

An analogical approach to the typology of inflectional complexity

Matías Guzmán Naranjo & Olivier Bonami

09.2022

Complexity in inflectional morphology

We follow Ackerman and Malouf (2013) in distinguishing two types of complexity in inflectional morphology:

- I complexity: the difficulty that a paradigmatic system poses for language users
- E complexity: the number of morphosyntactic distinctions that languages make and the strategies employed to encode them

We want to explore both types of complexity cross-linguistically.

Measuring E-complexity

Some intuitions for measuring E-complexity are:

- number of morphosyntactic distinctions (paradigm size)
(Lupyan and Dale, 2010)
- number of morphs
- amount of material expressing a morphosyntactic contrast

Most of these metrics require arbitrary decisions to be made and run into issues (e.g. segmentation problem).

A Word and Paradigm view

To sidestep the segmentation problem, and other issues related to finding and defining morphs, morphemes, stems, etc. we adopt a Word and Paradigm perspective.

We won't try to understand markers and forms in isolation, but contrasts between two cells in a paradigm.

Instead of segmenting:

- *cant-a-ba-mos* (Sp., 'sing', 1.PL.IMP)
 - ▶ *cant*: stem
 - ▶ *a*: theme vowel
 - ▶ *ba*: IMP
 - ▶ *mos*: 1.PL

We look at contrasts between different cells:

cell 1	form 1	from 2	cell 2	pattern
1.PL.IMP	cantaba <i>mos</i>	cantaba	1.SG.IMP	→ X <i>mos</i> ⇔ X
1.PL.IMP	cantaba <i>mos</i>	cantabas	2.SG.IMP	→ X <i>moY</i> ⇔ XY

Proportional analogies

We use a new framework to express proportions:

- Named variables with matching potential
- Segments

For the alternations:

- $canta :: canta\textcolor{red}{ba}$
 $[<X1,*> \rightleftharpoons <X1,*>ba]$
- $\textcolor{red}{carta} :: \text{ca}\textcolor{red}{tra}$
 $[<X1,*><X2,1><X3,1><X4,1>, \rightleftharpoons <X1,*><X3,1><X2,1><X4,1>]$

The alternation for a cell pair is the *class* we will use for the predictions.

Measuring E-complexity in analogical proportions

Using these proportions, it does not make sense to count morphs, but there is one aspect of the system we can use:

We can count the number of positions filled by contrastive sequences. We will call this **fragmentation** (Bonami and Beniamine, 2021)

- $\langle X1, * \rangle_o \Leftarrow \langle X1, * \rangle_a : 2 \text{ positions}$
- $\langle X1, * \rangle_o \langle X1, 2 \rangle_e \Leftarrow o \langle X, 1 \rangle_a \langle X1, 2 \rangle_s : 5 \text{ positions}$

Measuring I-complexity

There are roughly two proposals:

- Entropy-based approaches: measure the information one cell (or pair of cells) provides about a different cell in a paradigm (Ackerman and Malouf, 2013, 2016; Bonami and Beniamine, 2016).
- Accuracy-based approaches: calculate the accuracy of predicting one cell from another cell (Guzmán Naranjo, 2020).

We use an accuracy-based approach: For a cell form we need to predict its *class* (alternation pattern).

Measuring I Complexity: Analogical classification

The idea of analogical classification is simple:

- Words that look alike behave alike

This has been applied to gender assignment, inflection class assignment, derivation class assignment, etc.

There are many approaches to doing this:

- Analogical Modelling
- Neural Networks
- Boosting Trees
- etc.

But these can be slow. We use a *kNN* classifier with modified Levenshtein distances.

Measuring I Complexity: Calculating phonological distances

We use an edge-sensitive Levenshtein distance.

- $D(\text{casa}, \text{masa}) = 0.25$
- $D(\text{asac}, \text{asam}) = 1$

Measuring I Complexity: Classification based on phonological distance

We create a matrix of phonological distance

	lexeme 1	lexeme 2	lexeme 3	lexeme 4	lexeme 5
lexeme 1
lexeme 2
lexeme 3
lexeme 4
lexeme 5

And then simply run a k NN classifier (with $k = 5$).

The *accuracy* of the classification is the *complexity* of the cell pair.

The *accuracy* across all cell pairs is the *complexity* of the system.

