On the Complexity and Typology of Inflectional Morphological Systems

Ryan Cotterell, Christo Kirov, Jason Eisner, and Mans Hulden

SCIL 2018

Machine Learning ∩ Linguistics

- "**Computational linguistics** is analogous to computational biology or any other computational fill-in-the-blank. It develops computational methods to answer the scientific questions of linguistics."
- "**Natural language processing** is the art of solving engineering problems that need to analyze (or generate) natural language text. Here, the metric of success is not whether you designed a better scientific theory or proved that languages X and Y were historically related. Rather, the metric is whether you got good solutions on the engineering problem."

Statistical Computational Linguistics: machine learning meets linguistic theory

Introduction

- **Question:** What generalizations hold for the typology of morphological irregularity?
  - What makes an inflectional morphology system "complex"?
    - The size of the inflectional paradigms? (E-Complexity)
    - The predictability of inflected forms given other forms? (I-Complexity)
  - Hypothesis: There is a trade-off between E-Complexity and I-Complexity. Languages may have large paradigms, or highly irregular paradigms, but not both.
  - We formalize this hypothesis and verify it quantitatively in 31 diverse languages using machine learning tools.

Typology of Morphological Irregularity

- Intuition: smaller inflectional systems admit more irregularity than larger systems
- English Verbal System:
  - 5 forms
  - 300+ irregulars
- Turkish Verbal System
  - 100+ forms
  - 1 irregular

- Goal: Can we quantify this? Does it generally hold true?

What is an Irregular Verb?

- Spanish has three regular conjugations.
- But why is *poner* irregular? Many verbs pattern the same way…
- (yo pongo - yo tengo)

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

Word-Based Morphology (Aronoff 1976)

- An **inflected lexicon** is a set of word types, where each is a triple of:
  - **lexeme**: arbitrary index of a word's core meaning
  - **slot**: arbitrary index indicating the inflection of the word
  - **surface form**: a string over a fixed alphabet

- All words that share the same lexeme form a **paradigm**, with slots filled by surface forms. {go, goes, went}

- Each slot represents a morpho-syntactic bundle of representative features: [TENSE=PRESENT, MOOD=SUBJUNCTIVE, PERSON=2, NUMBER=SG]

Enumerative (E) Complexity (Ackerman & Malouf 2013)

- Complexity based on **counting**. Number of slots in a **paradigm** x number of exponents per slot.

- Here, for a particular part of speech, the average **paradigm** size across all **lexemes**.

- English verbs might have just a few paradigm slots, while Archi verbs might have thousands. Does this make Archi more complex?

Integrative (I) Complexity (Ackerman & Malouf 2013)

- How predictable is any given surface form given additional knowledge about the paradigm?

- Measures how **irregular** an inflectional system is.

The Low-Entropy Conjecture

“the hypothesis that enumerative morphological complexity is effectively unrestricted, as long as the average conditional entropy, a measure of integrative complexity, is low.” (Ackerman and Malouf, 2013)

E-complexity can be arbitrary, but I-complexity (irregularity) is low.

Here: There is a trade-off. Either E-Complexity or I-Complexity can be high, but not both.

Calculating I-Complexity (Ackerman & Malouf 2013)

	SINGULAR				PLURAL			
CLASS	NOM	GEN	ACC	VOC	NOM	GEN	ACC	VOC
1	*-os*	*-u*	*-on*	*-e*	*-i*	*-on*	*-us*	*-i*
2	*-s*	-∅	-∅	-∅	*-es*	*-on*	*-es*	*-es*
3	-∅	*-s*	-∅	-∅	*-es*	*-on*	*-es*	*-es*
4	-∅	*-s*	-∅	-∅	*-is*	*-on*	*-is*	*-is*
5	*-o*	*-u*	*-o*	*-o*	*-a*	*-on*	*-a*	*-a*
6	-∅	*-u*	-∅	-∅	*-a*	*-on*	*-a*	*-a*
7	*-os*	*-us*	*-os*	*-os*	*-i*	*-on*	*-i*	*-i*
8	-∅	*-os*	-∅	-∅	*-a*	*-on*	*-a*	*-a*

