On the Complexity and Typology of Inflectional Morphological Systems

Ryan Cotterell, Christo Kirov, Mans Hulden and Jason Eisner



JOHNS HOPKINS
UNIVERSITY

What’s *your* native language?

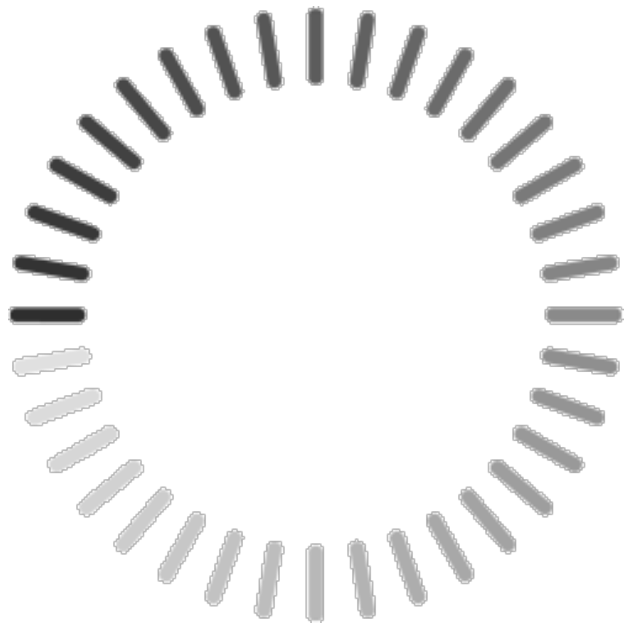
Johns Hopkins University

Do you think your native language more complex than English?

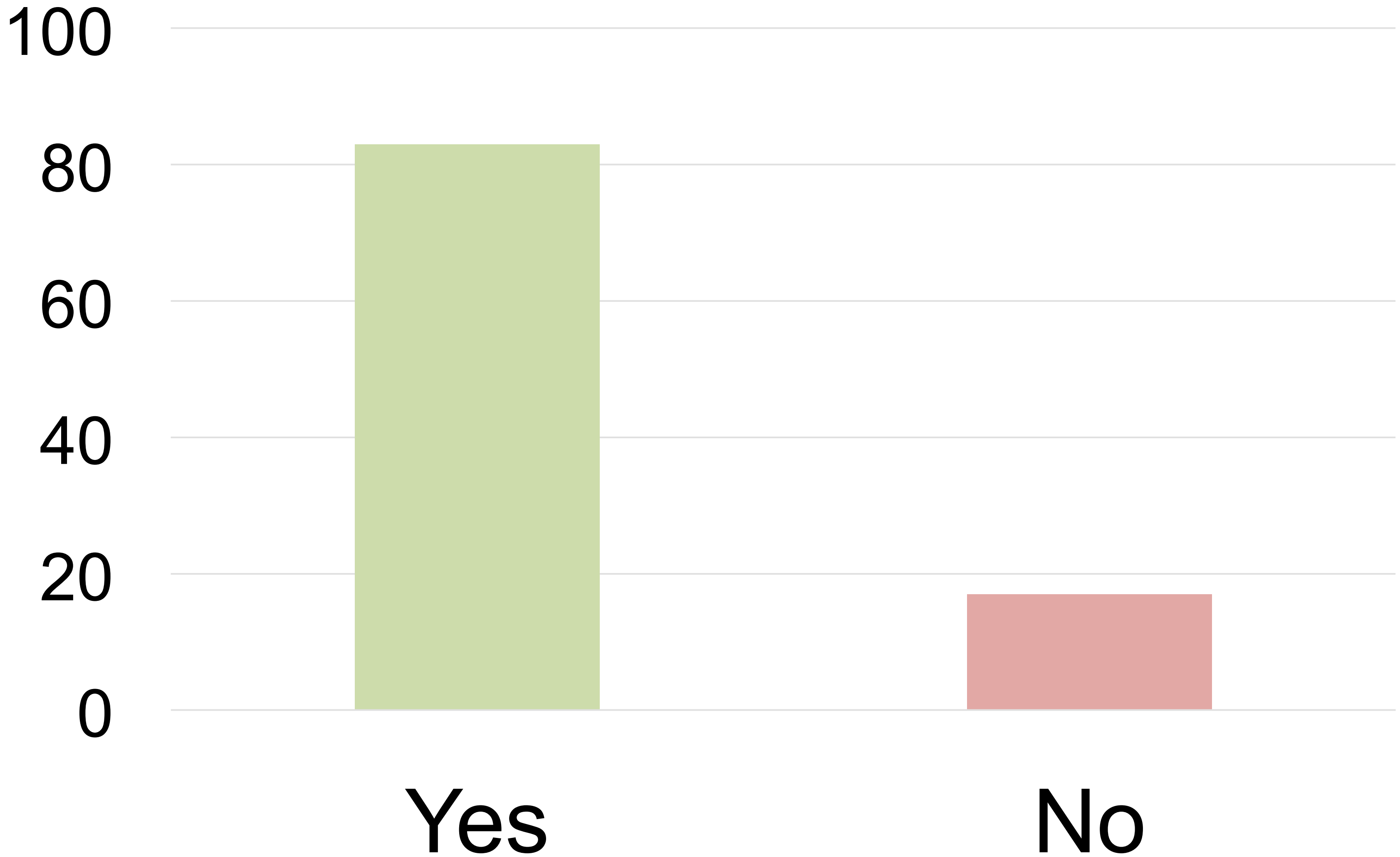
Audience Poll

Question: Who thinks their native language is more complex than English?

Loading…



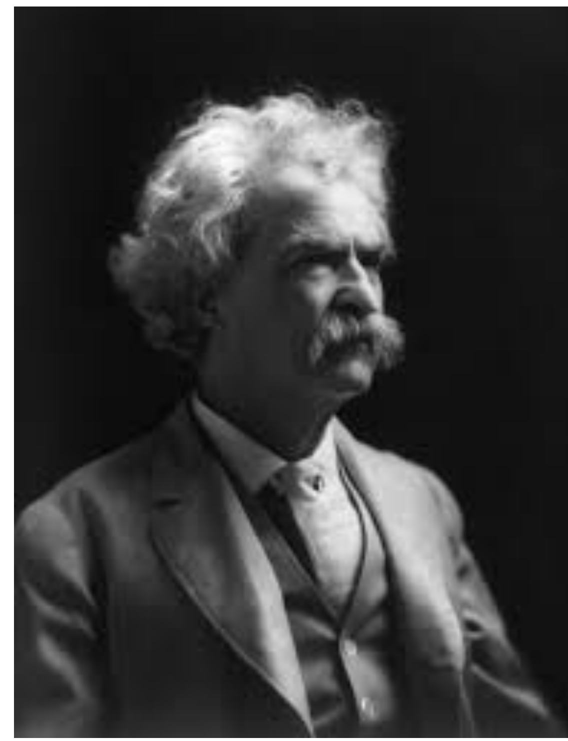
Instant Results



What makes your native language more complex than English?

More Morphological Variants = A More Complex Language

- I agree: a lot of morphological variants can make a language "difficult":
  - Mark Twain said it best: "I'd rather decline two drinks than one German adjective."



- Then, look at how simple English is!
  - Your average verb has four inflections: *run*, *runs*, *ran*, *running*
- Nouns and adjectives don't inflect in English according to case
  - The adjective *good* has one inflection

In Comparison: The Turkish Verb

- **koşmak** = Turkish verb ”to run”
  - (partial) paradigm →
- Tense, mood, evidentiality … marked through morphology
  - 100+ forms in Turkish!
- Archi (Kibrik 1998) has 1.5 million verb forms
  - It’s very, very agglutinative
- This makes a language more complex!

ben	koşıyorum
sen	koşıyorsun
o	koşıyor
biz	koşıyoruz
siz	koşıyorsunuz
onlar	koşıyorlar

ben	koşmadım
sen	koşmadın
o	koşmadı
biz	koşmadık
siz	koşmadınız
onlar	koşmadılar

JOHNS HOPKINS UNIVERSITY

Number of Forms is Only One Dimension of Morphological Complexity

- There are (at least) **two types** of morphological complexity
  - **Type 1:** how big are the paradigms? (seen before)
  - **Type 2:** how irregular are the paradigms?

- Ackerman and Malouf (2013) introduce the technical jargon
  - Enumerative Complexity (E-Complexity)
  - Integrative Complexity (I-Complexity)

English versus Turkish: # Forms

- **English**
  - 4 verbal slots
  - 2 nominal slots
  - 1 adjectival slot

7 Total

- **Turkish**
  - 350 verbal slots
  - 8 nominal slots
  - 1 adjectival slots

358 Total

Turkish is more morphologically complex under # forms per verb

English Versus Turkish: Irregularity

- **English**
  - 224 irregular verbs
  - 10 irregular nouns
  - 0 irregular adjectives

234 Total

- **Turkish**
  - 1 irregular verb
  - 0 irregular nouns
  - 0 irregular adjectives