We do this for all cell-pairs in both directions: $\text{cell-1} \Leftarrow\Rightarrow \text{cell-2}$

An example: 2sg.pres.ind → 1sg.pres.ind In Spanish

First, we build the alternations for all pairs:

1SG.PRES.IND	2SG.PRES.IND	alternation
toco	tocas	$\langle X1, * \rangle o \Leftarrow \langle X1, * \rangle as$
como	comes	$\langle X1, * \rangle o \Leftarrow \langle X1, * \rangle es$
barro	barres	$\langle X1, * \rangle o \Leftarrow \langle X1, * \rangle es$
tomo	tomas	$\langle X1, * \rangle o \Leftarrow \langle X1, * \rangle as$

We then calculate the phonological distances for all 2SG.PRES.IND:

	tocas	comes	barres	tomas
tocas	0.00	1.03	1.45	0.33
comes	1.03	0.00	0.95	0.70
barres	1.45	0.95	0.00	1.45
tomas	0.33	0.70	1.45	0.00

We first remove incompatible patterns ($\langle X1, * \rangle o \Leftarrow \langle X1, * \rangle as$ is not compatible with *comes*)

We then predict base on the nearest neighbour (with $k = 1$). In this case it is trivial and we get Acc = 1.

The datasets

We used datasets from the Unimorph as of 01.2021 (Kirov et al., 2018)

Additionally, we included some other datasets we had access to:

- Russian (Guzmán Naranjo, 2020)
- Kasem (Guzmán Naranjo, 2019)
- French verbs (Beniamine and Guzmán Naranjo, 2021)
- Arabic nouns (ibid.)
- Portuguese verbs (ibid.)
- English verbs (CELEX)
- Latin Nouns (ibid.)
- Latin Verbs (ibid.)
- Navajo Verbs (ibid.)
- Yaitepec Verbs (ibid.)
- Zenzotepec Verbs (ibid.)

A total of 137 datasets including verbs, nouns and adjectives, across 71 languages

Results

Fragmentation

Overall fragmentation by language I

lang	nouns					verbs					adjectives				
	200	500	1000	2000	5000	200	500	1000	2000	5000	200	500	1000	2000	5000
ady	1.76	1.76	1.8	1.8							1.76				
ang	3.17	2.88	3.15			3.14	3.45	3.55							
ara						4.46	4.78	4.79	4.79						
aze	2.77	2.82													
bak	2.04	2.03	1.99	1.99											
bel	2.97	3.06	3.06				3.13				2.19				
bul											3.42	3.43			
cat						2.56	2.63	2.85	2.8						
ces	2	2.09	2.15	2.35	2.38	2.71	2.93								
ckb	3.05					6.08									
crh	2.08	2.05	2.05	2.06											
cym						2.59									
dan	2.69	2.56	2.77	3.06	3.12	1.29									
deu	2.22	2.28	2.39	2.36	2.42	2.61	2.69	2.72	2.77	2.81					
dsb											1.96				
ell	3.33	3.5	3.48	3.42	3.28	6.05	6.16	6.14	6.11		1.94	1.94	1.96	2.06	2.05
eng						2.39	2.27	2.66	2.83	2.91					
est	2.47	2.62	2.63			4.4									
fao	2.67	2.76	2.95	3.01	3	2.75	2.76	2.76			2.32	2.34			
fas						3.86	3.93								
fin	2.51	2.53	2.63	2.66	2.62	4.36	4.48	4.55	4.58	4.64	2.4	2.45	2.48	2.5	2.51
fre						1.83	1.89	1.89	1.94	1.94					
frm						2.61	2.67	2.68							
fro						2.85	3	3.03	3.09						
fur						2.22									
gal						2.61	2.69								