Modern Greek Analysis

Probability of swapping one exponent for another:

$$r(m_i \mid m_j)$$

$$r(m_{\text{GEN;SG}} = \text{-}us \mid m_{\text{ACC;PL}} = \text{-}i) = 1$$

$$r(m_{\text{GEN;SG}} = \text{-}o \mid m_{\text{ACC;PL}} = \text{-}a) = \frac{1}{3}$$

$$r(m_{\text{GEN;SG}} = \emptyset \mid m_{\text{ACC;PL}} = \text{-}a) = \frac{2}{3}$$

Calculating I-Complexity (Ackerman & Malouf 2013)

	SINGULAR				PLURAL			
CLASS	NOM	GEN	ACC	VOC	NOM	GEN	ACC	VOC
1	*-os*	*-u*	*-on*	*-e*	*-i*	*-on*	*-us*	*-i*
2	*-s*	-∅	-∅	-∅	*-es*	*-on*	*-es*	*-es*
3	-∅	*-s*	-∅	-∅	*-es*	*-on*	*-es*	*-es*
4	-∅	*-s*	-∅	-∅	*-is*	*-on*	*-is*	*-is*
5	*-o*	*-u*	*-o*	*-o*	*-a*	*-on*	*-a*	*-a*
6	-∅	*-u*	-∅	-∅	*-a*	*-on*	*-a*	*-a*
7	*-os*	*-us*	*-os*	*-os*	*-i*	*-on*	*-i*	*-i*
8	-∅	*-os*	-∅	-∅	*-a*	*-on*	*-a*	*-a*

Modern Greek Analysis

Probability of swapping one exponent for another:

$$r(m_i \mid m_j)$$

Conditional entropy between slots:

$$H(i \mid j) = -\sum_{m_i \in \Sigma^*} r(m_i) \log r(m_i \mid m_j)$$

Average of conditional entropies:

$$\frac{1}{n^2 - n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} H(i \mid j)$$

Calculating I-Complexity (Ackerman & Malouf 2013)

Calculation is analysis-dependent.

- Only assigns probabilities to limited set of suffixes/prefixes in analysis tables, rather that arbitrary strings. This precludes assigning probability to e.g., suppletive forms.

Average conditional entropy overestimates I-Complexity.

- Implies all cell-2-cell transformations are equally likely.
- Predicting German Händen (DAT, PL) from Hand (NOM, SG) is difficult, but easy from Hände (NOM, PL)

Joint Entropy as I-Complexity

If we had joint distribution over all cells in a paradigm:

$$p(m_{\text{LEMMA}}, m_{\text{3PS}}, m_{\text{PAST}}, m_{\text{GERUND}})$$

Then complexity could be calculated as the entropy of this distribution H(p):

$$-\sum_{i=1}^{n} \sum_{\vec{m} \in (\Sigma^*)^n} p(m_1, \ldots, m_n) \log_2 p(m_1, \ldots, m_n)$$

Morphological Knowledge as a Distribution

$p(run, runs, running, ran)$ close to unigram frequency

$p(run, \cancel{snur}, running, \cancel{nar})$ close to 0

$p(sprint, sprints, sprinting \mid sprinted)$ close to 1

$p(wug, wugs, wugging \mid wugged)$ close to 1

A Variational Upper Bound on Entropy

True joint distribution (and its entropy) are horribly intractable!