1 Total

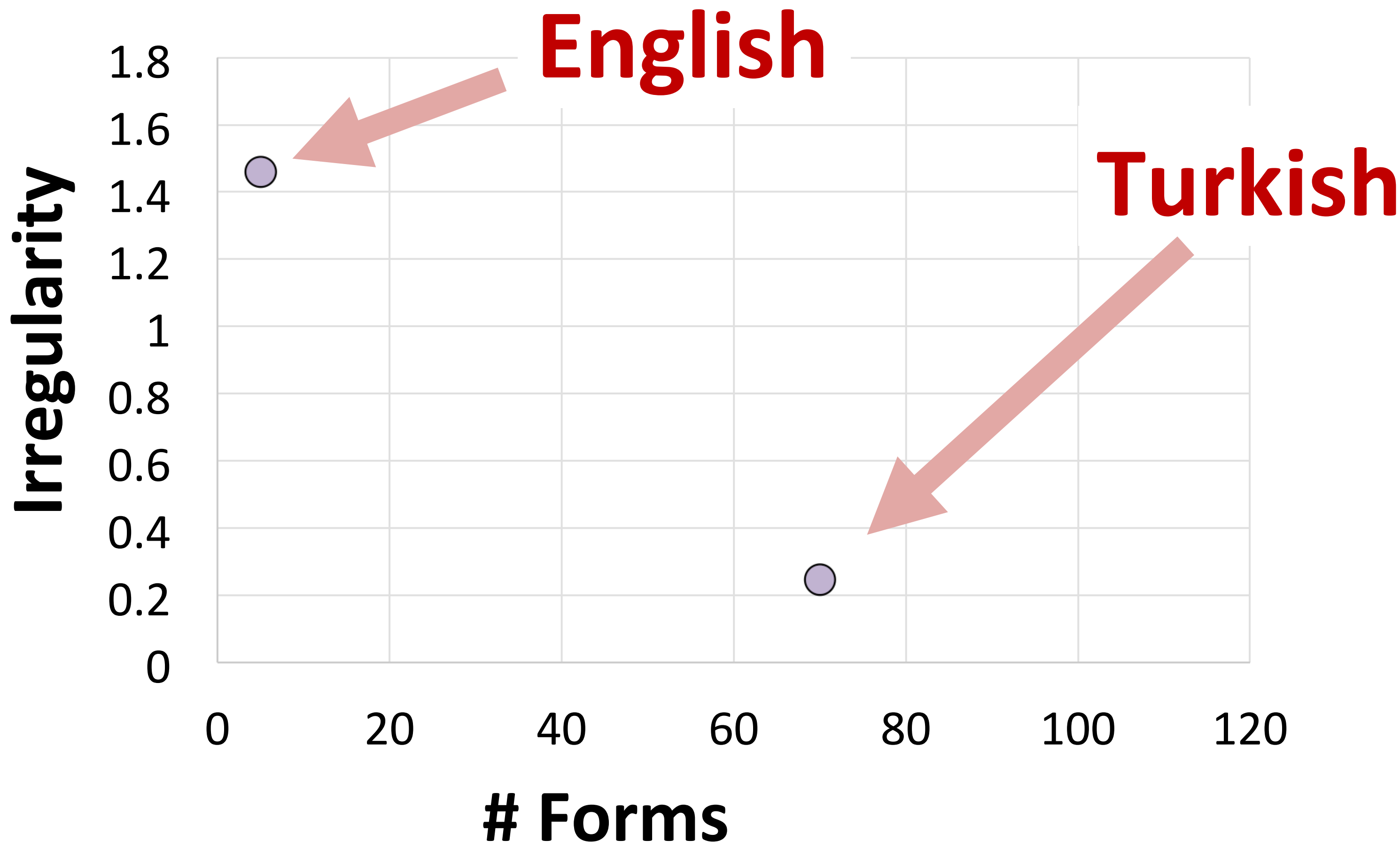
English is more morphologically complex under amount of irregularity

What's *This* Paper About?

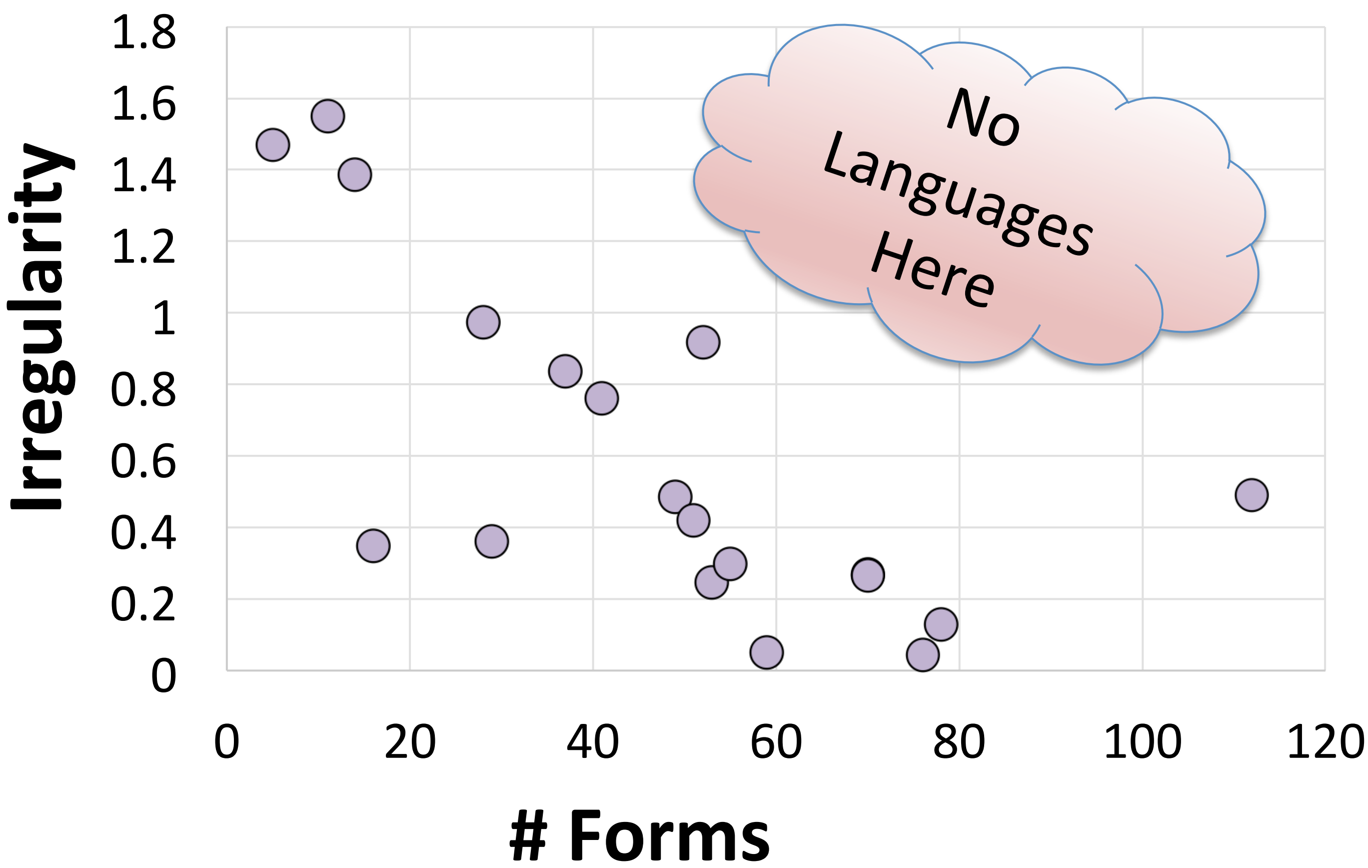
Good Linguistic Question:

How do the # morphological variants and morphological irregularity interact?

Plotting English and Turkish Verbs

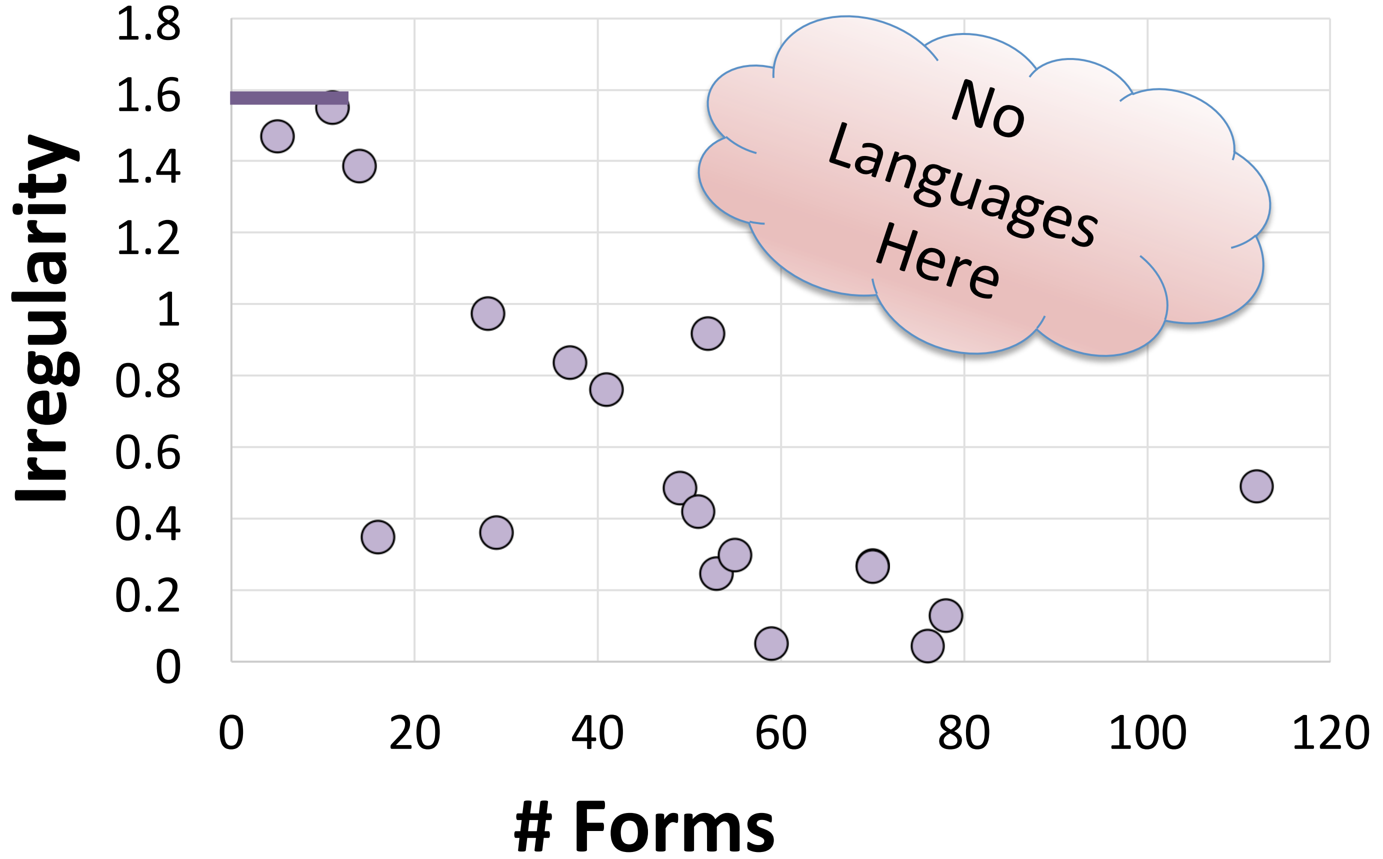


What about the Rest of the World's Languages? (For which we have data)