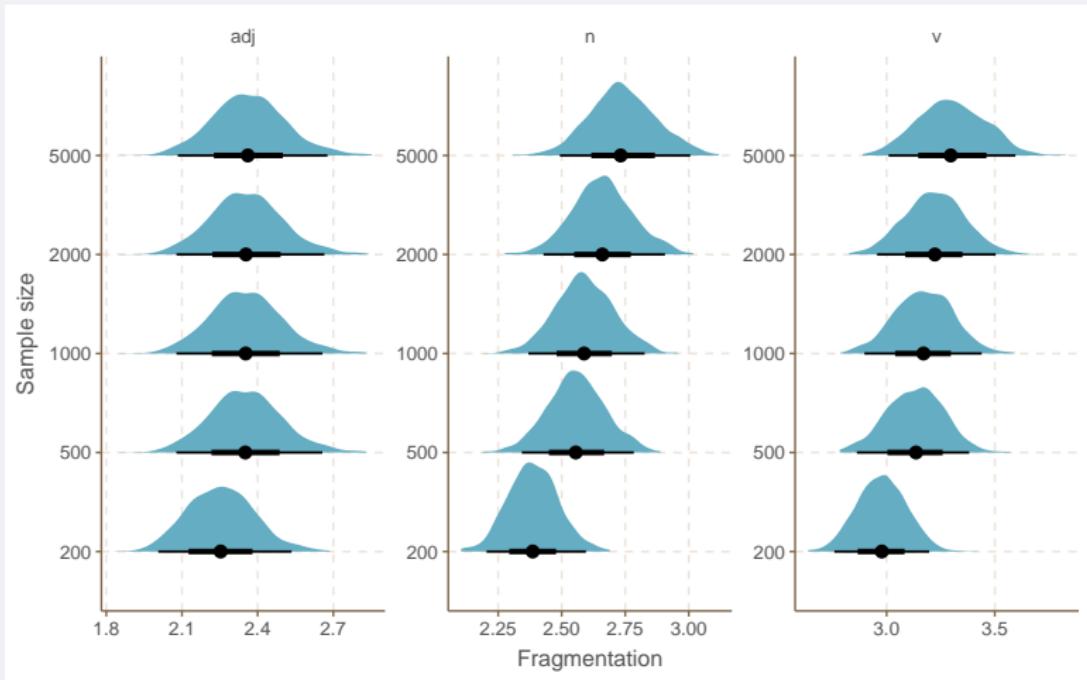
Quantifying variation across sample sizes

We can quantify the uncertainty about how sample size affects these values with Bayesian lognormal models:

```
mean_fragmentation ~ mo(sample_size) * pos + (1 +
mo(sample_size) || lang/pos)
```

(we fitted all our models with brms (Bürkner, 2017, 2018) and Stan (Carpenter et al., 2017))

