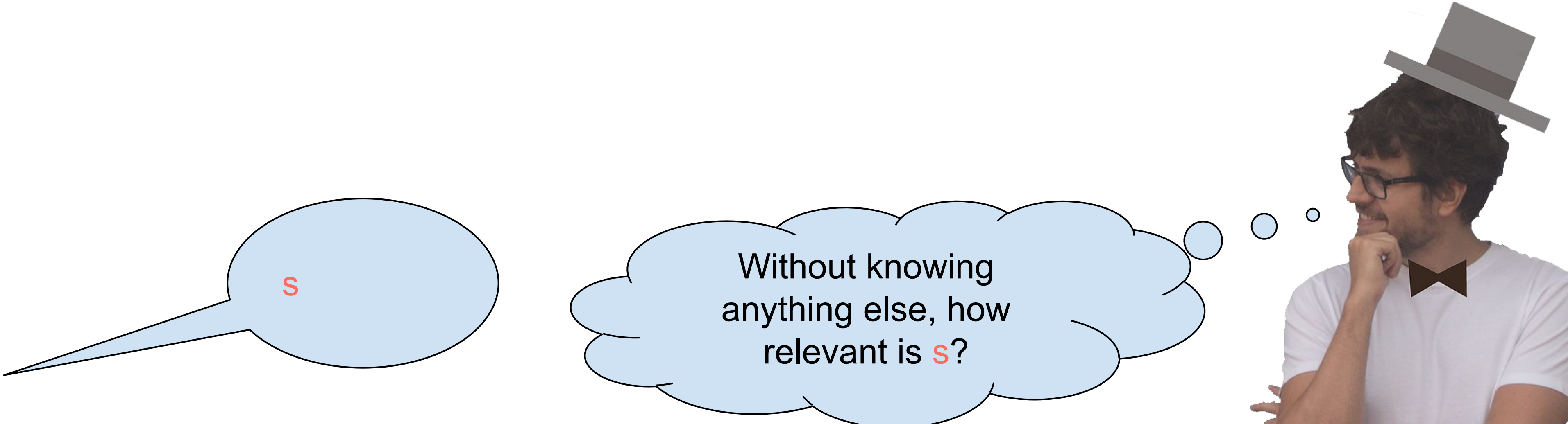
- Unigram surprisal: $\mathrm{H}(W_t)$



Controlling for Context Size

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

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



Controlling fo ntext Size

Full information a segment conveys

We propose the use of th o control for context size:

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

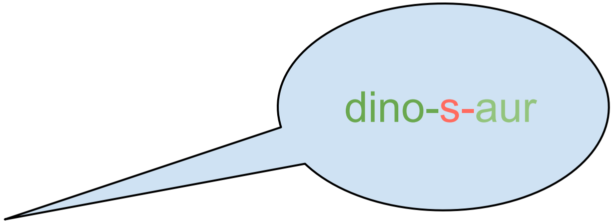
s

Without knowing anything else, how relevant is s?

Controlling for Context Size

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

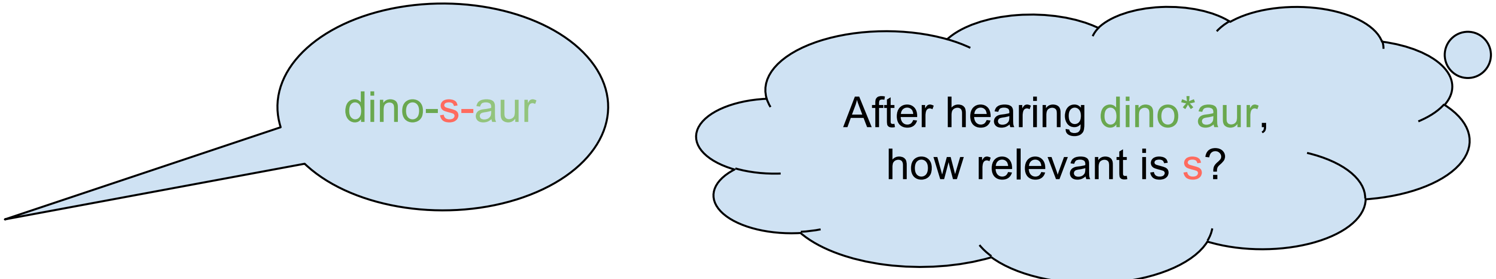
- Unigram surprisal: $\mathrm{H}(W_t)$
- Cloze Surprisal: $\mathrm{H}(W_t \mid W_{\neq t})$



Controlling for Context Size

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

- Unigram surprisal: $\mathrm{H}(W_t)$
- Cloze Surprisal: $\mathrm{H}(W_t \mid W_{\neq t})$



Controlling for Co… xt Size

We propose th… e measures to control for conte…

Non-redundant information a segment conveys

- Unigra… urprisal: …(W_t)
- Cloze Surprisal: $\mathrm{H}(W_t \mid W_{\neq t})$

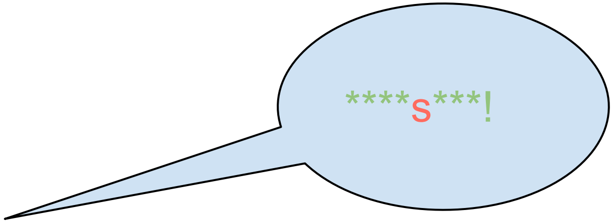
dino-s-aur

After hearing dino*aur,
how relevant is s?

Controlling for Context Size

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

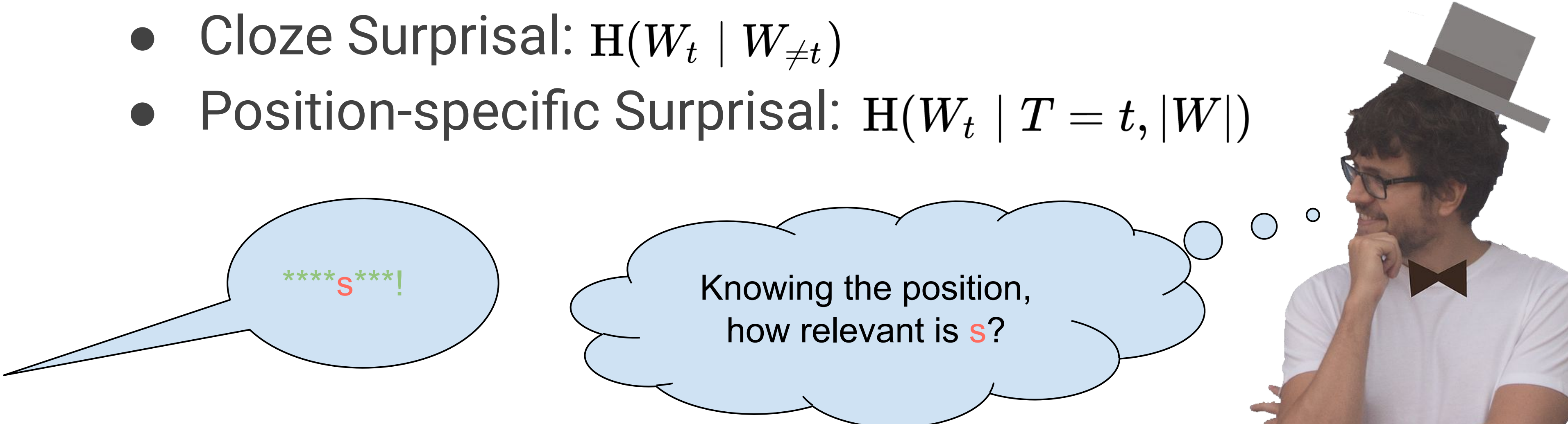
- Unigram surprisal: $\mathrm{H}(W_t)$
- Cloze Surprisal: $\mathrm{H}(W_t \mid W_{\neq t})$
- Position-specific Surprisal: $\mathrm{H}(W_t \mid T = t, |W|)$