We use a stand-in distribution q in place of the true joint p, attempting to minimize their KL-divergence:

$$-\sum_{\vec{m}\in(\Sigma^*)^n} p(m_1,\ldots,m_n)\log q(m_1,\ldots,m_n)$$

By maximizing the likelihood of some training data according to q:

$$\sum_{\vec{m}\in\mathcal{D}_{\text{train}}} \log q(m_1,\ldots,m_n).$$

We can estimate i-complexity from test data:

$$H(p,q) \approx -\frac{1}{d}\sum_{\vec{m}\in\mathcal{D}_{\text{test}}} \log q(m_1,\ldots,m_n)$$

A Generative Model of the Paradigm

Tree-structured Bayesian graphical model provides variational approximation (q) of joint paradigm distribution (p):

$$q_{\boldsymbol{\theta}}(m_1, \ldots, m_n) = \prod_{i=1}^{n} q_{\boldsymbol{\theta}}(m_i \mid m_{\mathrm{pa}_{\mathcal{T}}(i)})$$

A Generative Model of the Paradigm

- Start with pair-wise probability distributions

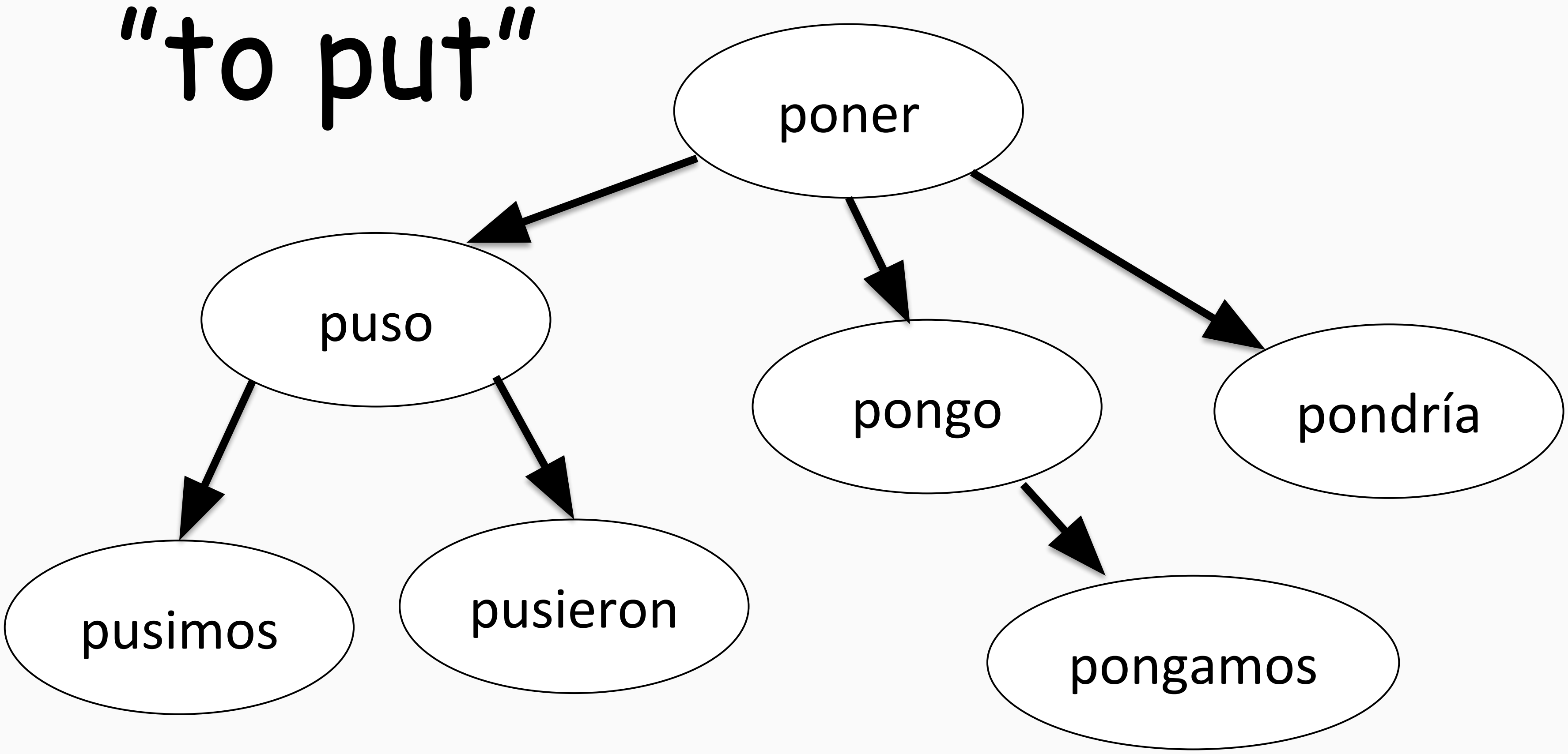
$$p\left(\text{pongamos} \mid \text{pongo}\right)$$