Fragmentation



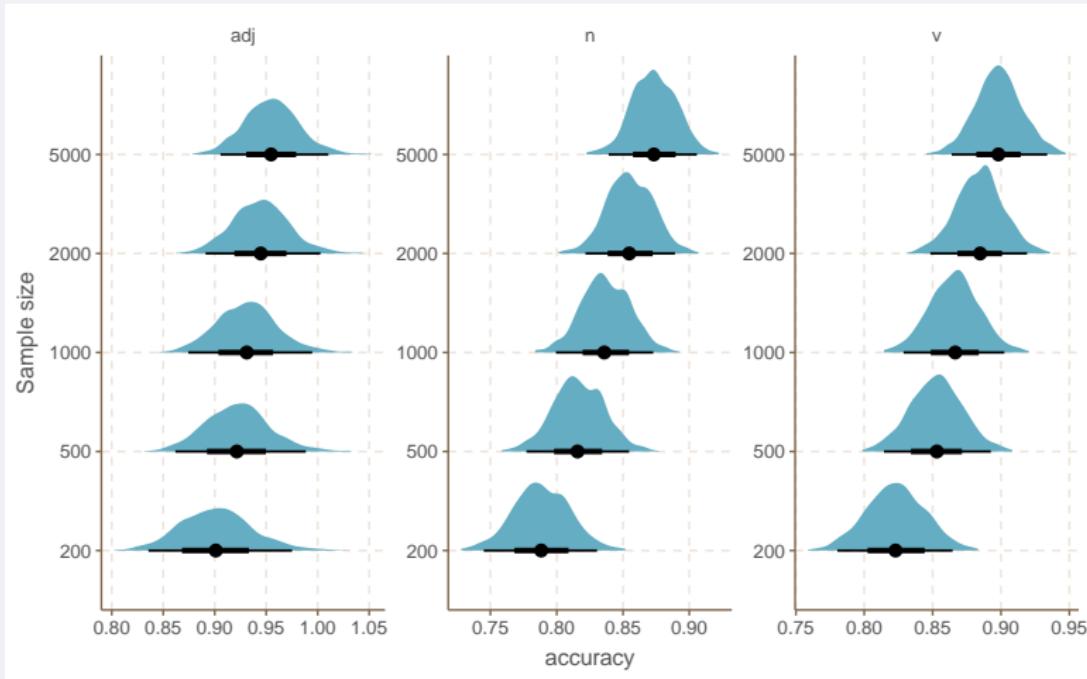
Results

I-complexity: Mean accuracy by language

Overall accuracy by language I

lang	nouns					verbs					adjectives				
	200	500	1000	2000	5000	200	500	1000	2000	5000	200	500	1000	2000	5000
ady	0.97	0.97	0.97	0.98							0.97				
ang	0.64	0.68	0.69			0.76	0.77	0.78							
aze	0.94	0.94													
bak	0.96	0.98	0.98	0.98											
bel	0.61	0.66	0.66				0.73				0.99				
bul											0.95	0.96			
cat						0.91	0.93	0.94	0.95						
ces	0.76	0.78	0.8	0.82	0.83	0.98	0.98								
ckb	0.97					0.7									
crh	0.96	0.97	0.97	0.98											
cym						0.86									
dan	0.75	0.77	0.8	0.82	0.84	1									
deu	0.6	0.68	0.7	0.73	0.79	0.77	0.81	0.81	0.85	0.86					
dsb											0.99				
ell	0.69	0.74	0.75	0.78	0.81	0.71	0.77	0.8	0.8		0.98	0.98	0.98	0.98	0.98
eng						0.92	0.95	0.95	0.96	0.97					
est	0.89	0.91	0.92			0.83									
fao	0.7	0.72	0.75	0.79	0.8	0.71	0.76	0.77			0.89	0.91			
fas						0.86	0.86								
fin	0.75	0.79	0.84	0.87	0.9	0.85	0.91	0.93	0.94	0.96	0.9	0.92	0.93	0.95	0.96
fre						0.9	0.91	0.93	0.94	0.96					
frm						0.95	0.96	0.96							
fro						0.83	0.85	0.88	0.9						
fur						0.92									