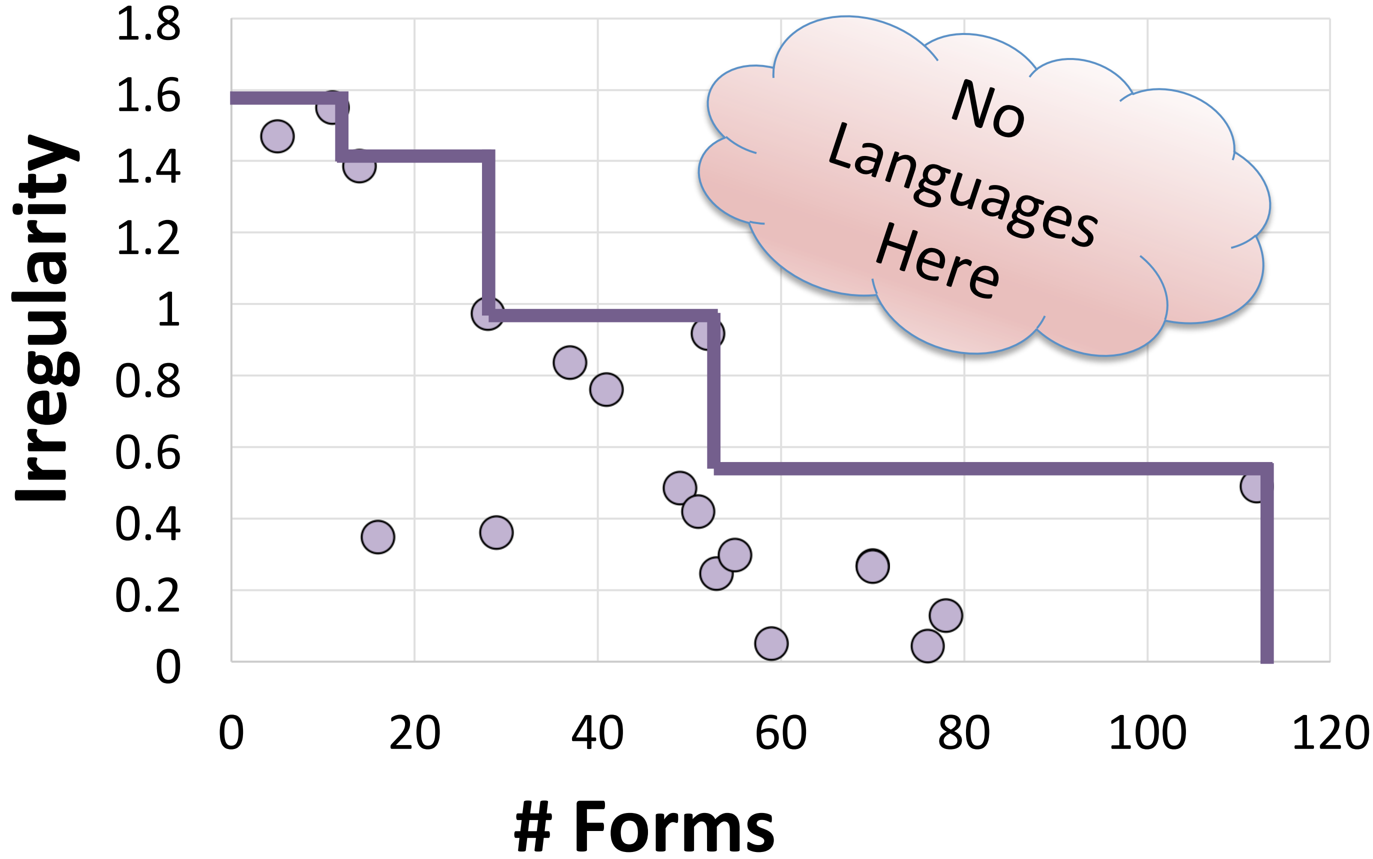


JOHNS HOPKINS UNIVERSITY

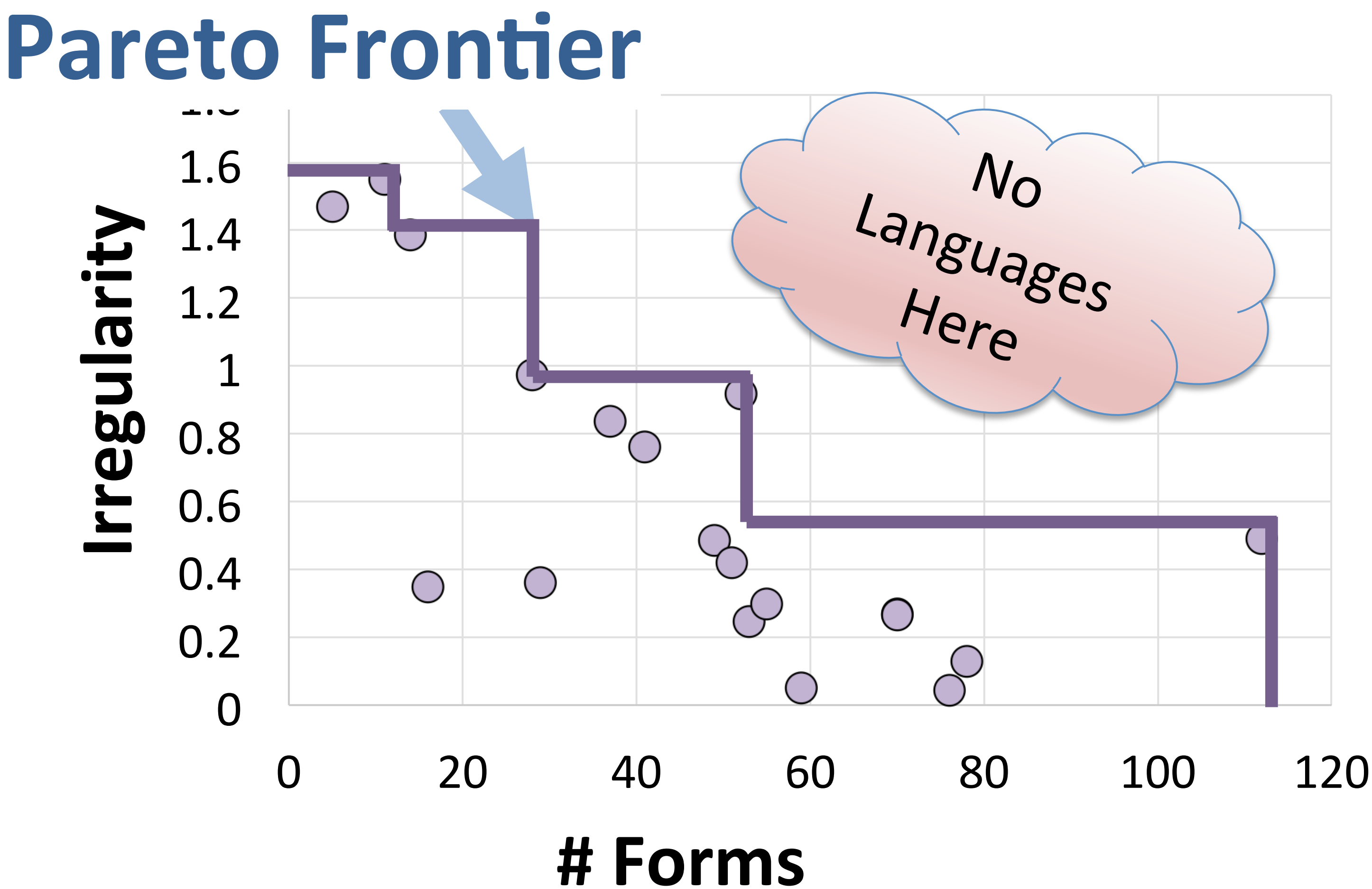
What about the Rest of the World's Languages? (For which we have data)



What about the Rest of the World's Languages? (For which we have data)

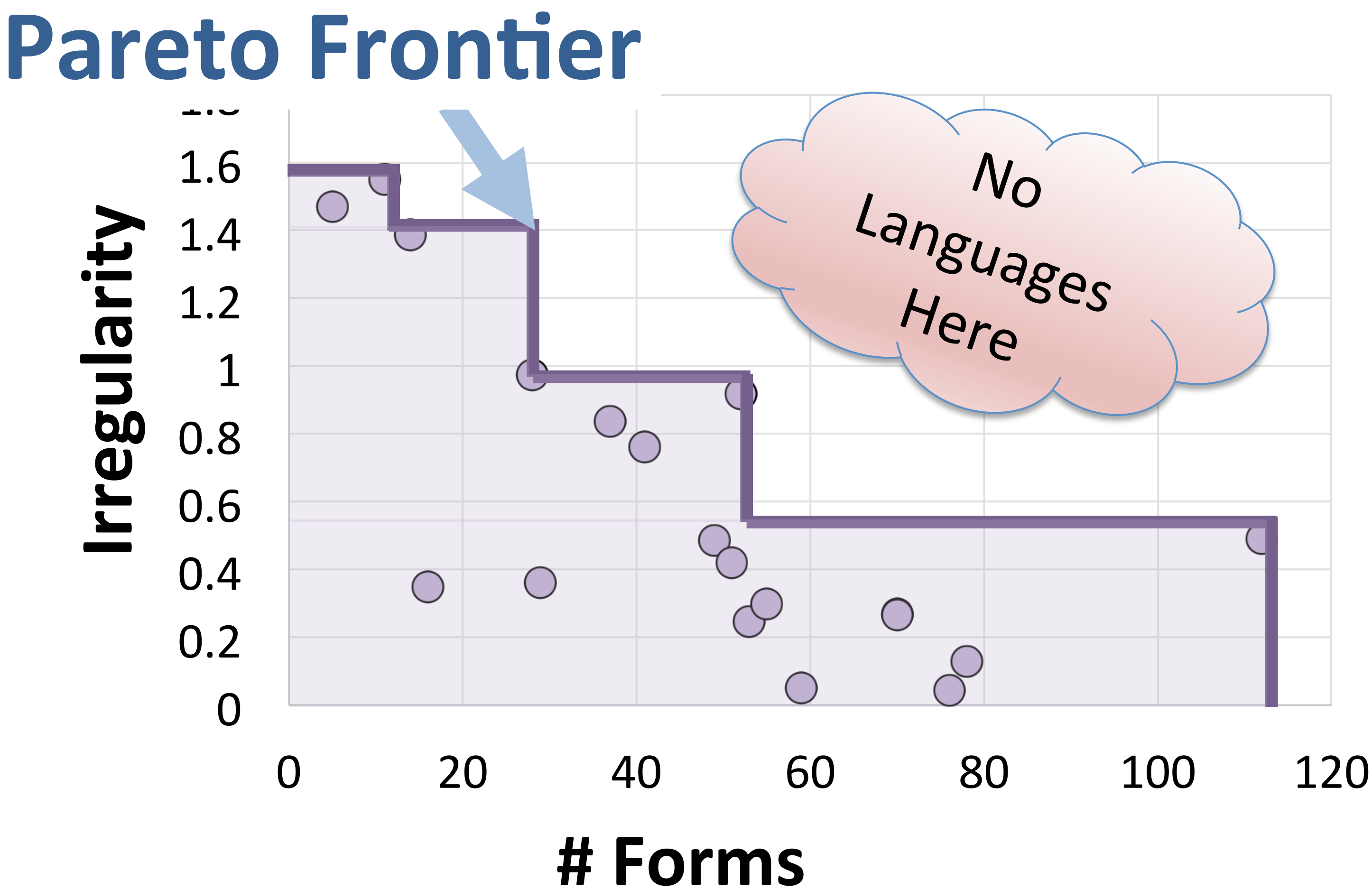


What about the Rest of the World's Languages? (For which we have data)