Controlling for Context Size

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

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

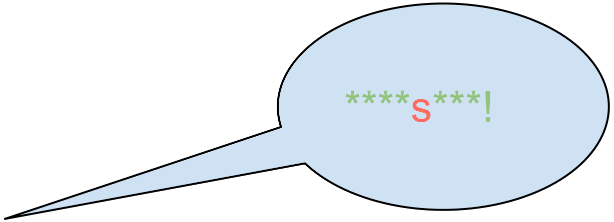


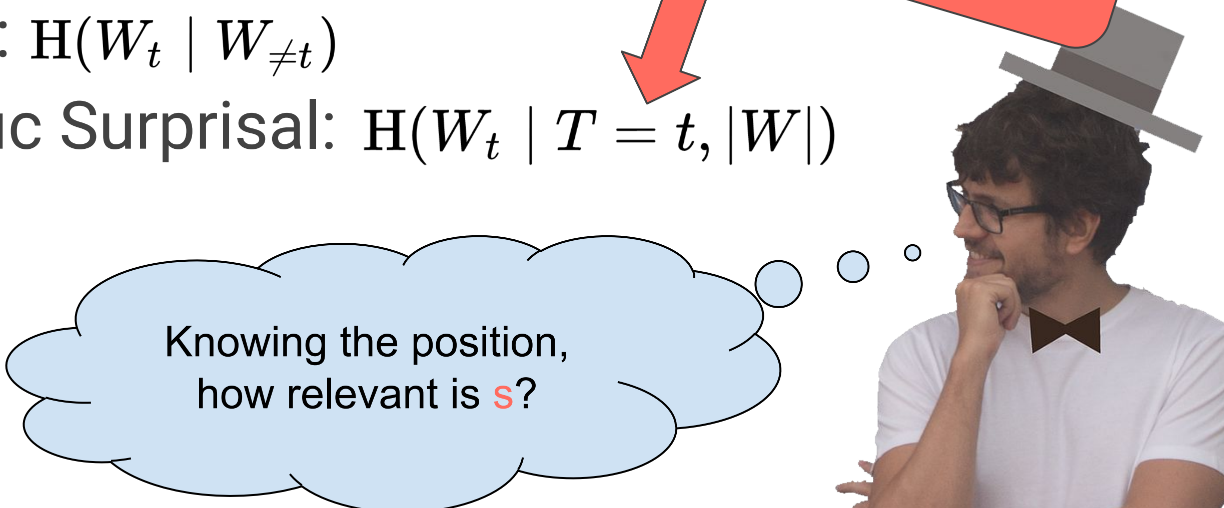
Controlling for Context Size

We propose the use of three mea… …o control for context size:

Information when primed by position and length

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





Controlling for Context Size

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

- Unigram surprisal: $\mathrm{H}(W_{...})$
- Cloze Surprisal...
- Position-spe... $(W_t \mid T = t, |W|)$

Informatio... by p... primed ...ngth

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

****s***!

Knowing the position, how relevant is s?

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)

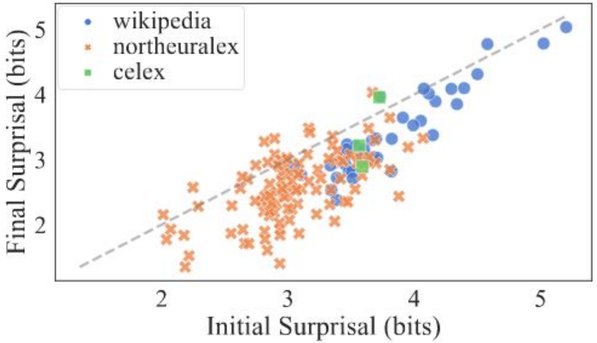
Results

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

Dataset	Languages	Forward		Backward		Unigram		Position-Specific		Cloze	
CELEX	3	3	0	0	3	2	0	2	1	2	1
NorthEuraLex	107	106	0	11	31	71	1	24	4	45	1
Wikipedia	41	41	0	0	39	39	1	31	1	35	2

Results

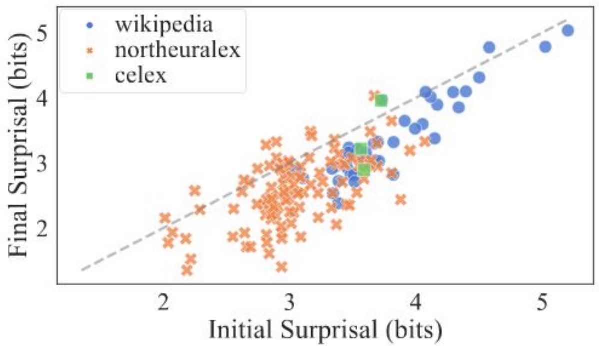


From the controlled metrics we see:

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

Dataset	Languages	Forward		Backward		Unigram		Position-Specific		Cloze	
CELEX	3	3	0	0	3	2	0	2	1	2	1
NorthEuraLex	107	106	0	11	31	71	1	24	4	45	1
Wikipedia	41	41	0	0	39	39	1	31	1	35	2

Results

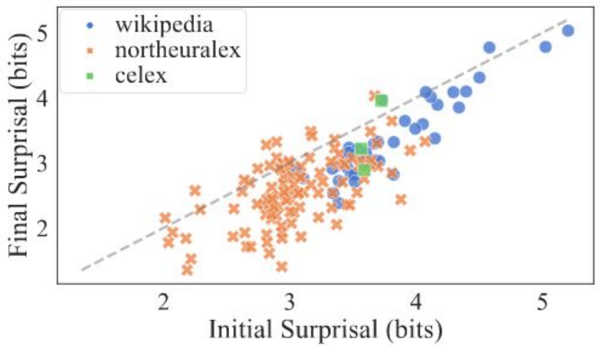


From the controlled metrics we see:

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

Dataset	Languages	Forward		Backward		Unigram		Position-Specific		Cloze	
CELEX	3	3	0	0	3	2	0	2	1	2	1
NorthEuraLex	107	106	0	11	31	71	1	24	4	45	1
Wikipedia	41	41	0	0	39	39	1	31	1	35	2

Results

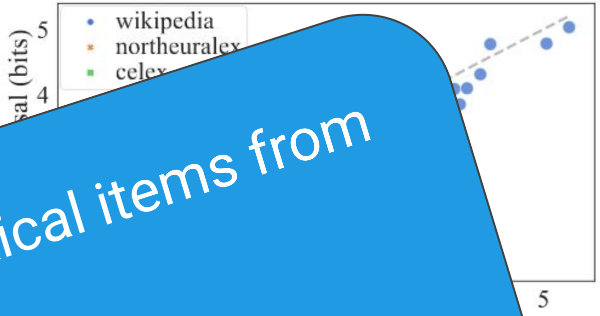


From the controlled metrics we see:

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

Dataset	Languages	Forward		Backward		Unigram		Position-Specific		Cloze	
CELEX	3	3	0	0	3	2	0	2	1	2	1
NorthEuraLex	107	106	0	11	31	71	1	24	4	45	1
Wikipedia	41	41	0	0	39	39	1	31	1	35	2

Results



From the controlled metri…

- a cross…

…ave

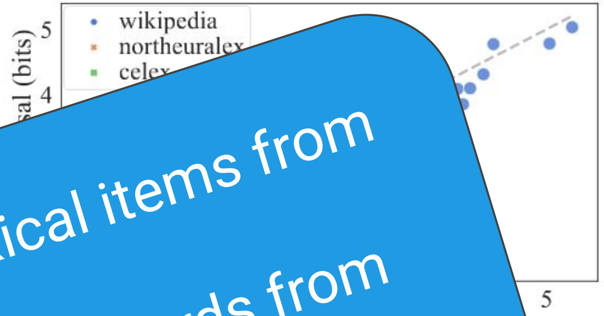
Prior work investigated the ease of recovering lexical items from its beginnings or endings:

Da…					Unigram		Position-Specific		Cloze	
CE…				3	2	0	2	1	2	1
Nort…		0	11	31	71	1	24	4	45	1
Wikip…	41	0	0	39	39	1	31	1	35	2

Results

From the controlled metri

- a cross



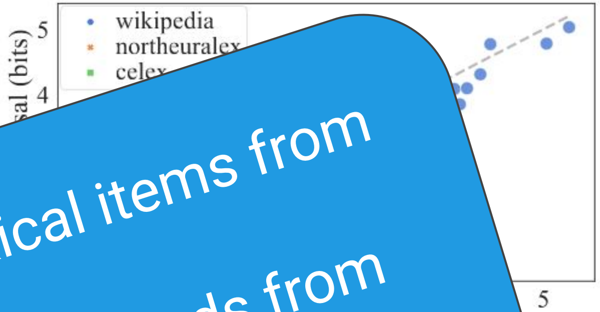
Prior work investigated the ease of recovering lexical items from its beginnings or endings:

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

ave

Da					Unigram		Position-Specific		Cloze	
CE				3	2	0	2	1	2	1
Nort		0	11	31	71	1	24	4	45	1
Wikip	41	0	0	39	39	1	31	1	35	2

Results



From the controlled metri

● a cross-li

ave

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

Da					Unigram		Position-Specific		Cloze	
CE				3	2	0	2	1	2	1
Nort		0	11	31	71	1	24	4	45	1
Wikip	41	0	0	39	39	1	31	1	35	2

An Informative Exploration of the Lexicon

Tiago Pimentel

Part 3

UNIVERSITY OF CAMBRIDGE

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

Introduction

Saussure claimed the association between wordforms and me is arbitrary.

- The arb
- Ex *cachorro* in Po
- **Sig** e arbitrarily attached to **signifieds**

Not the whole story!

Cachorro

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.

ACL 2019

Meaning to Form: Measuring Systematicity as Information

Tiago Pimentel, Arya D. McCarthy, Damián Blasi, Brian Roark, Ryan Cotterell

UNIVERSITY OF CAMBRIDGE
JOHNS HOPKINS WHITING SCHOOL of ENGINEERING
HARVARD UNIVERSITY
Google
ETH zürich

Research Question

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)

What does meaning tell us about forms?