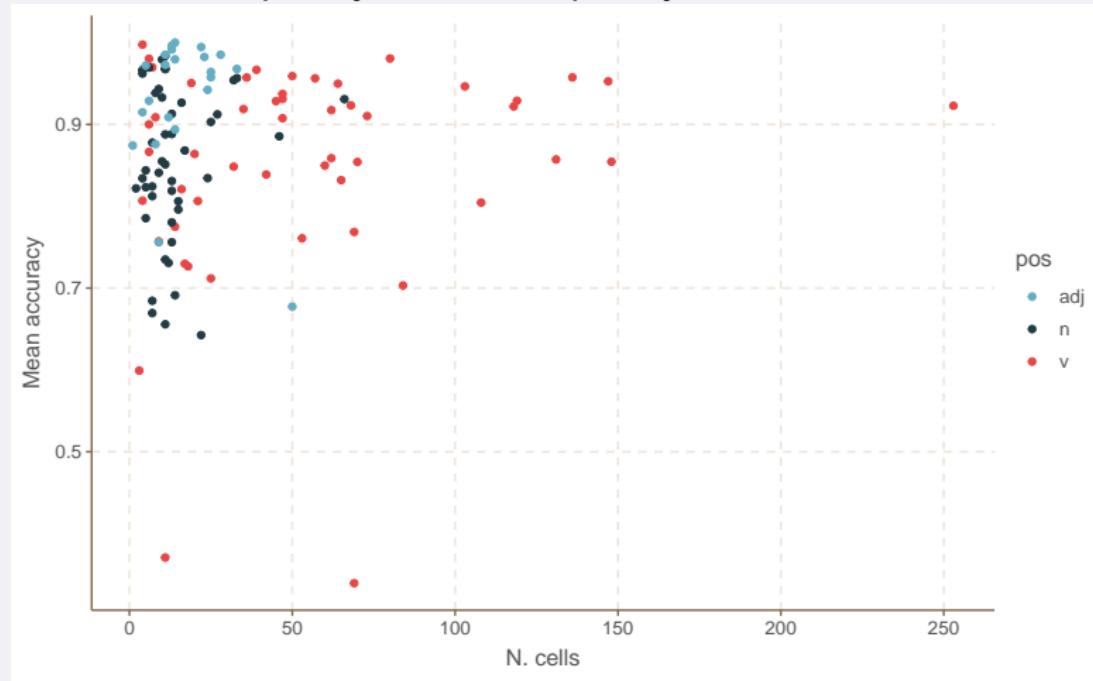
Sample size effects



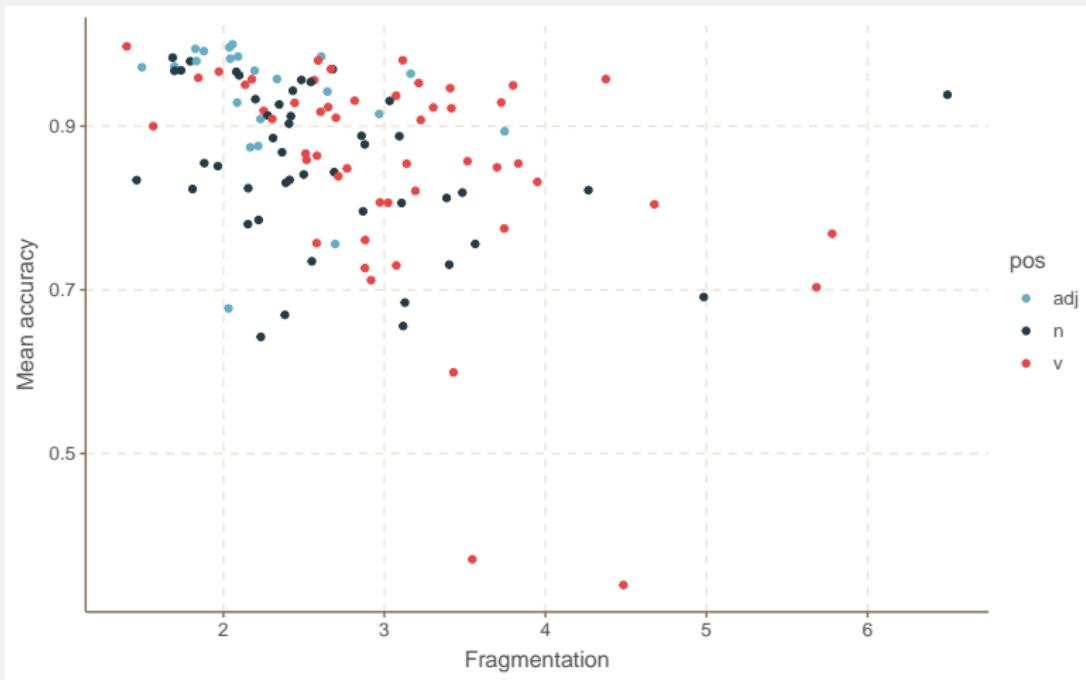
E-complexity vs I-complexity

Mean E-complexity vs mean I-complexity

A claim found in Cotterell et al. (2019) is that there is a tradeoff between I-complexity and E-complexity.



Mean E-complexity vs mean I-complexity



E-complexity vs I-complexity in finer detail

However, for each language, different cells have different accuracies and different levels of fragmentation.

Instead of looking at averages across languages, we look at the mean accuracy for cell→cell pair vs its mean fragmentation:

correct | total ~ fragmt. + (1 + fragmt. | language)

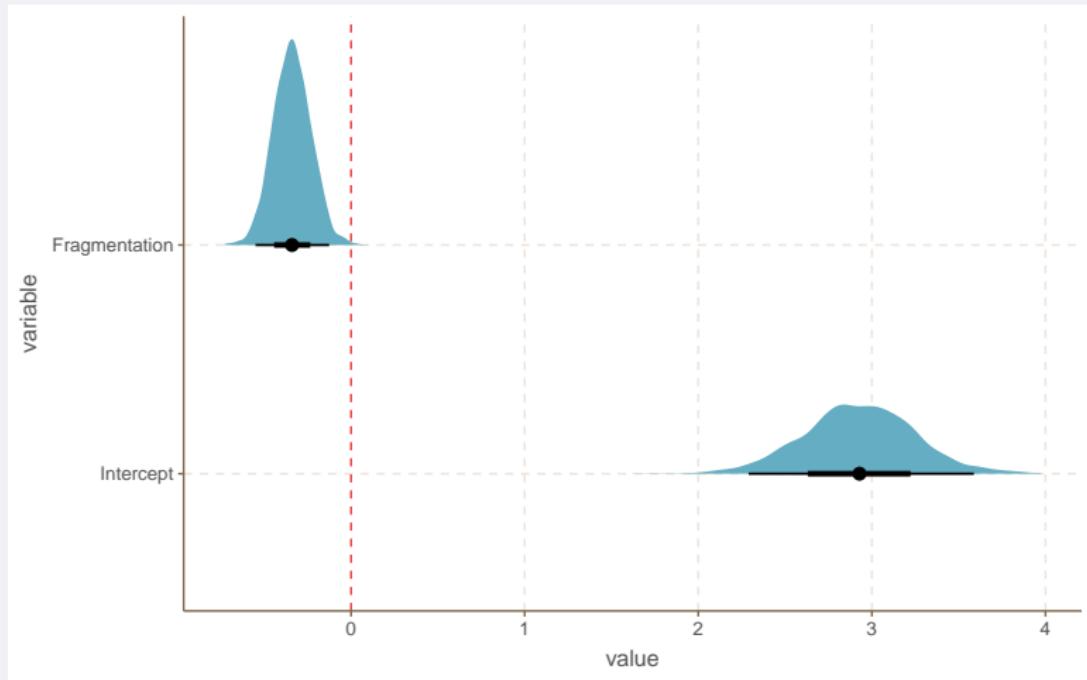
This should tell us whether cells with patterns like