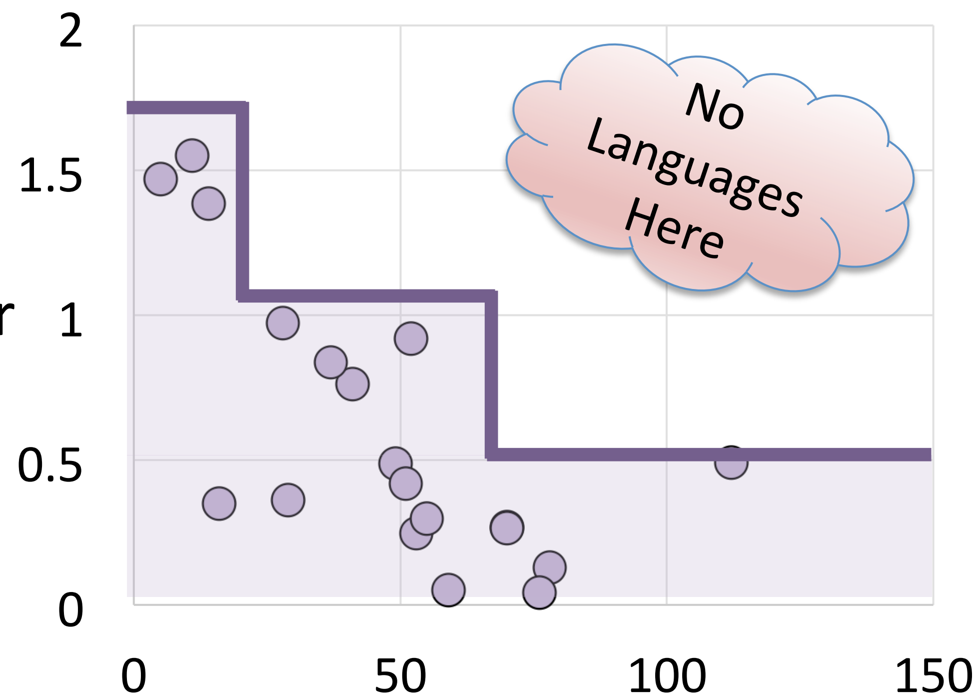
JOHNS HOPKINS
UNIVERSITY

What about the Rest of the World's Languages? (For which we have data)



Scientific Hypothesis about Language

- Use machine learning techniques to test hypothesis about Language
- Morphological systems can have *either* a lot of forms or lot of irregularity
  - But not both!
- Why? Speculative reason: memorizing a lot of irregulars would tax human memory



JOHNS HOPKINS UNIVERSITY

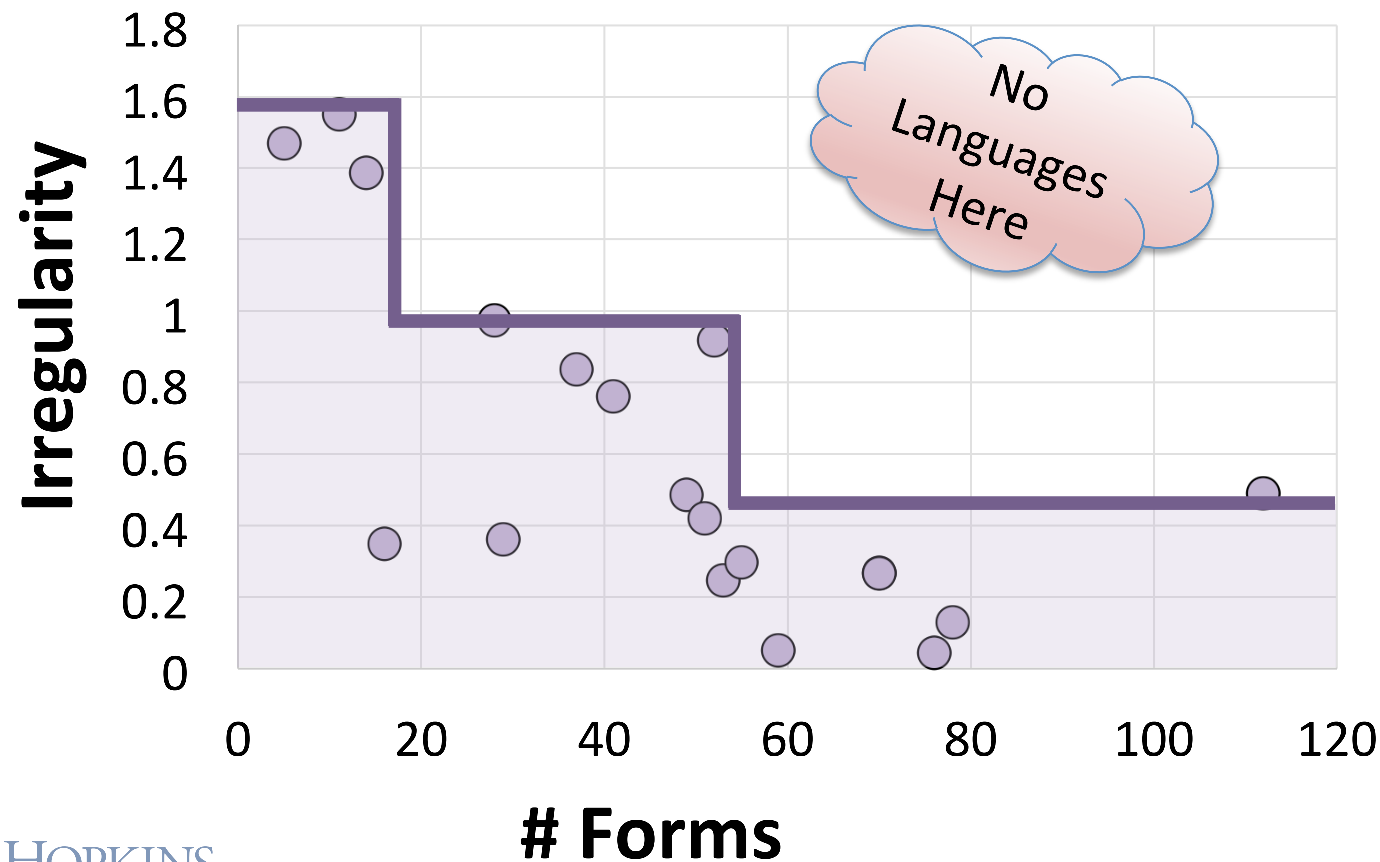
Chinese is Low on *Both* Dimensions of Morphological Complexity

- **Morphology in a language is not necessary!**
- Let’s look at the Chinese verb “to drink”
  - drink = 喝
  - drinking = 喝
  - drank = 喝

- Look mommy, no inflection!

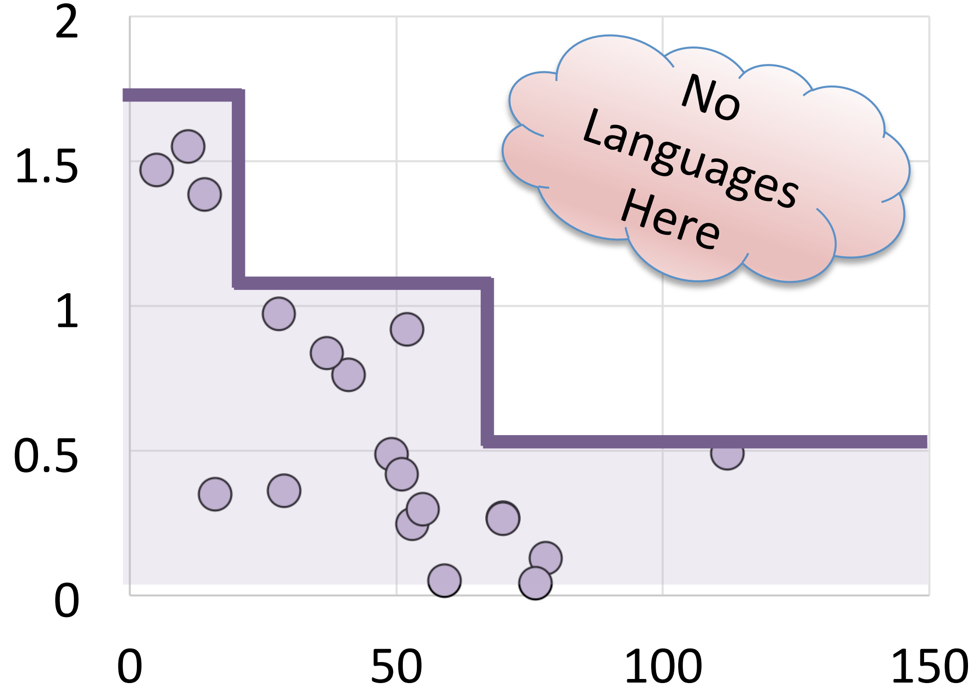
Our Hypothesis Again

- Inflectional, morphological systems have a lot of forms, or a lot irregularity, but not both