1ps;prs;sbjv;pl

1ps;prs;ind;sg

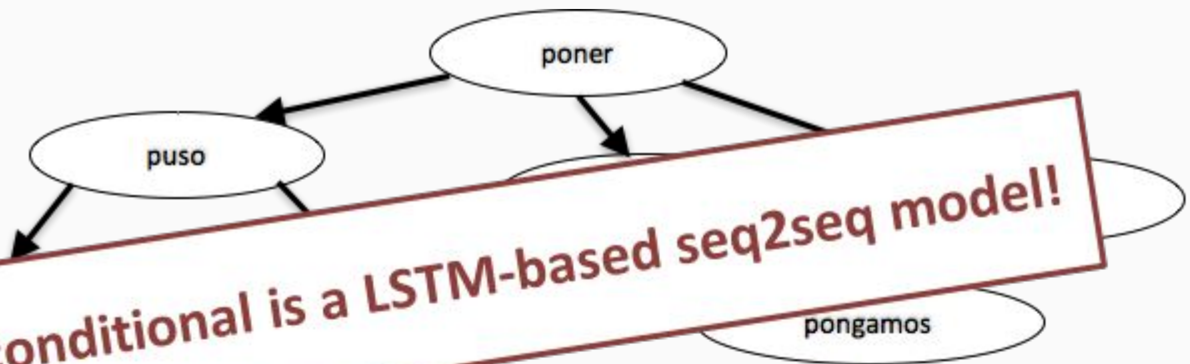
- In NLP, this task is known as morphological reinflection
  - Three shared tasks: SIGMORPHON (2016), CoNLL (2017, 2018)
  - Cotterell et al. (2016,2017) for overview of the results
  - State of the art: LSTM seq2seq model with attention (Bahdanau 2015)

A Generative Model of the Paradigm



Cotterell et al. (2017) @ EACL 2017

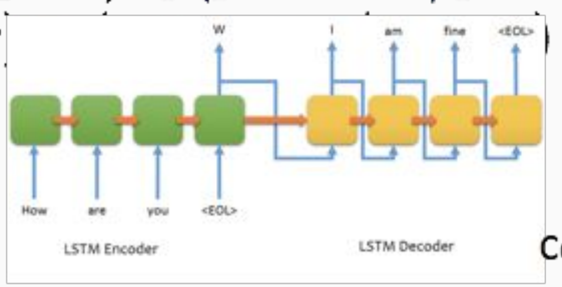
Generative Model of the Paradigm



Each conditional is a LSTM-based seq2seq model!