- $\langle X1, * \rangle a \Leftarrow \langle X1, * \rangle o$

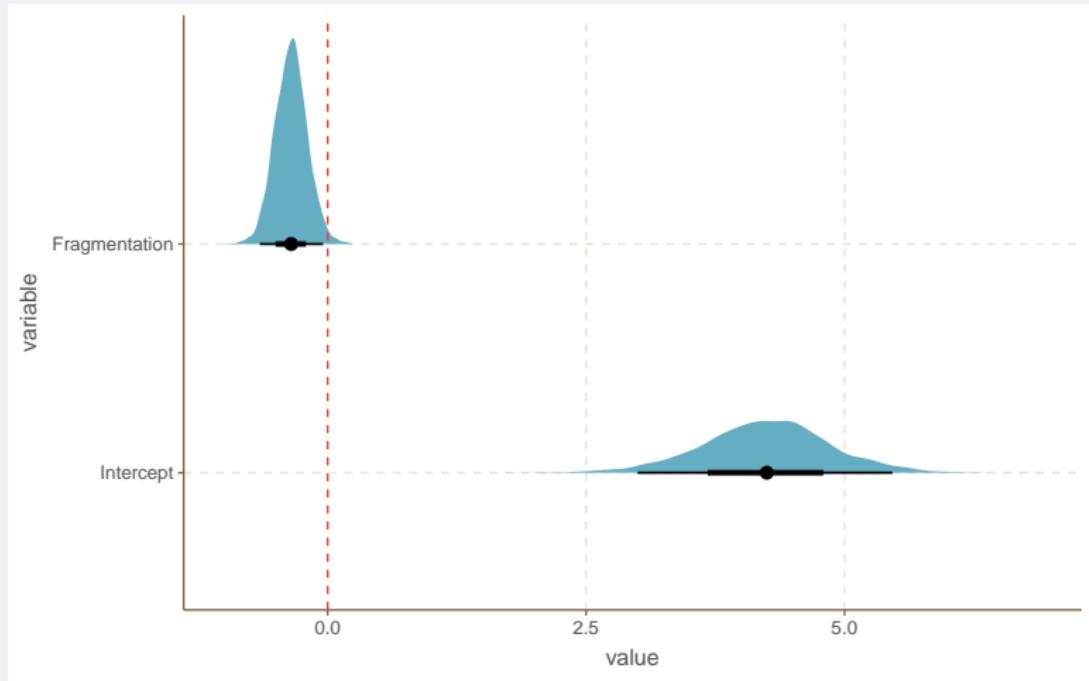
are overall easier to predict than cells with patterns like

- $u \langle X1, * \rangle a \langle X2, 1 \rangle s \Leftarrow i \langle X1, * \rangle o \langle X2, 1 \rangle l$