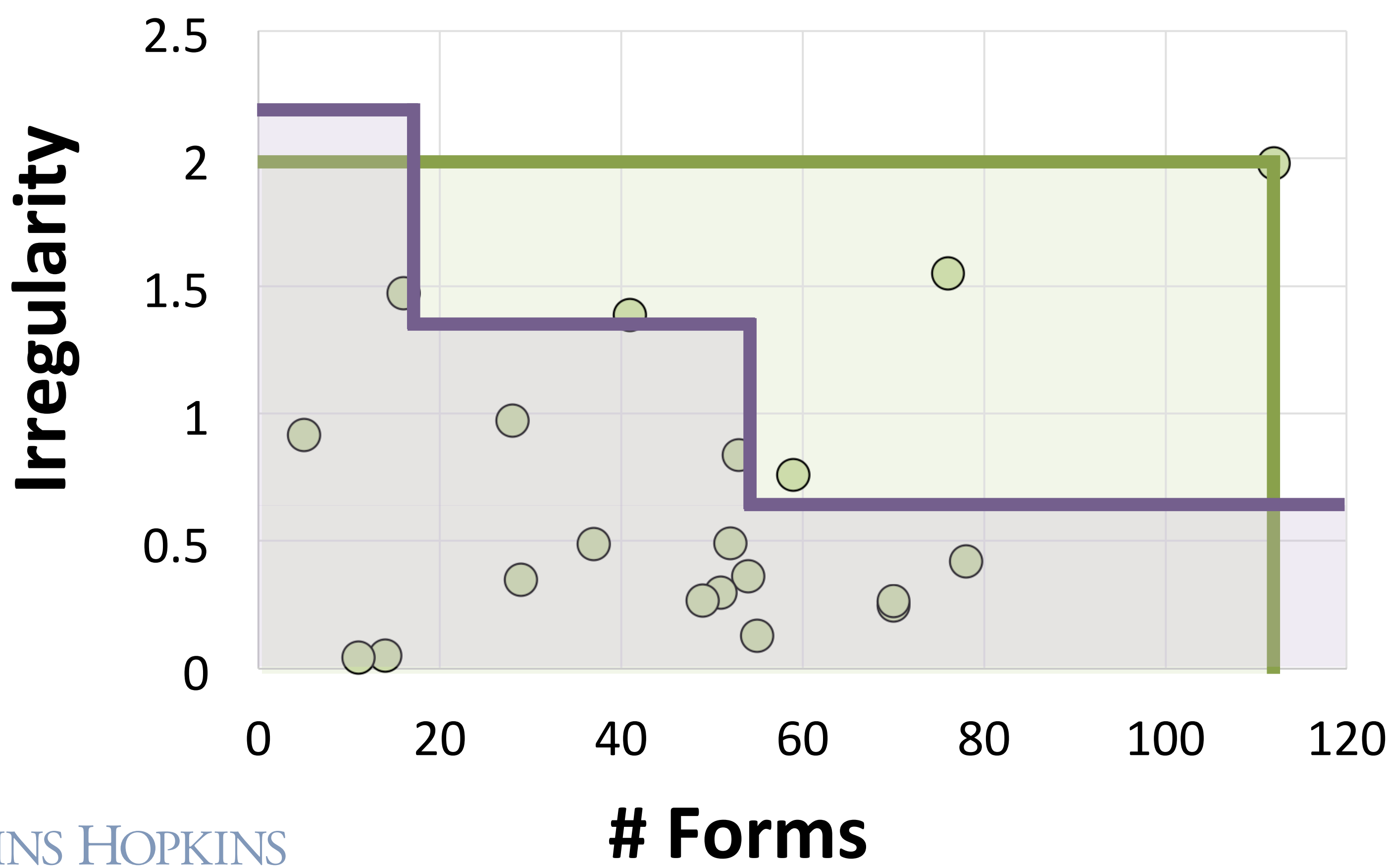


A Paired Permutation Test

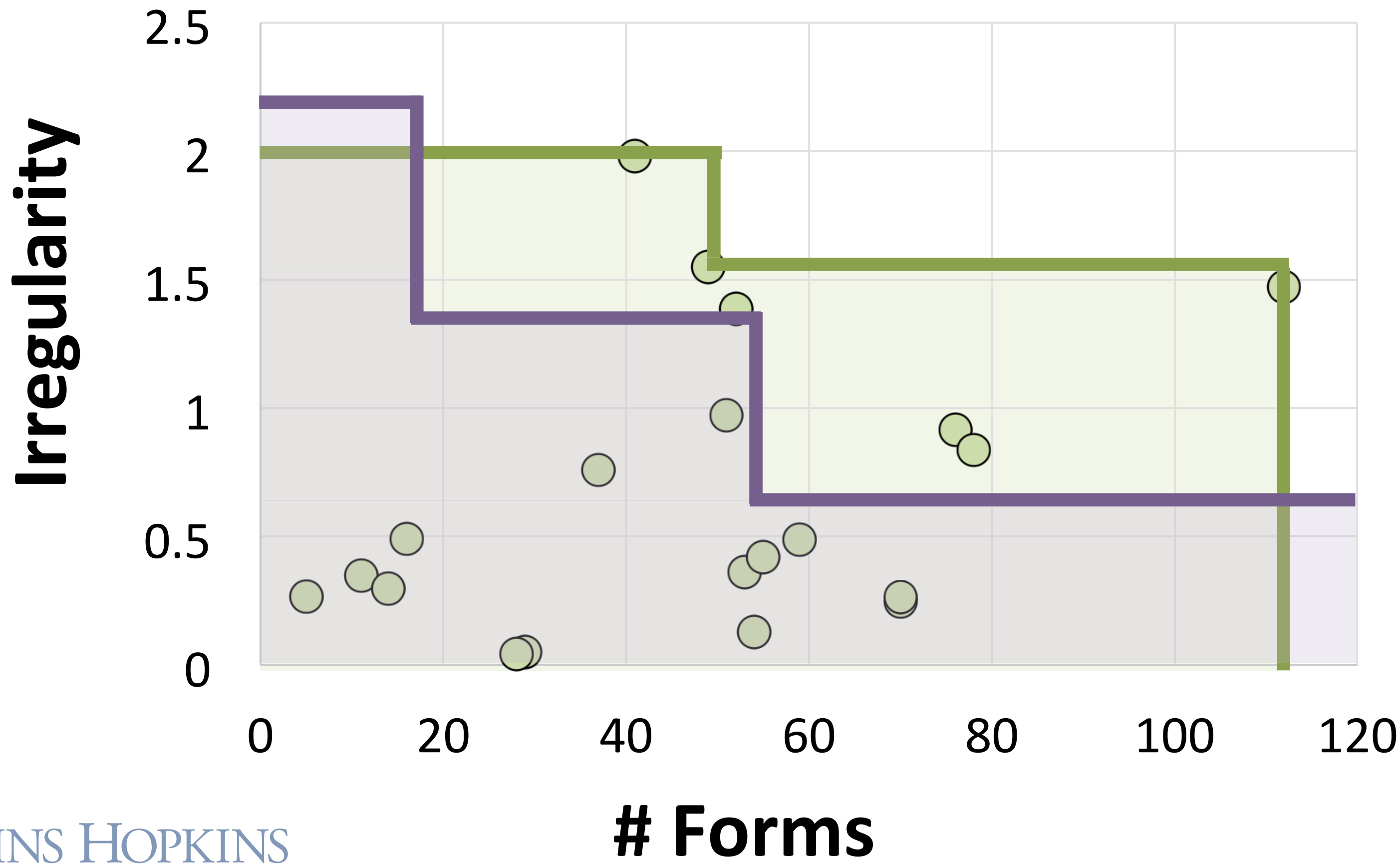
- There *appears* to be a trend, but is it significant?
  - Is the upper right-hand corner more empty that it would be by chance?
- New Significance Test
  - Keep x-axis in tact, shuffle y-axis
  - compare area under the Pareto curve
  - Non-parametric test