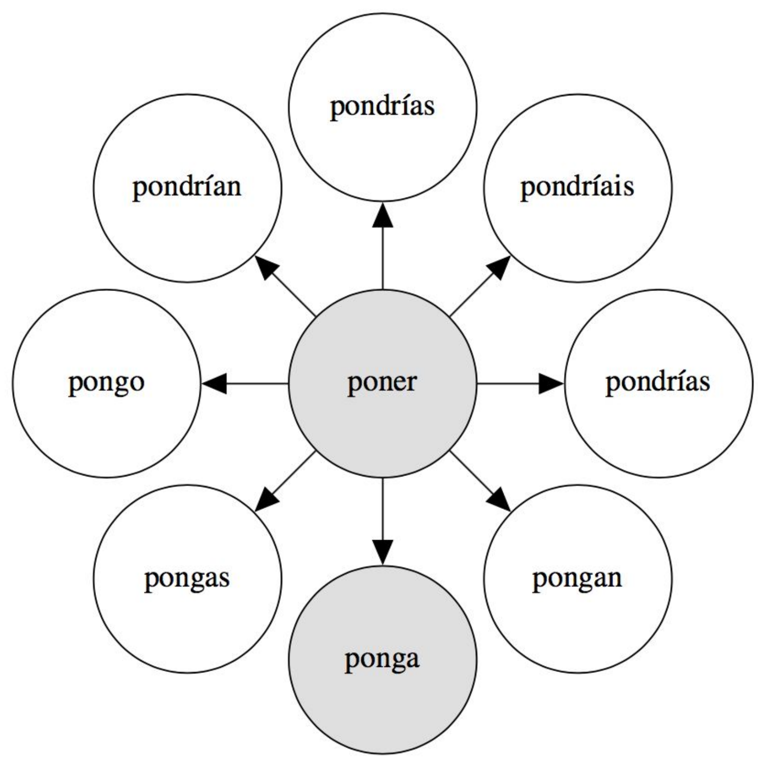
$$p(\text{pusieron} \mid \text{puso})$$

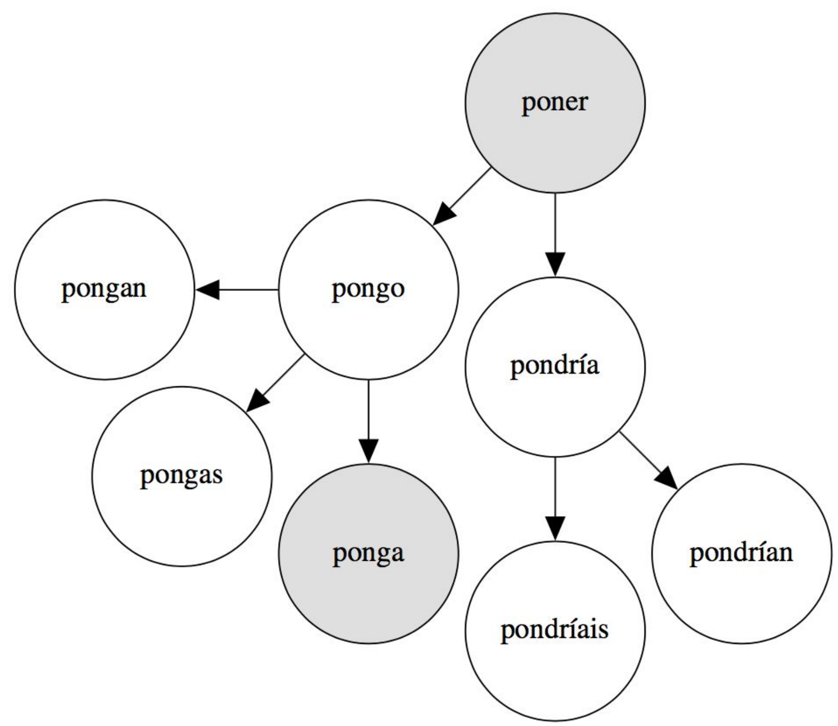
$$p(\text{pusimos} \mid \text{puso}) \cdot p(\text{pusieron} \mid \text{puso}) \cdot p(\text{puso} \mid \text{poner}) \cdot$$
$$p(\text{pongo} \mid \text{poner} \ldots) \cdot p(\text{pongamos} \mid \text{poner})$$



Cotterell et al. (2017) @ EACL 2017

Tree-structured Graphical Model for Paradigms





Selecting a Tree Structure

Use Edmonds (1967) algorithm to select the highest weighted directed spanning tree over all paradigms.