Overall form uncertainty

Form uncertainty given meaning

define systematicity as mutual information:

MI(**meanings**; **forms**) = H(**forms**) - H(**forms** | **meanings**)

Our work

Advantages:

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

MI(**meanings**; **forms** | **POS**) = H(**forms** | **POS**) - H(**forms** | **meanings**, **POS**)

But, how can we measure H(**forms**) and H(**forms** | **meanings**)?

Our work

We use two LSTMs to get the language's entropy

1. H(**forms**):

- Predict phone given previous ones;
- $p_\theta(\textbf{form}) = \prod p_\theta(w_t \mid w_{t-1})$
- $H(\textbf{forms}) \leq H_\theta(\textbf{forms}) \approx - \sum \log p_\theta(\textbf{form}) / N$

Our work

We use two LSTMs to get the language's entropy

2. H(**forms** | **meanings**)

- Condition LSTM on meaning (word2vec embedding);
- $p_\theta(\textbf{form} \mid \textbf{meaning}) = \prod p_\theta(\textbf{w}_t \mid \textbf{w}_{t-1}, \textbf{m})$
- $H_\theta(\textbf{forms} \mid \textbf{meanings}) \approx - \sum \log p_\theta(\textbf{form} \mid \textbf{meaning}) \,/\, N$

Our work

We now estimate the MI with the cross-entropies:

$$\mathrm{MI}(\textbf{meanings}; \textbf{forms}) \approx H_\theta(\textbf{forms}) - H_\theta(\textbf{forms} \mid \textbf{meanings})$$

We also compute the uncertainty coefficient:

$$\mathrm{Unc}(\textbf{forms} \mid \textbf{meanings}) = \mathrm{MI}(\textbf{meanings}; \textbf{forms}) \,/\, H(\textbf{forms})$$

Results - CELEX

Used only monomorphemic words.

Results:

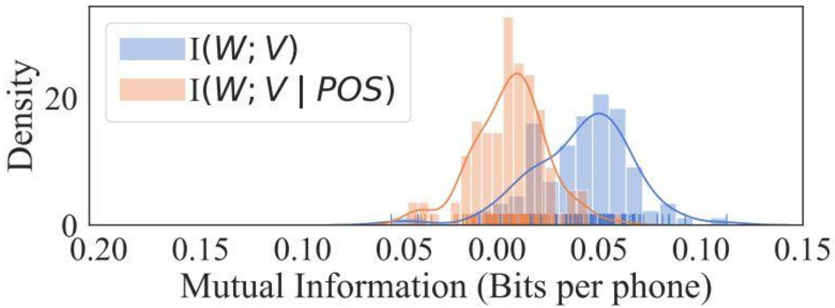
- Statistically significant systematicity in all three languages.
- Systematicity effect is reduced when we condition on POS.

		Systematicity		Systematicity given POS tags	
Language	H(forms)	MI	Unc	MI	Unc
English	3.401	0.11	3.24%	0.084	2.50%
German	3.195	0.168	5.26%	0.154	4.84%
Dutch	3.245	0.156	4.82%	0.089	2.84%

Results - NorthEuraLex

Lexicon consists of “basic” concepts;

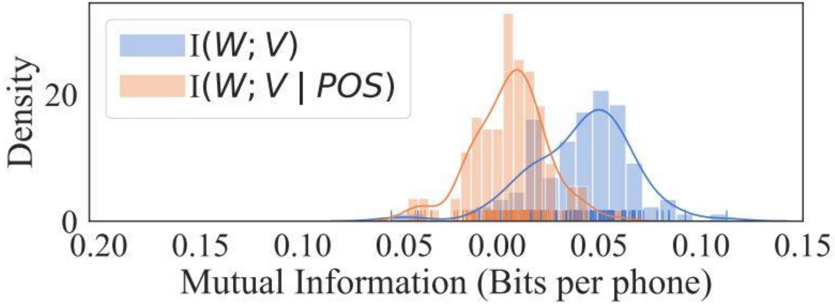
- We assume words are not multi-morphemic.



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):

- 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 - Phonesthemes

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.

Language	Phonestheme	Examples
Dutch	/sx/-	schelp, schild, schot, shacht, schaar
	-/əl/	kegel, nevel, beitel, vleugel, zetel
	-/xt/	beicht, nacht, vocht, plicht, licht
	-/ɔp/	stop, shop, drop, top, bob

Language	Phonestheme	Examples
English	/ɪn/-	infidel, intellect, institute, enigma, interim
	/sl/-	slop, slough, sluice, slim, slush
	-/kt/	aspect, object, fact, viaduct, tact
	-/mə/	panorama, asthma, trachoma, eczema, magma

Language	Phonestheme	Examples
German	/gə/-	geschehen, Gebiet, gering, Geruecht, gesinnt
	-/əln/	rascheln, rumpeln, tummeln, torkeln, mogeln
	-/ln/	rascheln, rumpeln, tummeln, torkeln, mogeln
	-/ən/	goennen, saeen, besuchen, giessen, streiten

NAACL 2021

Finding Concept-specific Biases in Form–Meaning Associations

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

UNIVERSITY OF CAMBRIDGE
Google
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
ETH zürich
HARVARD UNIVERSITY

Research Question

Are there **cross-linguistic** associations between the **forms** and **meanings** of words?

And how do we find them?

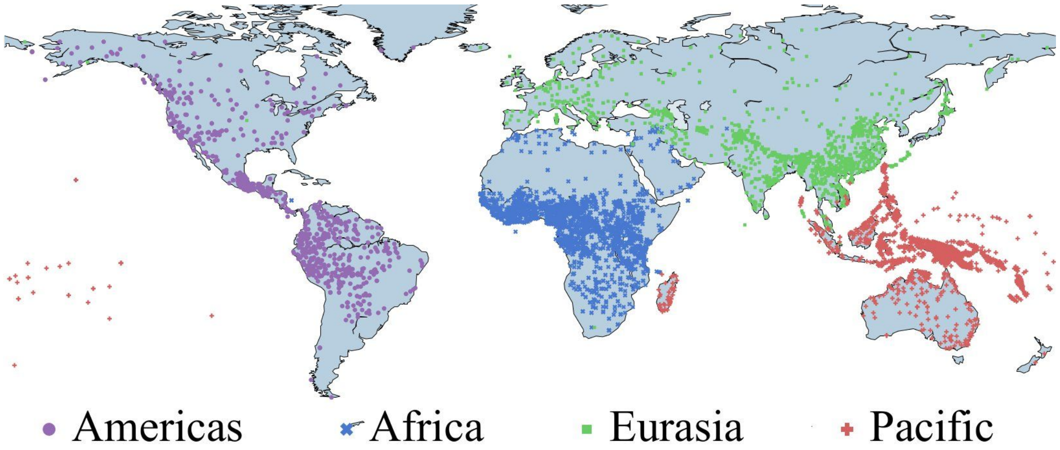
Example

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 ¾ 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 d... tion as bef...

MI(**mean**...

How do we...

Cross-entropy

Issue:
Cross-entropy needs to be computed on data independent from training!
Languages have been in contact (not i.i.d.)!

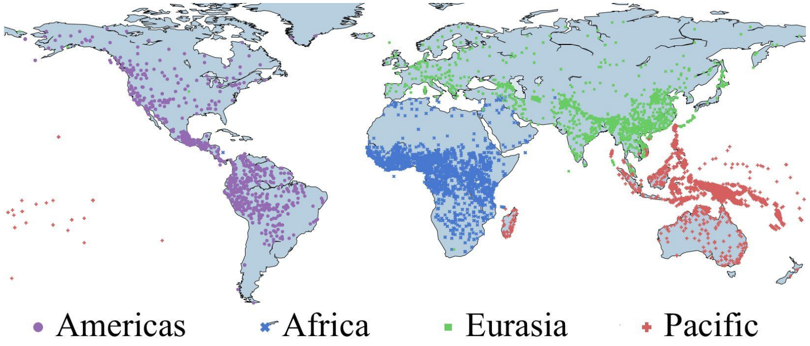
...g p_θ(**form**) / N

H(...) ≤ - ∑ log p_θ(**form** | **meaning**) / N

Method

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

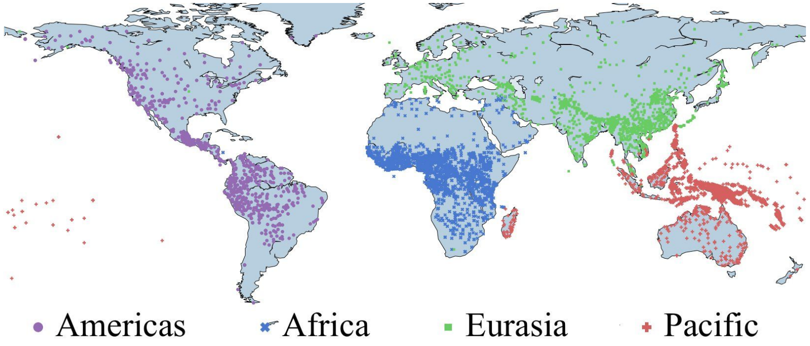
- 2 areas for training, 1 development, 1 test;
- Cross-validation.