Random Morphological Trade-Off

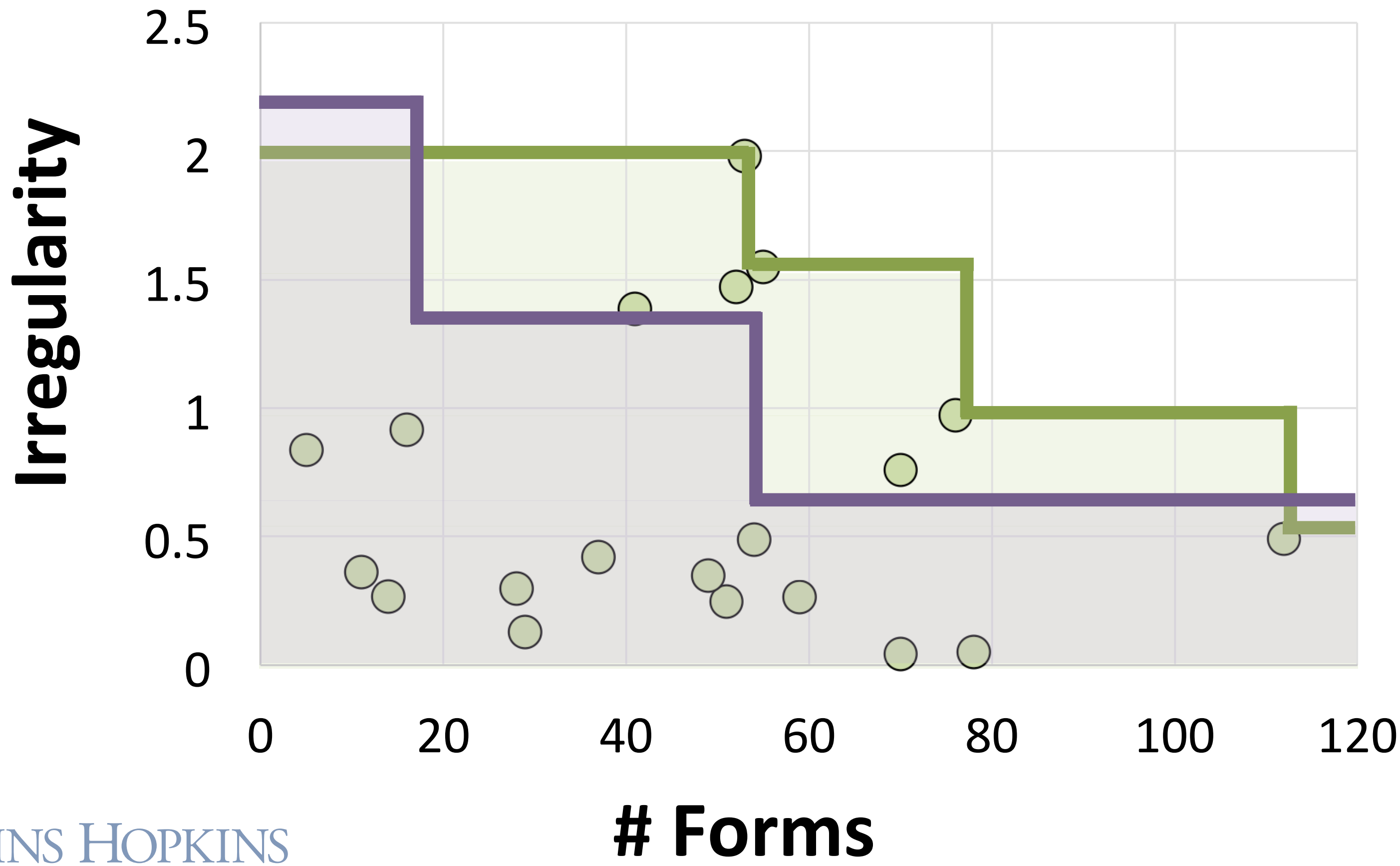


Random Morphological Trade-Off



JOHNS HOPKINS UNIVERSITY

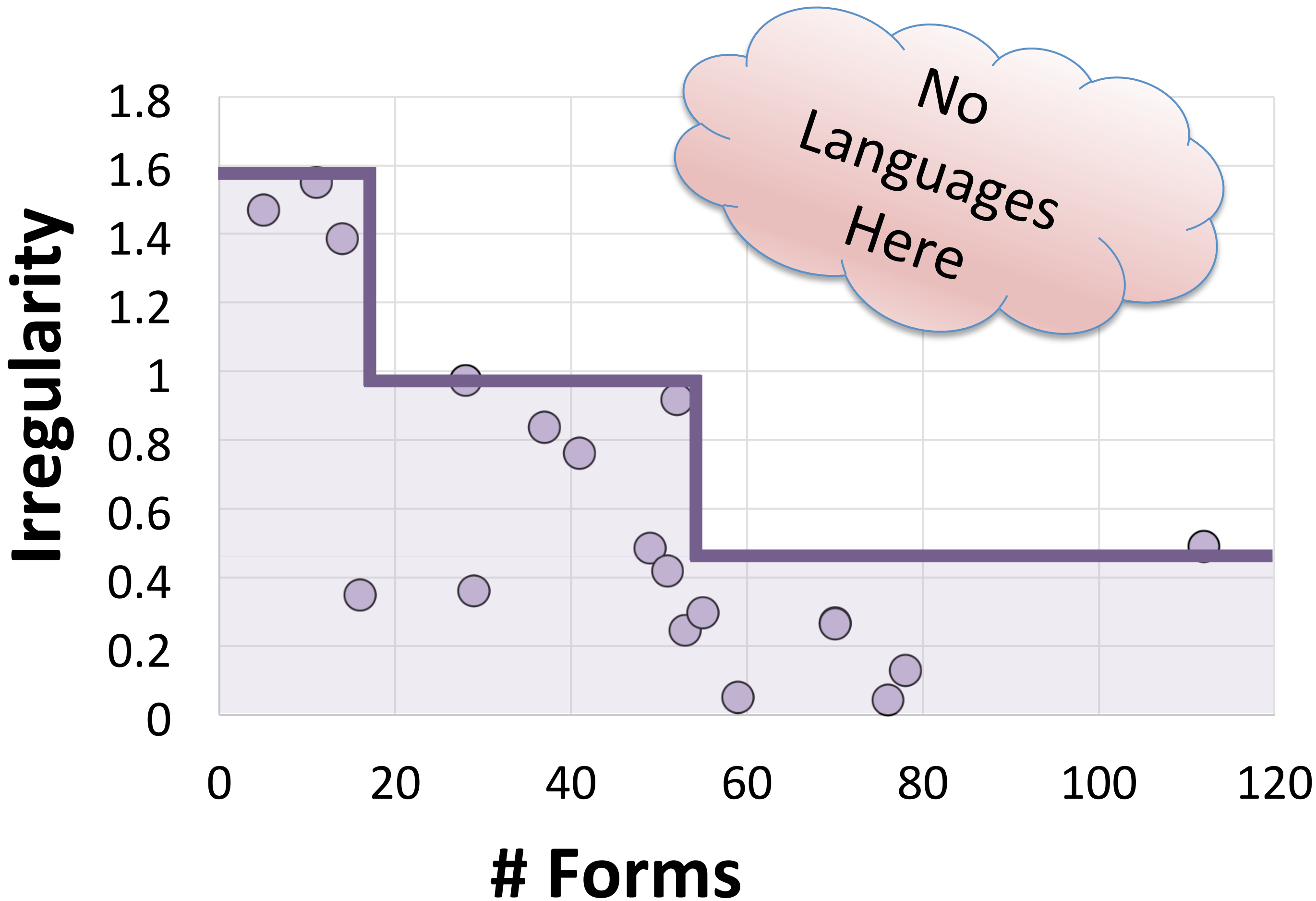
Random Morphological Trade-Off



JOHNS HOPKINS
UNIVERSITY

Scientific Finding

Gap in the upper right-hand size with $p < 0.05$



Caution: Limited # Of Languages

- We need to be very cautious about reporting the results!

- The languages are not i.i.d.
  - Some of them are genetically related
  - Focus on Western European Languages

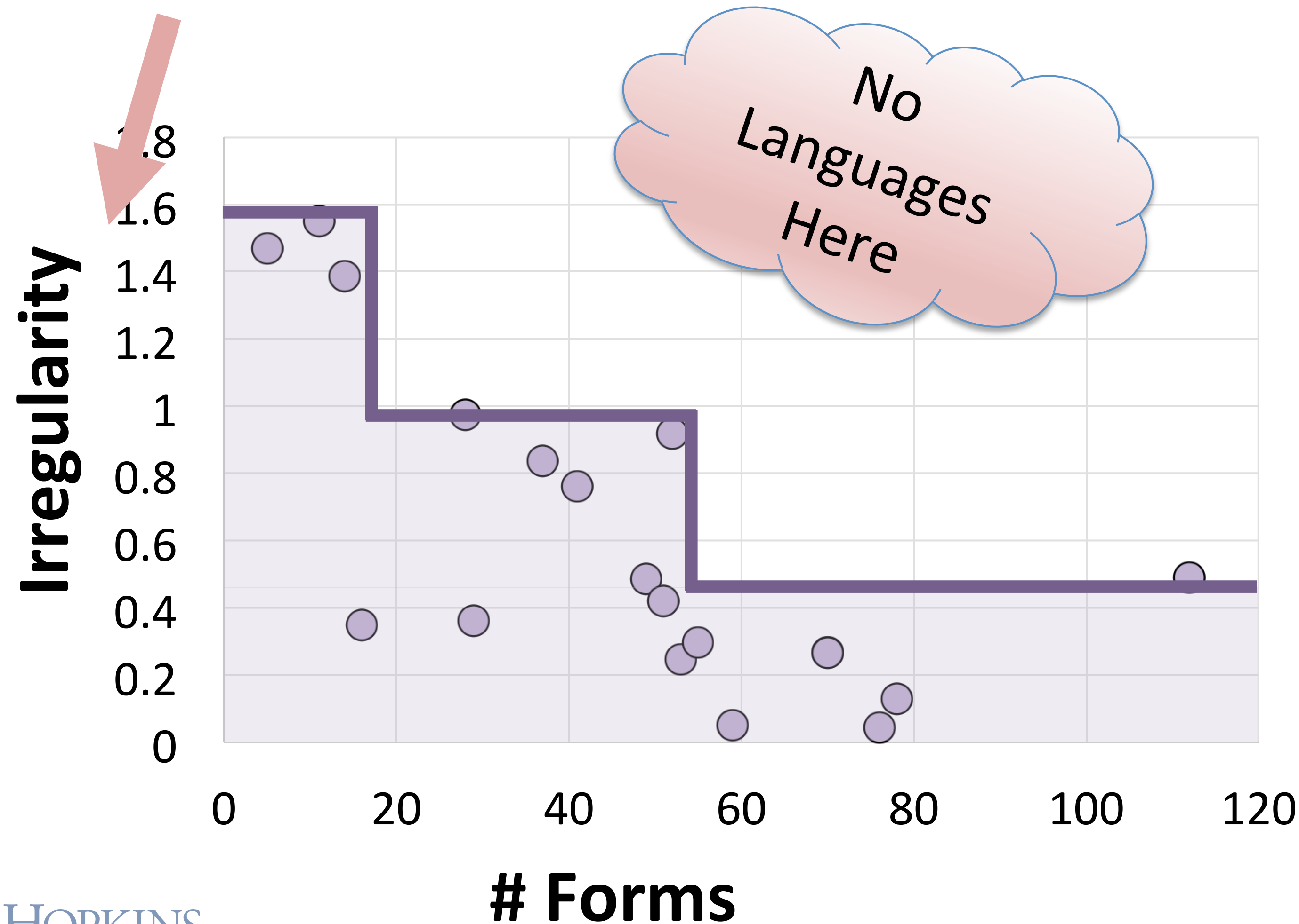
- We have a small sample of size of languages
  - There might be unobserved counterexamples
  - For *this sample*, the Pareto frontier leaves an unusually large gap in the upper right

Technical Contribution:

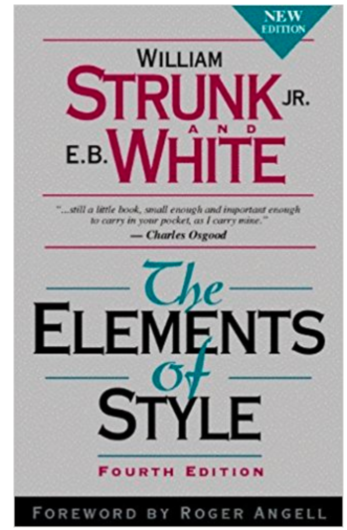
Operationalizing Morphological Irregularity

Where did the *y*-axis come from?

What do those numbers mean?



What’s an Irregular Verb?



- **TL;DR:** some grammarian said so
- **Example:** Spanish has three types of regular verbs
  - *ar, er, ir*
- The rest are “irregular”
  - Why???
- Are they equally irregular?
  - Or are some verbs more irregular than others?

CANTAR	BEBER	VIVIR
cant-é	beb-í	viv-í
cant-aste	beb-iste	viv-iste
cant-ó	beb-ió	viv-ió
cant-amos	beb-imos	viv-imos
cant-asteis	beb-isteis	viv-isteis
cant-aron	beb-ieron	viv-ieron

New Insight: We will tackle morphological irregularity *probabilistically*

Regularity = Predictability

- For each language
  - **Step 1:** Build a really good generative probability model *p* of the morphological paradigm
  - **Step 2:** Train its parameters on some data
  - **Step 3:** Irregularity = -log *p*(**held-out data**)

Morphological Reinflection

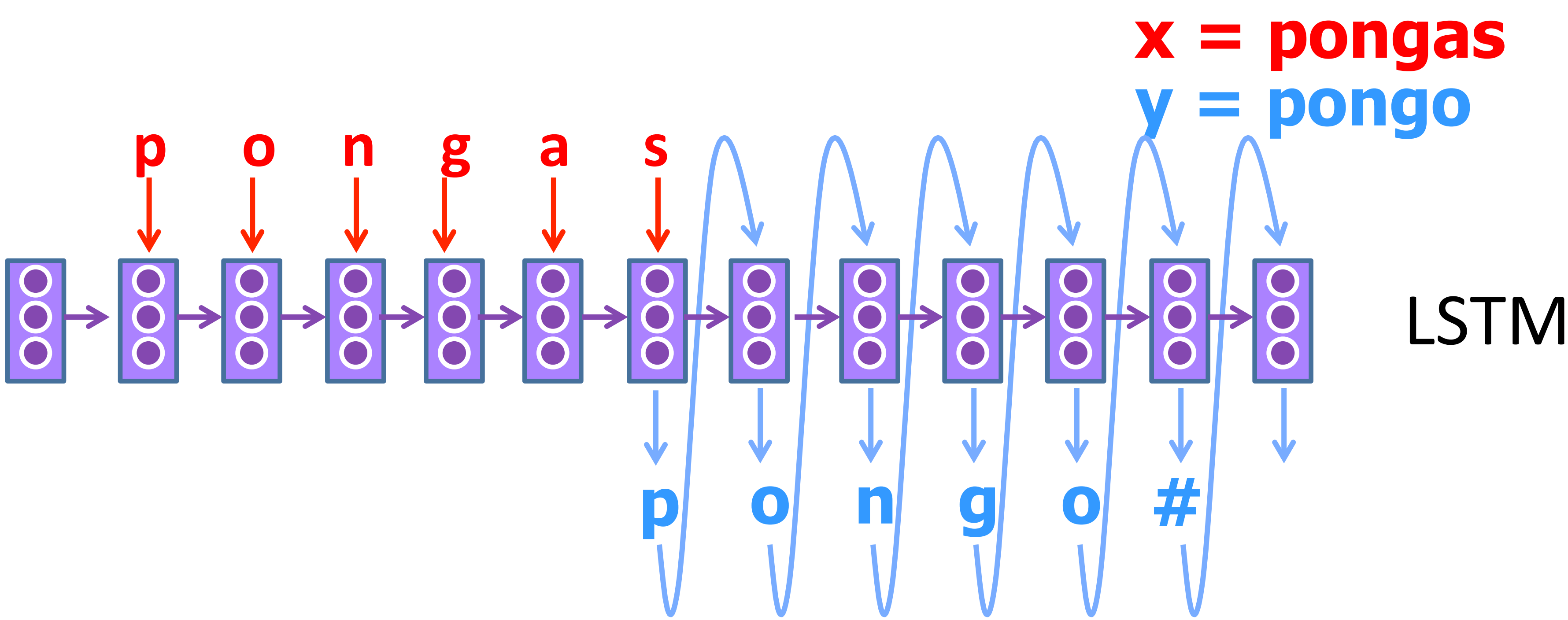
- Start with pair-wise probability distributions