Edge weights:

$$\frac{1}{d} \sum_{\vec{m} \in \mathcal{D}_{\text{dev}}} \log q(m_i \mid m_j)$$

Vertex weights:

$$\frac{1}{d} \sum_{\vec{m} \in \mathcal{D}_{\text{dev}}} \log q(m_i \mid \text{empty string})$$

Data and Annotation

```
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

...
```

Annotated paradigms sources from the UniMorph Dataset (Kirov et al. 2018). https://unimorph.github.io/

Paradigm slot feature bundles annotated in UniMorph Schema (Sylak-Glassman et al. 2015)

23 languages sourced for verb paradigms. 31 languages sourced for noun paradigms.

	Nouns		Verbs	
Language	$\lvert\pi\rvert$	$H(p, q_{\boldsymbol{\theta}})$	$\lvert\pi\rvert$	$H(p, q_{\boldsymbol{\theta}})$
Arabic	112	0.44	36	0.21
Armenian	–	–	34	0.23
Bulgarian	52	0.666	9	0.22
Catalan	53	0.24	–	–
Czech	–	–	14	0.61
Danish	–	–	6	1.67
Dutch	16	0.24	–	–
English	5	0.27	2	0.10
Estonian	–	–	30	0.38
Faroese	14	1.24	16	0.21
Finnish	–	–	28	0.11
French	49	0.32	–	–
Georgian	–	–	19	0.61
German	29	0.32	8	0.77
Hungarian	59	0.04	34	0.38
Icelandic	–	–	16	0.66
Irish	–	–	13	0.06
Latin	100	0.59	12	0.12
Latvian	–	–	12	0.12
Lithuanian	–	–	14	1.04
Lower Sorbian	–	–	18	0.84
Macedonian	79	0.33	11	0.17
Northern Kurdish	–	–	20	0.67
Northern Sami	54	1.23	13	0.80
Norwegian Bokmål	5	2.12	3	0.71
Norwegian Nynorsk	–	–	3	0.46
Polish	–	–	14	0.80
Romanian	37	0.76	6	1.54
Russian	25	0.27	12	1.67
Serbo-Croatian	70	0.08	14	1.41
Slovak	–	–	12	1.64
Slovenian	–	–	18	0.69
Spanish	–	–	70	0.30
Swedish	11	1.04	8	0.15
Turkish	120	0.65	108	0.26
Ukrainian	–	–	14	0.85

Neural Sequence-2-Sequence Model

Encoder-Decoder architecture with attention, parameterized as in Kann & Shutze (2016)

- Bidirectional LSTM encoder.
- Unidirectional LSTM decoder.
- 100 hidden units
- 300 units per character embedding

Single network learns all mappings between paradigm slots:

H a n d IN=NOM IN=SG OUT=NOM OUT=PL -> H ä n d e

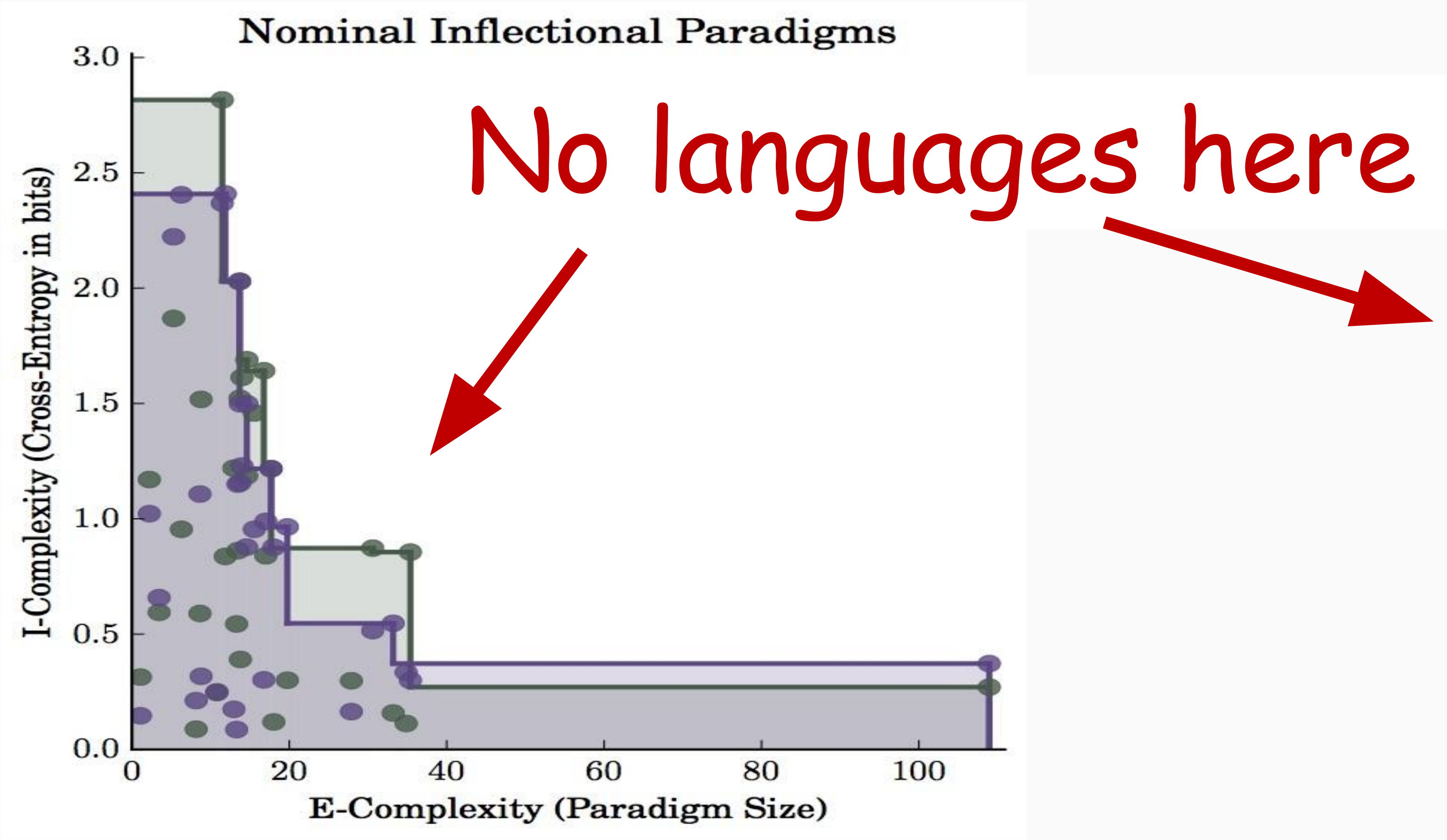
Experimental Details

For all experiments:

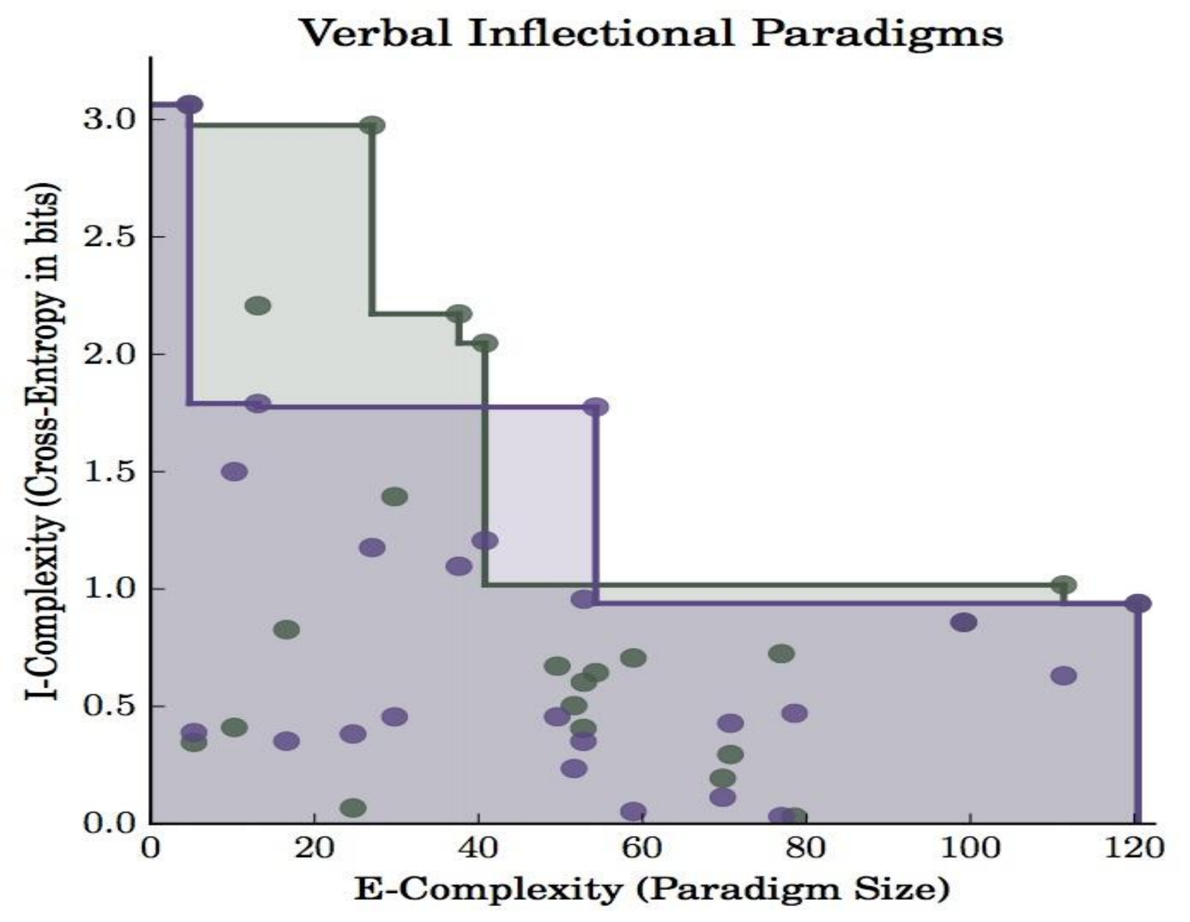
Held out 50 full paradigms for Dev set, 50 for Test set.

- Regime 1: Equal Number of Paradigms (Purple):
  - 600 complete paradigms for training (all n^2 mappings)
  - More training data for languages with larger paradigms
- Regime 2: Equal Number of Transformation Pairs (Green):
  - 60,000 mappings for training sampled at uniform from all mappings
  - Fewer examples per mapping for languages with larger paradigms

Noun Results



Verb Results



Discussion and Analysis

There appears to be a trade-off between between paradigm size and irregularity. Upper-right area of graph is NOT empty by chance.

Non-parametric test:

- Create 10,000 graph permutations by randomly assigning existing y coordinates to x coordinates
- Check how often upper-right area of true curve is emptier (contains fewer points) than random permutation.

$p < 0.05$ for both parts-of-speech and both training regimes

Next Steps

- We still have to explain why this trend exists!
- How much is due to model choices (seq2seq)?

- Is there a relationship between irregularity and learnability?
- **Conjecture**: only frequent irregular forms can exist and large systems dilute frequency of individual types
  - Evolutionary model in progress!

- Formulation of complexity that does not require paradigmatic treatment?
  - Derivational morphology, for example, is often seen as syntagmatic (but, e.g., Bonami & Strnadova 2016).

Thank You!

Questions?

“It would be good to return some emphasis within NLP to cognitive and scientific investigation of language rather than almost exclusively using an engineering model of research.” (Manning, 2016)