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 contributi… meaning into form.

- Approxima… lan…

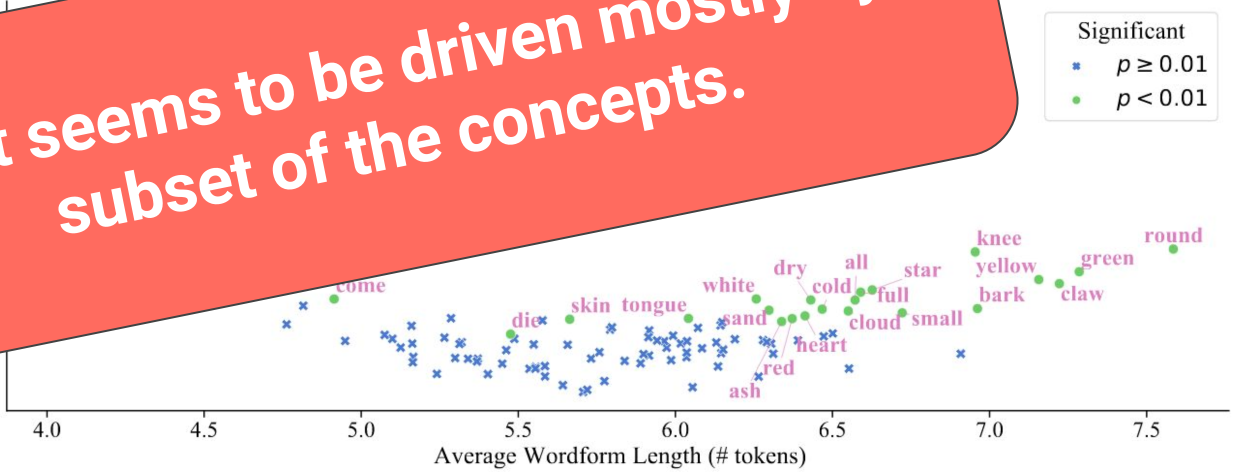
There are very small, but significant biases.
But we draw no strong conclusions at this level of the analysis.

Trai…			H(W)	MI(W;V)	U(W|V)
		…a	3.773	0.011	0.28%
		Americas	3.901	0.007	0.17%
A…	…ricas	Eurasia	3.999	0.015‡	0.38%
Ar…	Africa	Pacific	3.755	0.016‡	0.42%
Ave…			3.857	0.012‡	0.31%

Results - Per concept

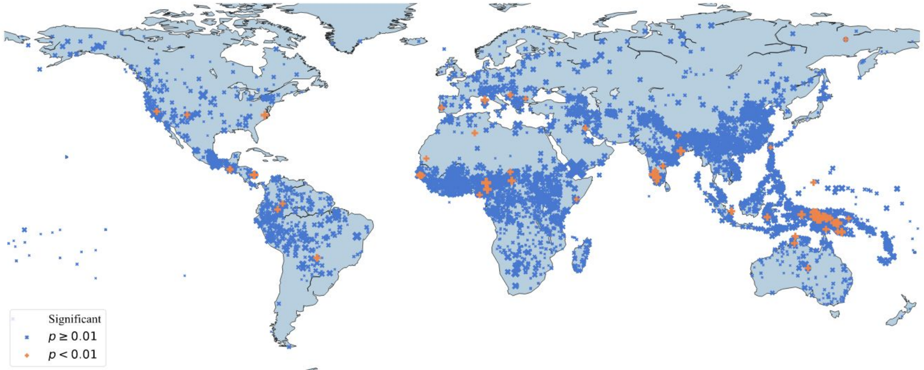
- Out of 100 concepts, 26 have significantly positive MI;
    - **Pronouns** present the highest values.
    - Most **colours** show statistically posi…
    - Some concepts rel… …ed to the

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 [#]

Concept	Tokens	Concept	Tokens	Concept	Tokens	Concept	Tokens	Concept	Tokens
blood	s	eye	i	liver	k l r t	path	d t	tree	#
bone	s u	fire	# t	louse	m n	say	#	two	r
breast	m u	fish	a s	mountain	b d g l o r	see	# e	water	#
come	# e	full	l o p t	name	# i	skin	k l p r t	we	# e i n
die	t	give	#	neck	o	star	k l o r s t u w	who	#
dog	k	horn	k r	new	a	stone	k t	you	# a i n
drink	# u	I	# a n	night	N d i l m p r t u	sun	e		
ear	e l t	knee	N b g k m o r t u	nose	N i u	tongue	d e l n r		
eat	#	leaf	a l p t	one	k t	tooth	e i		

An Informative Exploration of the Lexicon

Tiago Pimentel

Part 4

UNIVERSITY OF CAMBRIDGE

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

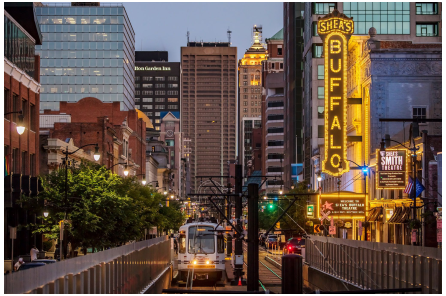
Tiago Pimentel, Rowan Hall Maudslay, Damián Blasi, Ryan Cotterell

UNIVERSITY OF CAMBRIDGE
HARVARD UNIVERSITY
ETH zürich

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

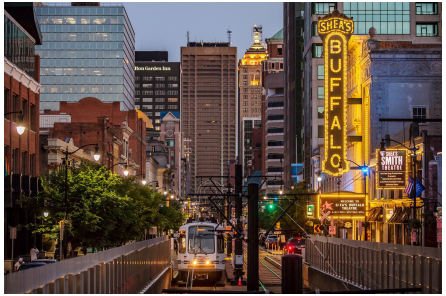
Words can mean more than one thing 😲

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 bisons intimidate other NY bisons



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 accomodate the listeners?

Image adapted from https://pngtree.com/so/woman

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 accomodate 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 accomodate the listeners?

→ **Put differently**: Is there a **negative** een
contextual uncertainty and l ?

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:

Meaning

Word

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

A Measure of Lexical Ambiguity

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

Meaning Word

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

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

$$I(M; W = w) = \underbrace{H(\mathbf{M})}_{\text{constant}} - H(M \mid W = w)$$

A Measure of Lexical Ambiguity

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

Meaning

Word

$$\mathbf{H(M \mid W{=}w)}$$

Information a word provides about meaning

Eq[illegible] up to an additive constant, to [illegible]al information (MI)

$$\mathbf{I(M;\, W = w)} = \underbrace{H(\mathbf{M})}_{\text{constant}} - \mathbf{H(M \mid W = w)}$$

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

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

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)

$$H(M \mid W=w) \approx H(N(\mu w, \Sigma w)) \quad . \\ = \tfrac{1}{2} \log_2 \det(2\pi e \Sigma w)$$

Validating our Measures of Lexical Ambiguity

How well do the BERT and WordNet measures of lexical ambiguity correlate ***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!

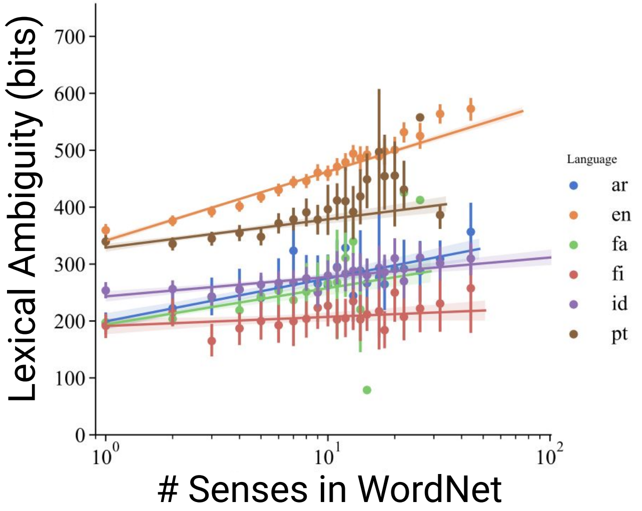
Language	# Types	Pearson	Spearman
Arabic	836	0.25**	0.30**
English	6995	0.40**	0.40**
Finnish	1247	0.06*	0.07*
Indonesian	3308	0.12**	0.13**
Persian	2648	0.14**	0.13**
Portuguese	3285	0.13**	0.13**

Validating our Measures of Lexical Ambiguity

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

→ Relatively well!