$$p\,(\text{pongamos} \mid \text{pongo})$$

1ps;prs;sbjv;pl 1ps;prs;ind;sg

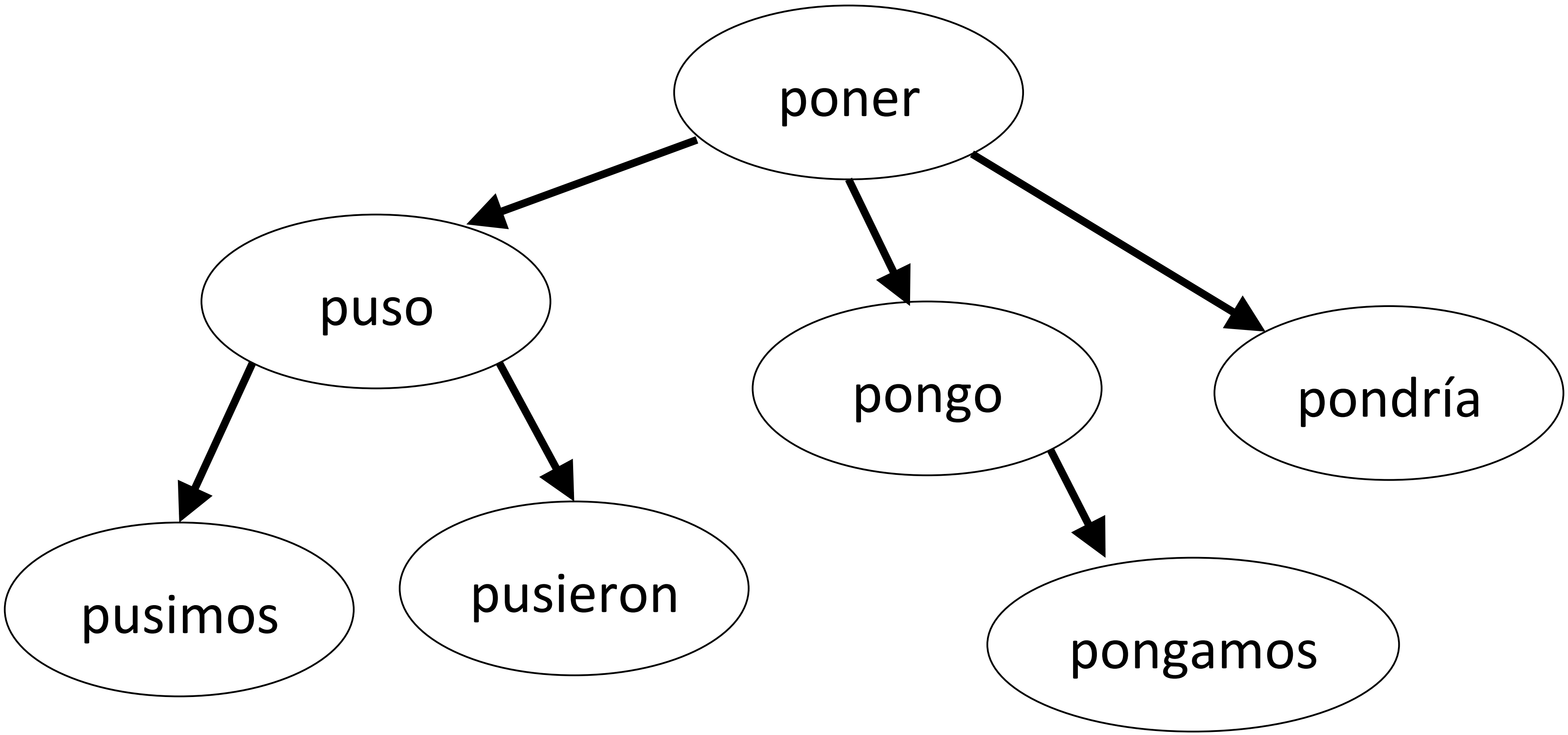
- In NLP, this task is known as **morphological reinflection**
  - Three shared tasks: SIGMORPHON (2016), CoNLL (2017, 2018)
  - Cotterell et al. (2016,2017) for overview of the results
  - State of the art: LSTM seq2seq model – same as MT

Sequence-to-Sequence Model

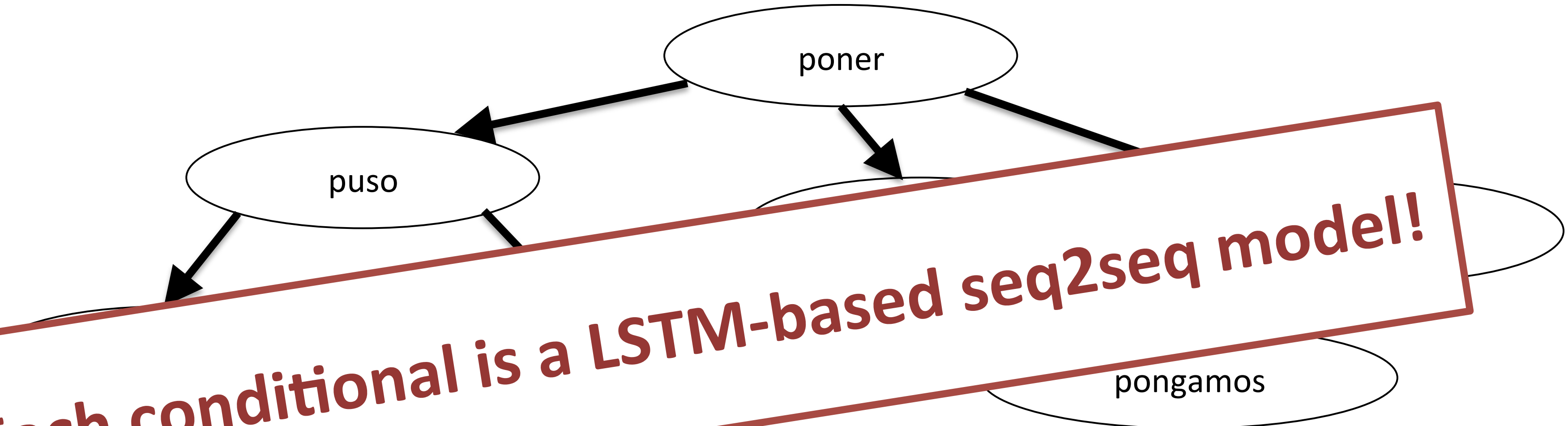


p(**y** | **x**) reads **x**, then stochastically emits chars of **y**, 1 by 1, like a language model

Arrange into a Bayesian Network



Neural Graphical Model Over Strings

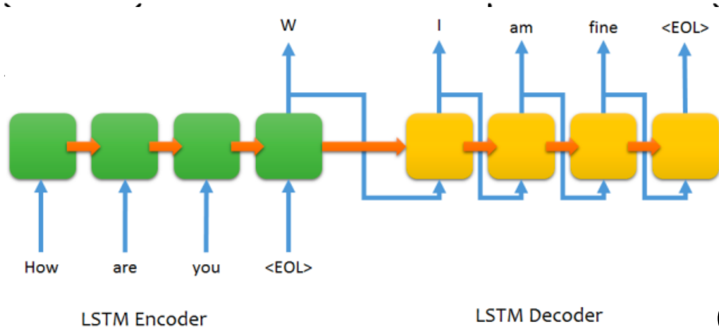


Each conditional is a LSTM-based seq2seq model!