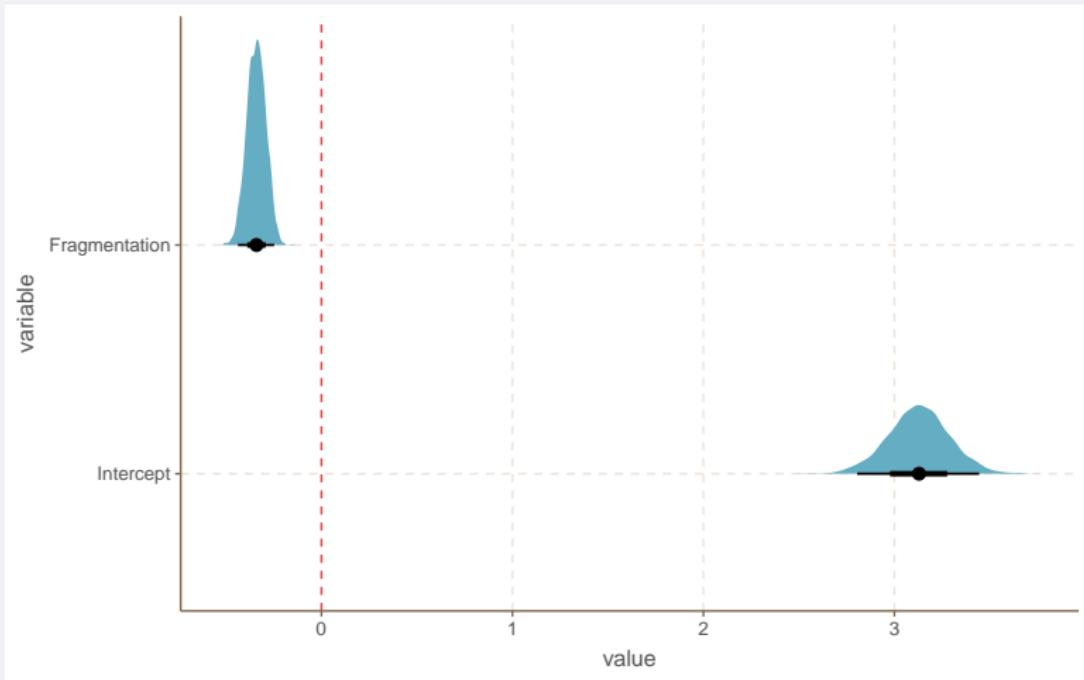
Nouns: coefficients population effects



Adjectives: coefficients population effects



Verbs: coefficients population effects



What about morphemes?

We want to explore what happens from the morpheme-based perspective, however:

- It is unclear how to produce morpheme segmentation automatically
- Counting morphemes is a difficult, theory dependent task

Here we explore one of many possibilities: Morfessor

Morfessor

We trained Morfessor on bible corpora.

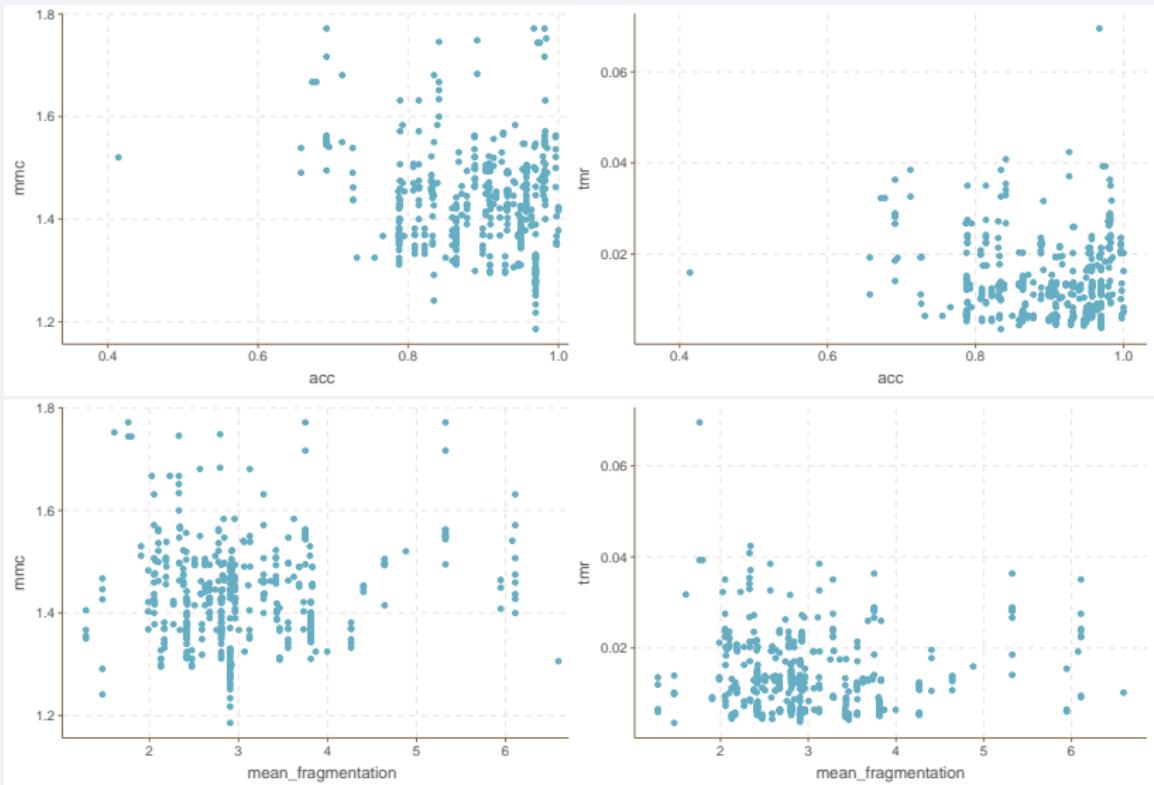
Index of synthesys vs I- and E-complexity