Language	# Types	Pearson	Spearman
Arabic	836	0.25**	0.30**
English	6995	0.40**	0.40**
Finnish	1247	0.06*	0.07*
Indonesian	3308	0.12**	0.13**
Persian	2648	0.14**	0.13**
Portuguese	3285	0.13**	0.13**

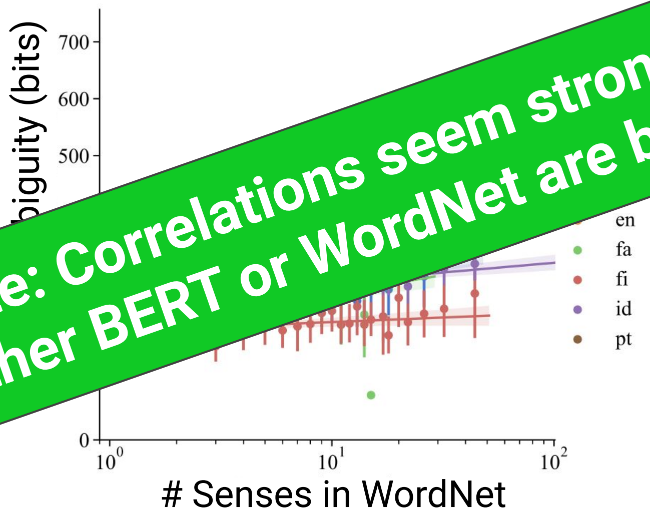


Validating our Measures of Lexical Ambiguity

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

→ Relatively well!

Language	# Types	Pearson	Spearman
Arabic	836	0.25**	0.30**
English	6995	0.40**	0.40**
Finnish	1247	0.06*	0.07
Indonesian	3308	0.12**	[illegible]
Persian	2648	0.14**	[illegible]
Portuguese	3285	0.13**	0.13**



Side note: Correlations seem stronger when either BERT or WordNet are better.

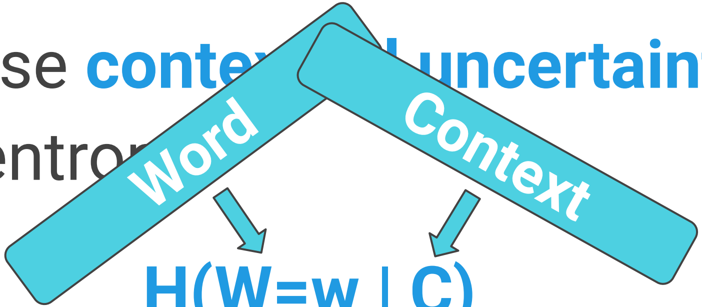
A Measure of Contextual Uncertainty

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

$$\mathbf{H(W{=}w \mid C)}$$

<u>A Measure</u> of Contextual Uncertainty

We operationalise **conte**[...]**l uncertainty** as the half-pointwise entro[...]

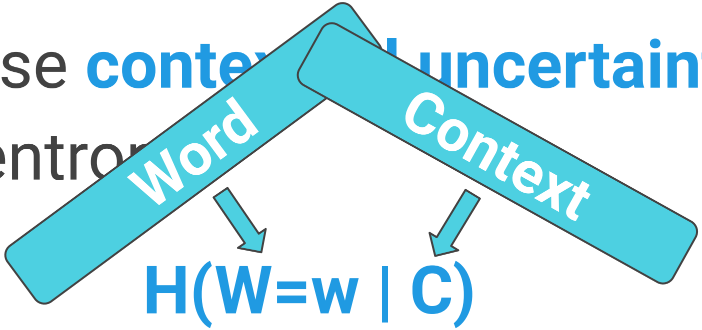


$$H(W{=}w \mid C)$$

Average uncertainty of a word in all its contexts

A Measure of Contextual Uncertainty

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



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!

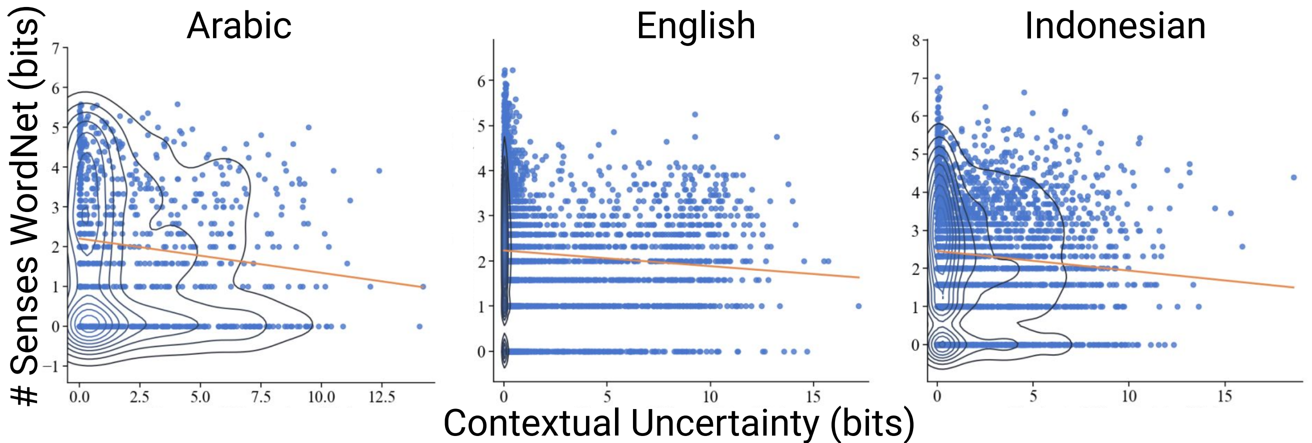
Language		# Types	Pearson	Spearman
Arabic	(ar)	836	-0.14**	-0.15**
English	(en)	6995	-0.07**	-0.11**
Finnish	(fi)	1247	0.01	0
Indonesian	(id)	3308	-0.09**	-0.14**
Persian	(fa)	2648	-0.11**	-0.12**
Portuguese	(pt)	3285	-0.10**	-0.11**

Empirical Evidence for the Trade-off

In WordNet

- Yes, in 5 of the 6 analysed languages!

Language		# Types	Pearson	Spearman
Arabic	(ar)	836	-0.14**	-0.15**
English	(en)	6995	-0.07**	-0.11**
Finnish	(fi)	1247	0.01	0
Indonesian	(id)	3308	-0.09**	-0.14**
Persian	(fa)	2648	-0.11**	-0.12**
Portuguese	(pt)	3285	-0.10**	-0.11**



Empirical Evidence for the Trade-off

In BERT

Empirical Evidence for the Trade-off

In BERT

- Yes, in all the 18 analysed languages!

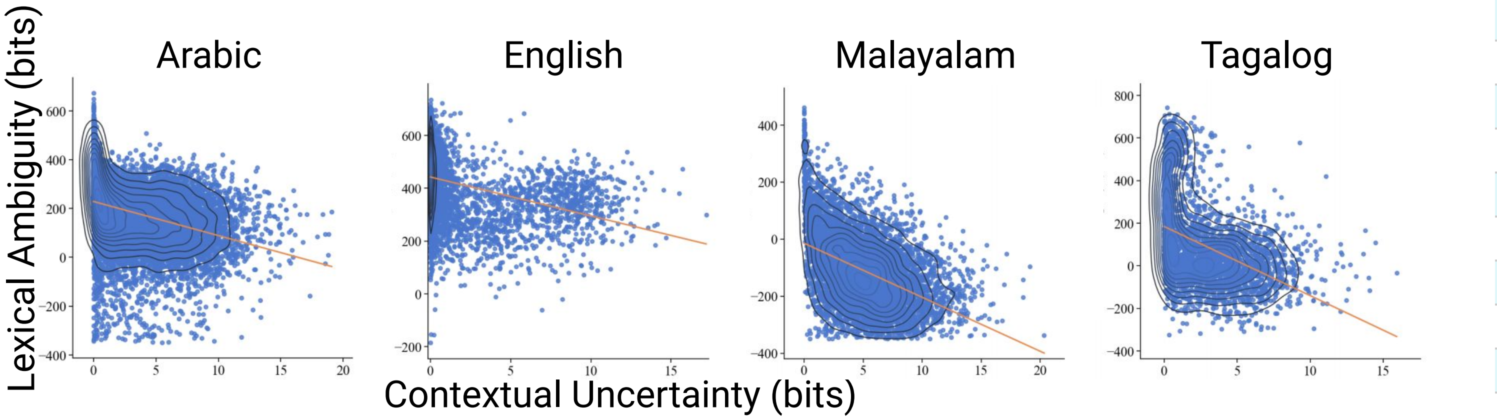
Language		# Types	Pearson	Spearman
Afrikaans	(af)	4505	-0.41**	-0.52**
Arabic	(ar)	10181	-0.33**	-0.41**
Bengali	(bn)	8128	-0.43**	-0.44**
English	(en)	7097	-0.33**	-0.35**
Estonian	(et)	4482	-0.40**	-0.44**
Finnish	(fi)	3928	-0.38**	-0.45**
Hebrew	(he)	13819	-0.34**	-0.37**
Indonesian	(id)	4524	-0.45**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

Empirical Evidence for the Trade-off

In BERT

- Yes, in all the 18 analysed languages!