$$p(\text{pusieron} \mid \text{puso})$$

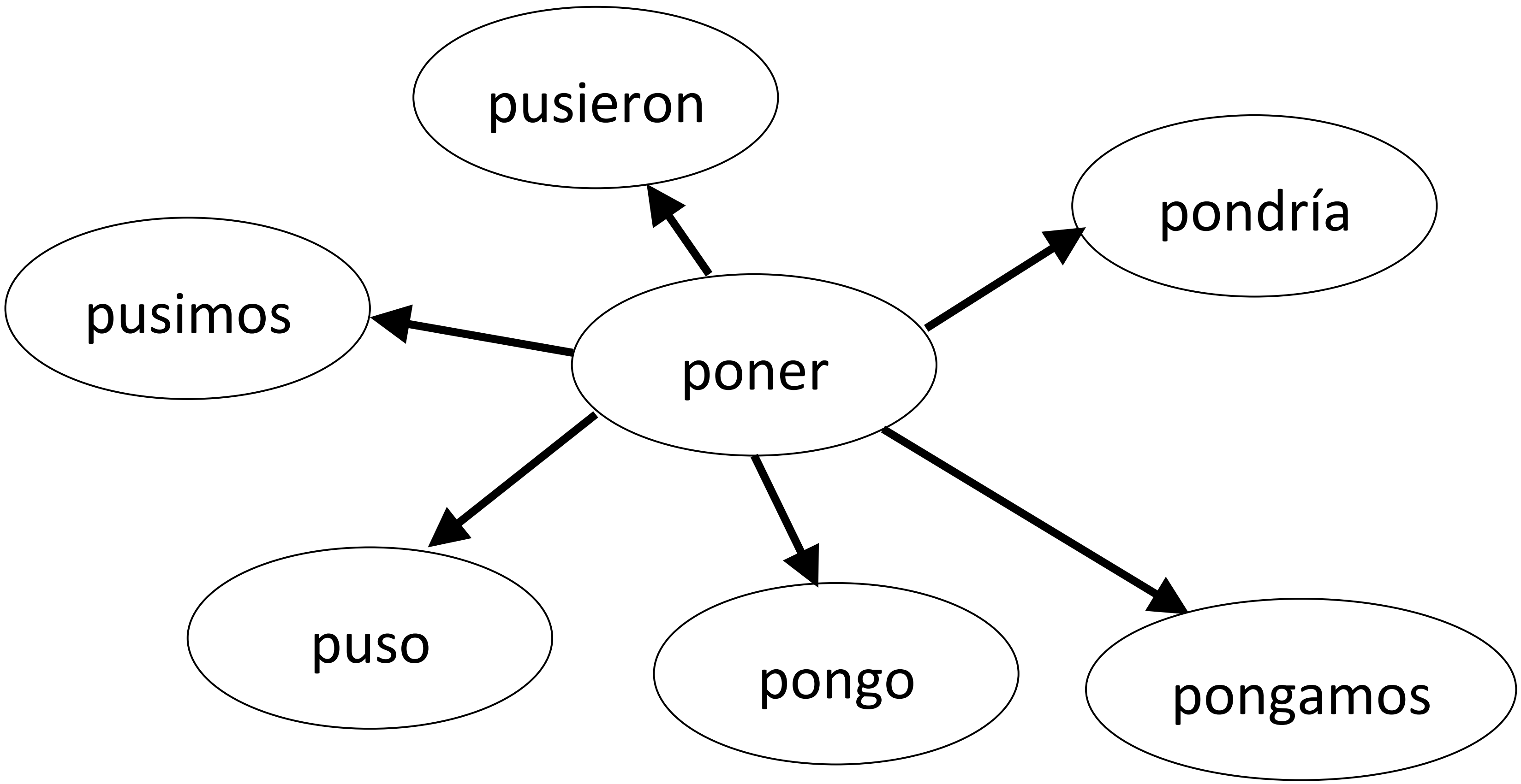
$$p(\text{pusimos} \mid \text{puso}) \cdot p(\text{pusieron} \mid \text{puso}) \cdot p(\text{puso} \mid \text{poner}) \cdot$$

$$p(\text{pongo} \mid \text{poner} \quad \ldots \quad) \cdot p(\text{pongamos} \mid \text{poner})$$



Cotterell et al. (2017) @ EACL 2017

Many Possible Networks



How to Choose Best Tree?

- Standard structure learning problem in graphical models
- Strategy: Tie parameters among all conditionals
  - Conditionals for every possible tree trained together
- Inspired by Chow-Liu Algorithm
  - Use Chu-Liu-Edmonds
  - Finds optimal directed spanning tree in $O(n^3)$ time

Experimental Languages

- Data from the UniMorph (Kirov et al. 2018)
- Selected languages with "enough" training examples
- Verbal Paradigms:
  - 23 languages / 3 families
- Nominal Paradigms
  - 31 languages / 3 families

Cross-linguistically Compatible Labels

Akademie	Akademie	N;ACC;SG
Akademie	Akademie	N;DAT;SG
Akademie	Akademie	N;GEN;SG
Akademie	Akademien	N;ACC;PL
Akademie	Akademien	N;DAT;PL
Akademie	Akademien	N;GEN;PL
Akademie	Akademien	N;NOM;PL
Akademie	Akademie	N;NOM;SG
Akademiker	Akademiker	N;ACC;PL
Akademiker	Akademiker	N;ACC;SG
Akademiker	Akademiker	N;DAT;SG
Akademiker	Akademiker	N;GEN;PL
Akademiker	Akademikern	N;DAT;PL
Akademiker	Akademiker	N;NOM;PL
Akademiker	Akademiker	N;NOM;SG
Akademiker	Akademikers	N;GEN;SG
...		

German Nominal Paradigms

Plug For UniMorph

- Now data for over 100 languages!
- Freely downloadable from unimorph.github.io

UniMorph
Universal Morphological Annotation

Schema Software Publications Contact

UniMorph

The Universal Morphology (UniMorph) project is a collaborative effort to improve how NLP handles complex morphology in the world's languages. The goal of UniMorph is to annotate morphological data in a universal schema that allows an inflected word from any language to be defined by its lexical meaning, typically carried by the lemma, and by a rendering of its inflectional form in terms of a bundle of morphological features from our schema. The specification of the schema is described here and in Sylak-Glassman (2016).

Estimating the Parameters

- Estimating morphological irregularity is now a standard machine learning problem

- Model is trained using gradient descent on UniMorph data
  - Best model selected on development data

- Irregularity = loss on held-out data

But why is there a trade-off?

- This paper shows the existence of a trade-off between two types of morphological complexity
- The real scientific question is *why*?
- On-going work guesses that it has to learnability and the learning infrequent, irregular forms
  - I.e., rare forms tend to regularize
- Artificial learnability study already available
  - Preliminary version on arXiv

Linguistic Complexity More Broadly

A Twitter Poll About Complexity

Ryan D. Cotterell
@_shrdlu_

Do you think your native language is more complex than English? (If you speak a language other than English natively.) #acl2018

Also, come to my talk at 17:00 tomorrow @acl2018 on that very topic!

79%	Yes
21%	No

109 votes • 1 day left

Equal Complexity Hypothesis

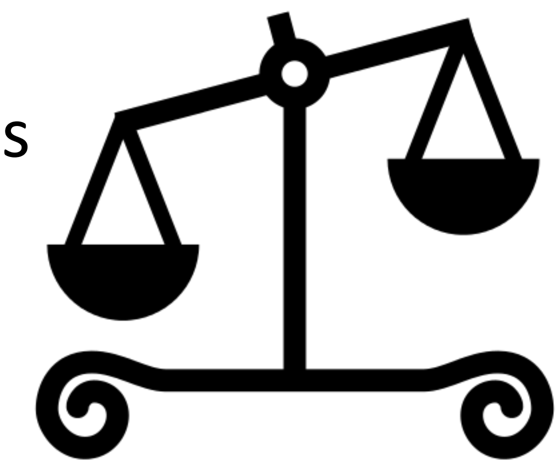
- Hockett (1958) argued that all languages are equally complex
- Idea goes back much further in the linguistics literature

- All languages appear to optimize for efficient communication subject to learnability

Complexity Trade-Offs

- **Corollary**: if one facet of a language is more complex, another is simpler to compensate

- **Trade-Off Example**:
  - German has more inflected forms than English (morphology)
  - English has a more complicated tense system (syntax)

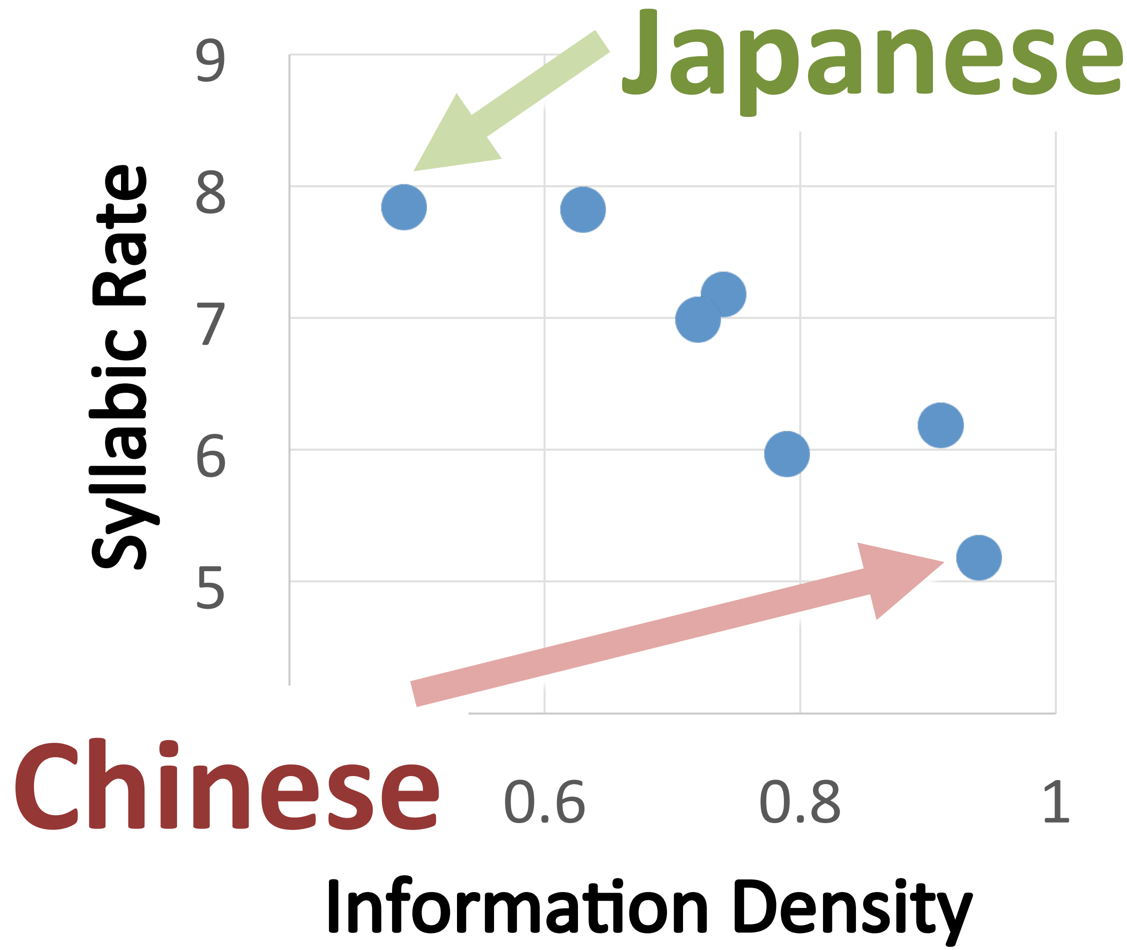


Example: Rate Of Speech (Pellegrino 2011)

- Are all languages spoken equally fast?

  No!

- Spoken Rapidly
  - Spanish, Japanese
- Spoken Slowly
  - English, Chinese



John McWhorter on Creoles

- McWhorter wrote the seminal paper in 2001
- Argues creoles are in fact less complex
- Complexity accretes over time
  - Creoles are new languages

Published Work

- Check out our NAACL 2018 paper: *All Are Languages Equally Hard to Language-Model?*

Future Work

- We only looked a specific trade-off in morphological complexity
  - Data-driven methods for trade-offs in other areas of linguistics

- Extensions look at language more holistically
  - Trade-offs between morphology and phonology
  - Trade-offs between morphology and syntax

- Why didn't linguistics already solve this problem?
  - No big data, no methods

Fin