Language		# Types	Pearson	Spearman
Afrikaans	(af)	4505	-0.41**	-0.52**
Arabic	(ar)	10181	-0.33**	-0.41**
Bengali	(bn)	8128	-0.43**	-0.44**
English	(en)	7097	-0.33**	-0.35**
Estonian	(et)	4482	-0.40**	-0.44**
Finnish	(fi)	3928	-0.38**	-0.45**
Hebrew	(he)	13819	-0.34**	-0.37**
			-0.45**	-0.57**
			-0.44**	-0.46**
			-0.42**	-0.41**
			-0.47**	-0.46**
			-0.39**	-0.40**
			-0.39**	-0.49**
			-0.31**	-0.45**
			-0.45**	-0.50**
			-0.40**	-0.46**
			-0.34**	-0.39**
Yoruba	(yo)	417	-0.55**	-0.64**

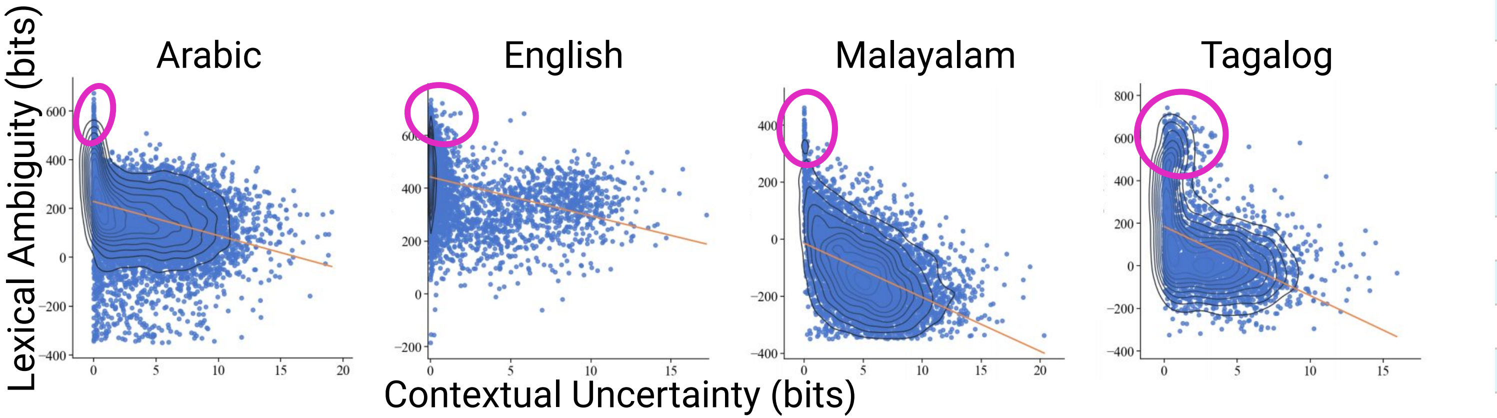


Empirical Evidence for the Trade-off

In BERT

- Yes, in all the 18 analysed languages!

Language		# Types	Pearson	Spearman
Afrikaans	(af)	4505	-0.41**	-0.52**
Arabic	(ar)	10181	-0.33**	-0.41**
Bengali	(bn)	8128	-0.43**	-0.44**
English	(en)	7097	-0.33**	-0.35**
Estonian	(et)	4482	-0.40**	-0.44**
Finnish	(fi)	3928	-0.38**	-0.45**
Hebrew	(he)	13819	-0.34**	-0.37**
			-0.45**	-0.57**
			-0.44**	-0.46**
			-0.42**	-0.41**
			-0.47**	-0.46**
			-0.39**	-0.40**
			-0.39**	-0.49**
			-0.31**	-0.45**
			-0.45**	-0.50**
			-0.40**	-0.46**
			-0.34**	-0.39**
Yoruba	(yo)	417	-0.55**	-0.64**



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

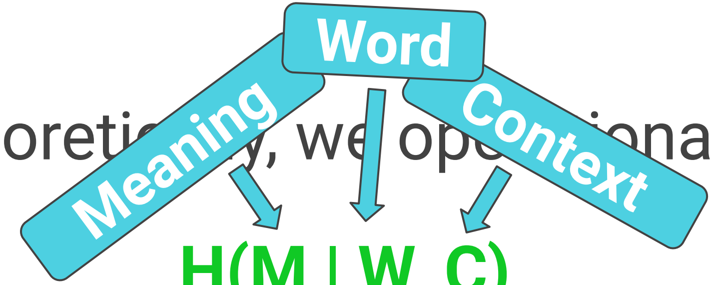
- Information-theoretically, we operationalise **clarity** as:

$$H(M \mid W, C)$$

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:

$$\mathbf{I(M; C; W = w)}$$

Robustness

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

- We operationalise **robustness** as a tripartite MI:

$$\mathbf{I(M; C; W = w)}$$

which is the **redundant** information shared by meaning, context and word

Why Should There Be a Trade-off?

- Assume language is **clear**, i.e. $\mathbf{H(M \mid W, C) = 0}$

Why Should There Be a Trade-off?

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

Why Should There Be a Trade-off?

- Assume language is **clear**, i.e. $\mathbf{H(M \mid W, C) = 0}$
- Assume language is **robust**, i.e. $\mathbf{I(M; C; W = w)} \geq \underbrace{k}_{\text{constant}}$

Why Should There Be a Trade-off?

- Assume language is **clear**, i.e. $\mathbf{H(M \mid W, C) = 0}$
- Assume language is **robust**, i.e. $\mathbf{I(M; C; W = w)} \geq \underbrace{k}_{\text{constant}}$

- We show:

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

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. $I(M; C; W = w) \geq \underbrace{k}_{\text{constant}}$

- We show:

$$I(M; C; W = w) = \underbrace{H(M)}_{\text{constant}} - \overbrace{H(M \mid W = w)}^{\text{Lexical ambiguity}} - \overbrace{H(W = w \mid C)}^{\text{Contextual uncertainty}}$$

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. $I(M; C; W = w) \geq \underbrace{k}_{\text{constant}}$

- We show:

$$I(M; C; W = w) = \underbrace{H(M)}_{\text{constant}} - \overbrace{H(M \mid W = w)}^{\text{Lexical ambiguity}} - \overbrace{H(W = w \mid C)}^{\text{Contextual uncertainty}}$$

- 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. $\mathbf{H(M \mid W, C) = 0}$
- Assume language is **robust**, i.e. $\mathbf{I(M; C; W = w)} \geq \underbrace{k}_{\text{constant}}$

- We show:

Lexical ambiguity

Contextual uncertainty

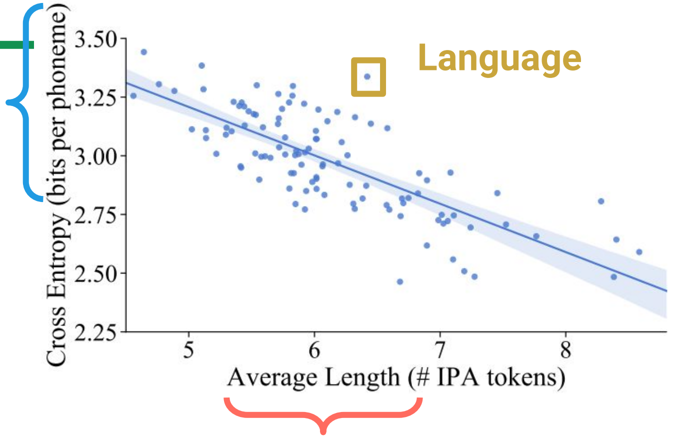
$$\mathbf{I(M; C; W = w)} = \underbrace{H(M)}_{\text{constant}} - \mathbf{H(M \mid W = w)} - \mathbf{H(W = w \mid C)}$$

There is a trade-off!!!

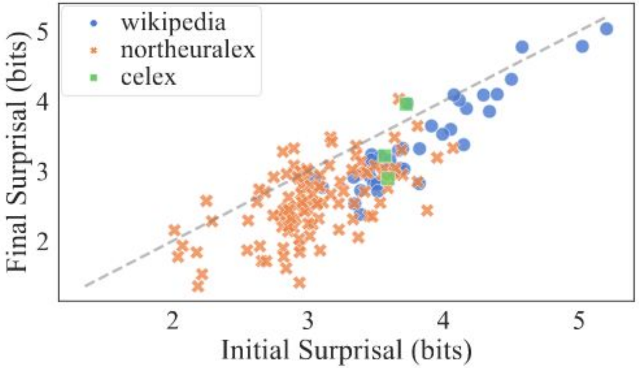
- Thus, $\mathbf{H(M \mid W = w)} + \mathbf{H(W = w \mid C)} \leq H(M) - k$

Takeaways

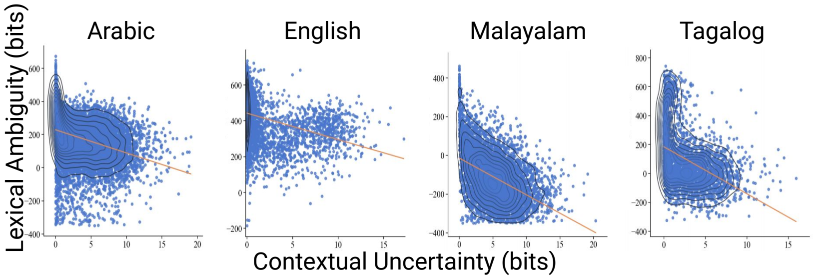
Spoiler



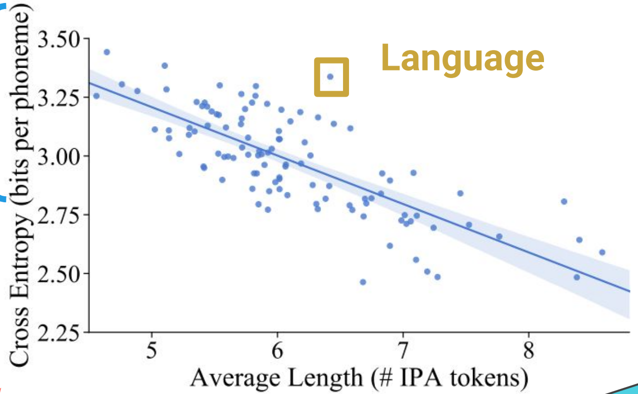
Language	H(W)	MI(W;V)	U(W|V)
English	3.401	0.11	3.24%
German	3.195	0.168	5.26%
Dutch	3.245	0.156	4.82%



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%

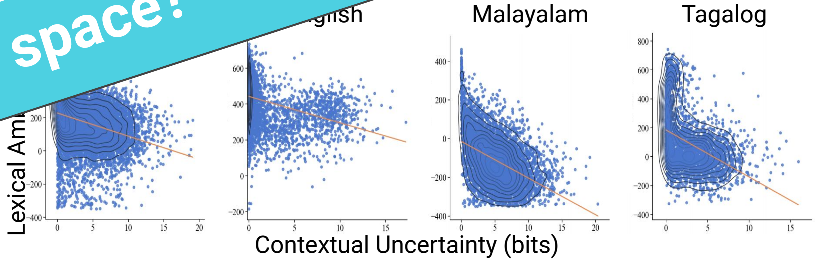


Spoiler



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%

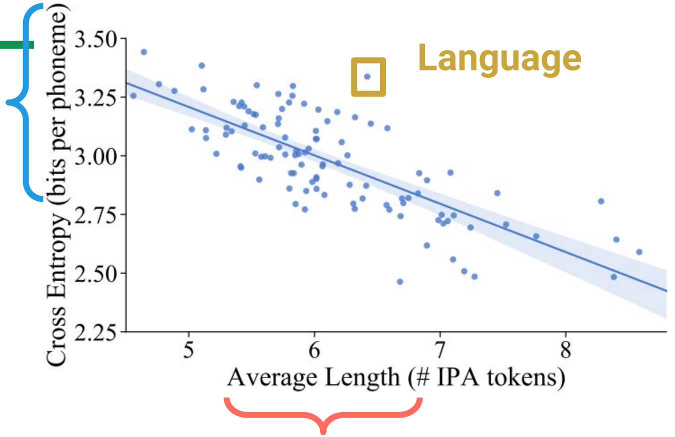
Test	H(W)	MI(W;V)
Africa	3.773	0.011
Americas	3.901	0.007
Eurasia	3.999	$0.015^{\ddagger}$
Pacific	3.755	$0.016^{\ddagger}$
Average	3.857	$0.012^{\ddagger}$

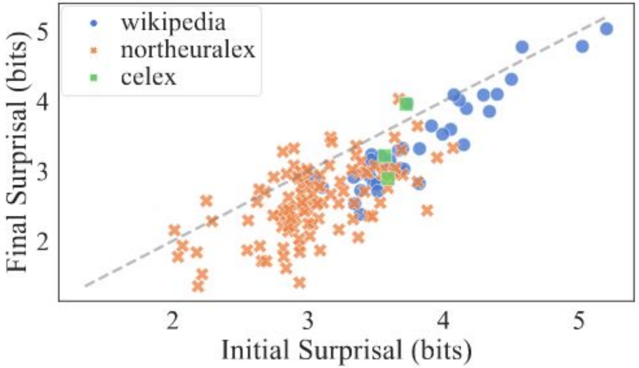


Phonotactic complexity correlates with word length

- **Pressure towards a roughly equal sized phonotactic space?**

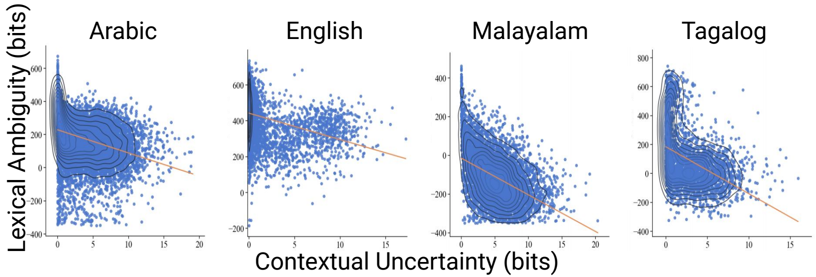
Spoiler



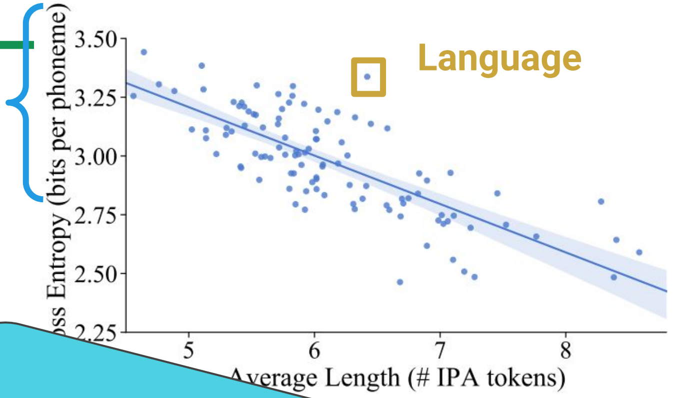


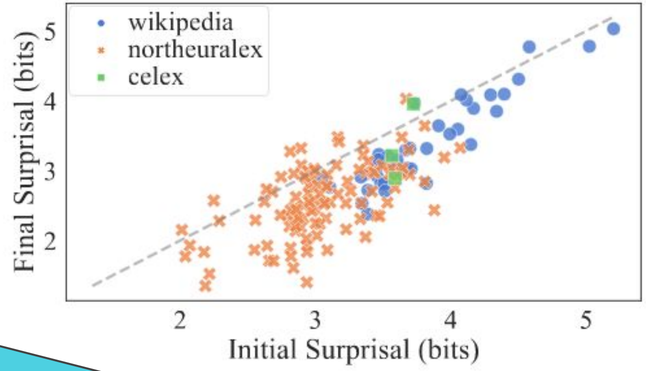
Language	H(W)	MI(W;V)	U(W|V)
English	3.401	0.11	3.24%
German	3.195	0.168	5.26%
Dutch	3.245	0.156	4.82%

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%



Spoiler





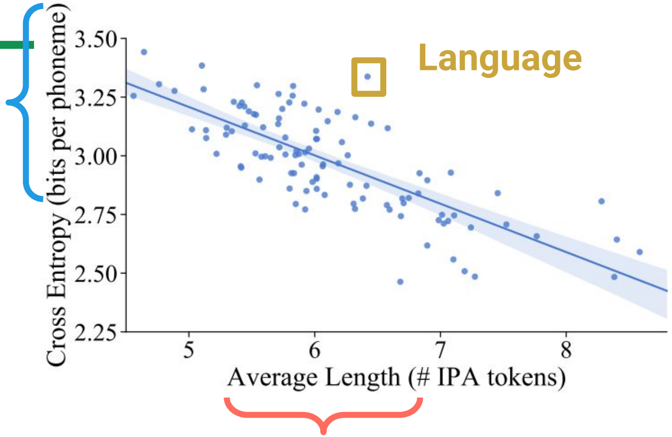
Language	H(W)	MI(
English	3.401	
German	3.195	
Dutch	3.245	

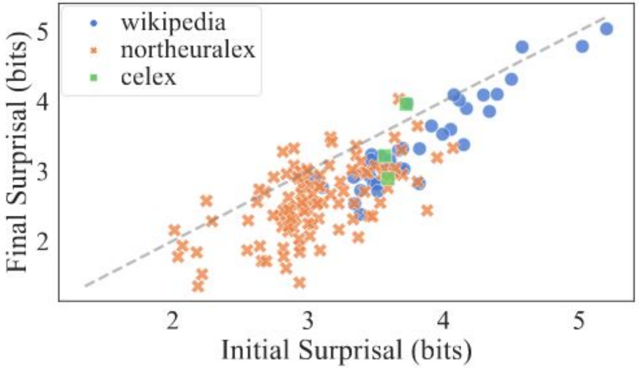
Test	H(W)	MI(	
Africa	3.773	0.011	
Americas	3.901	0.007	0.17%
Eurasia	3.999	0.015‡	0.38%
Pacific	3.755	0.016‡	0.42%
Average	3.857	0.012‡	0.31%

Cross-linguistic tendency to frontload information

- Pressure towards faster lexical access?

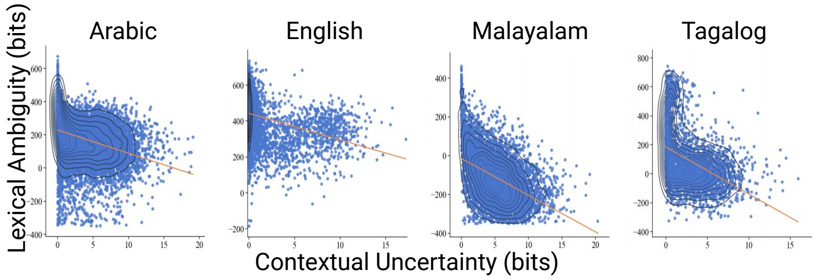
Spoiler



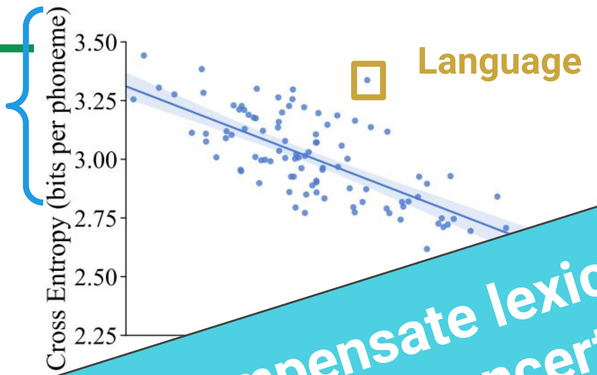


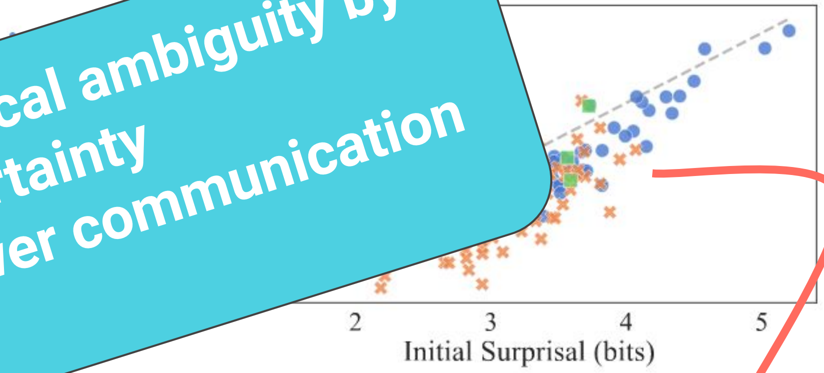
Language	H(W)	MI(W;V)	U(W|V)
English	3.401	0.11	3.24%
German	3.195	0.168	5.26%
Dutch	3.245	0.156	4.82%

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%



Spoiler



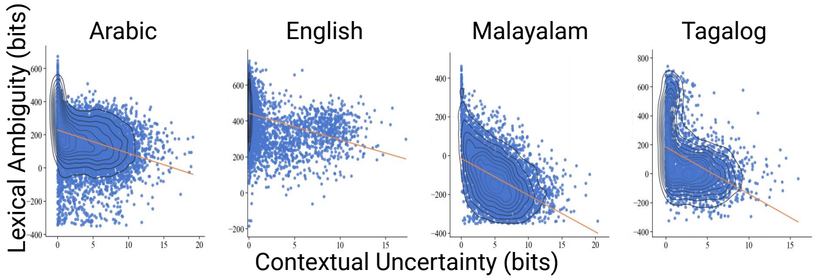


Speakers compensate lexical ambiguity by reducing contextual uncertainty

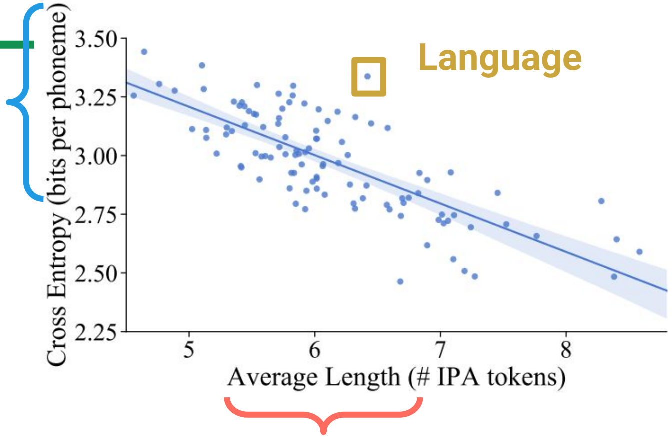
- **Pressure towards fewer communication errors?**

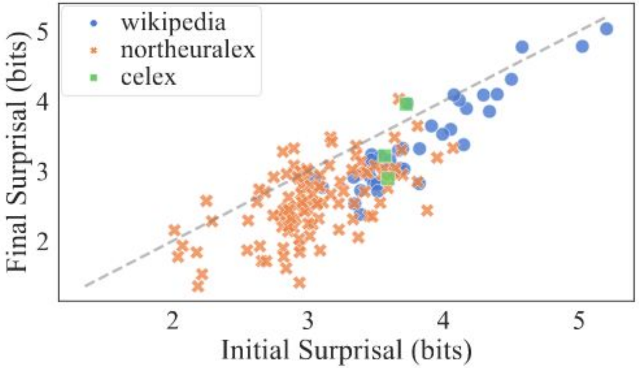
Language	H(W)
English	3.401
German	3.195
Dutch	3.245

Test	H(W)		(W|V)
Africa	3.773		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%



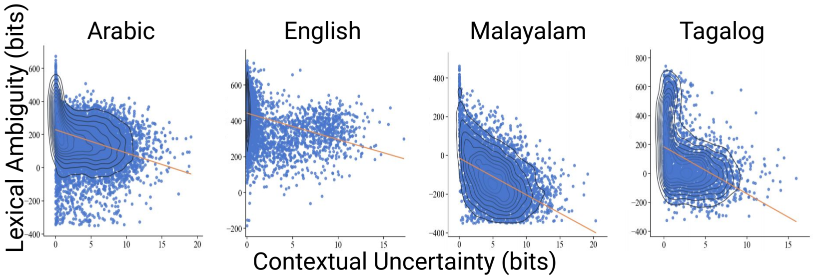
Spoiler



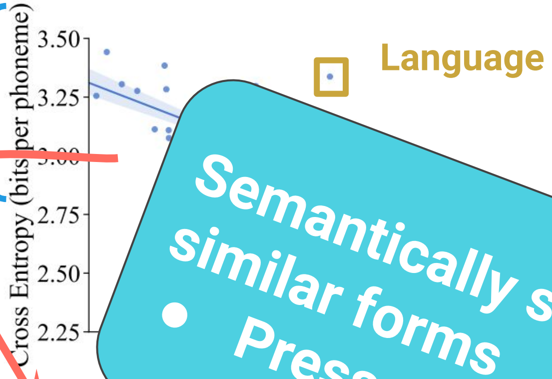


Language	H(W)	MI(W;V)	U(W|V)
English	3.401	0.11	3.24%
German	3.195	0.168	5.26%
Dutch	3.245	0.156	4.82%

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%

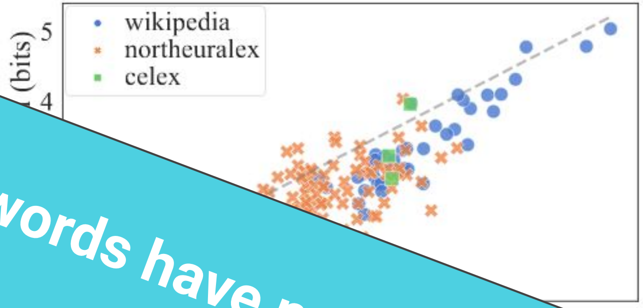


Spoiler



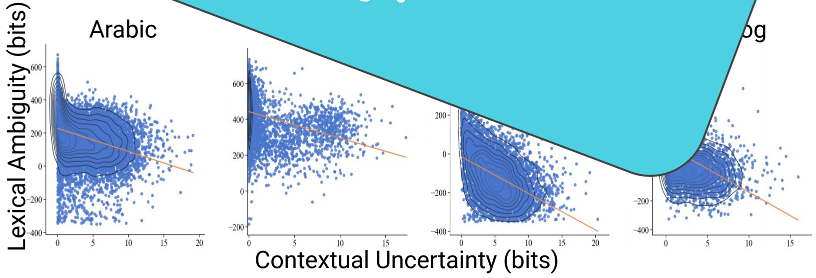
Semantically similar words have more similar forms

- Pressure towards learnability?



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%



Thanks :-)

And thanks to all co-authors:

Rowan Hall Maudslay, Brian Roark, Damián Blasi, Ryan Cotterell, Arya D. McCarthy, Søren Wichmann