From the segmentations provided by morfessor, we extract two metrics:

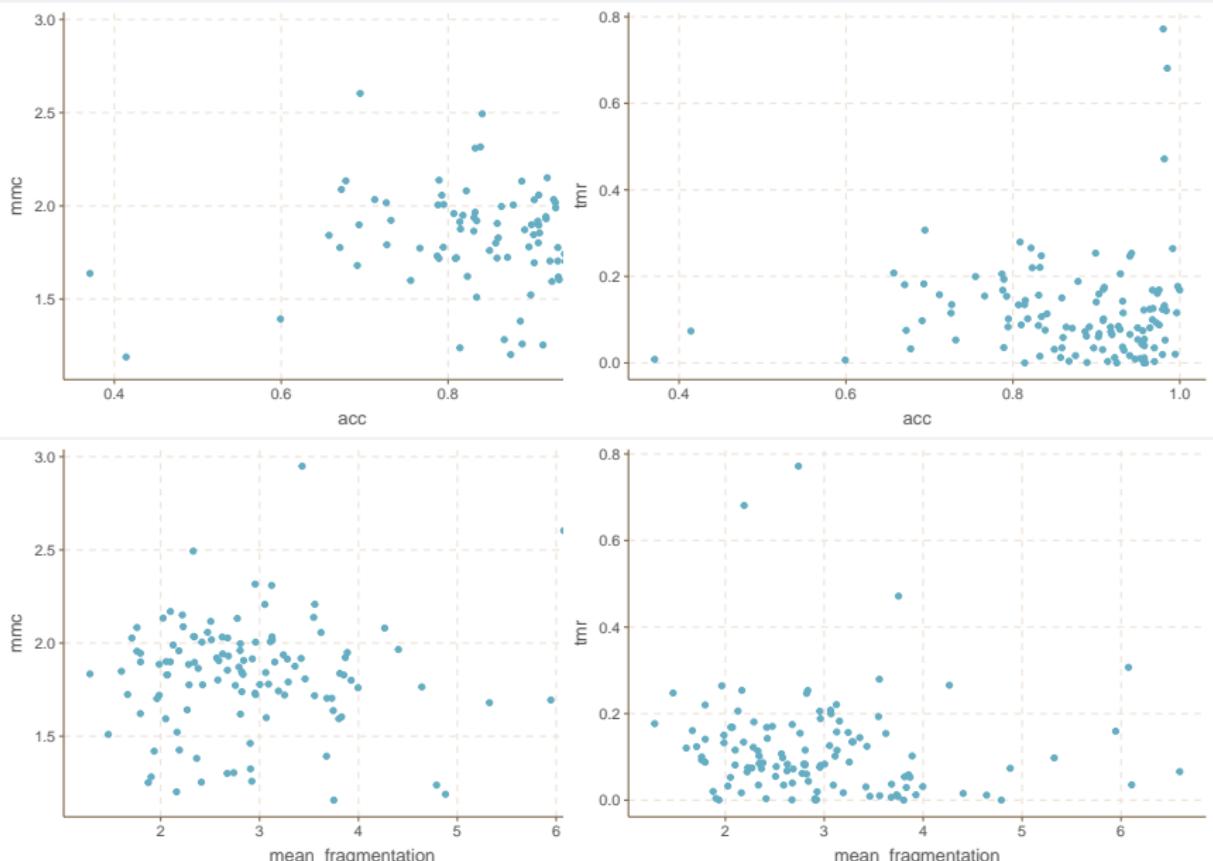
- mean morpheme count per word (mmc)
- total morpheme ratio (tmr)

Overall, we see no clear correlations

Morfessor on Bibles



Morfessor on Paradigms



Conclusion I

In this presentation we have presented:

- A W&P approach to morphological typology
- A formalism for proportional analogies (+ efficient computational implementation)
- A simple technique for analogical classification

Conclusion II

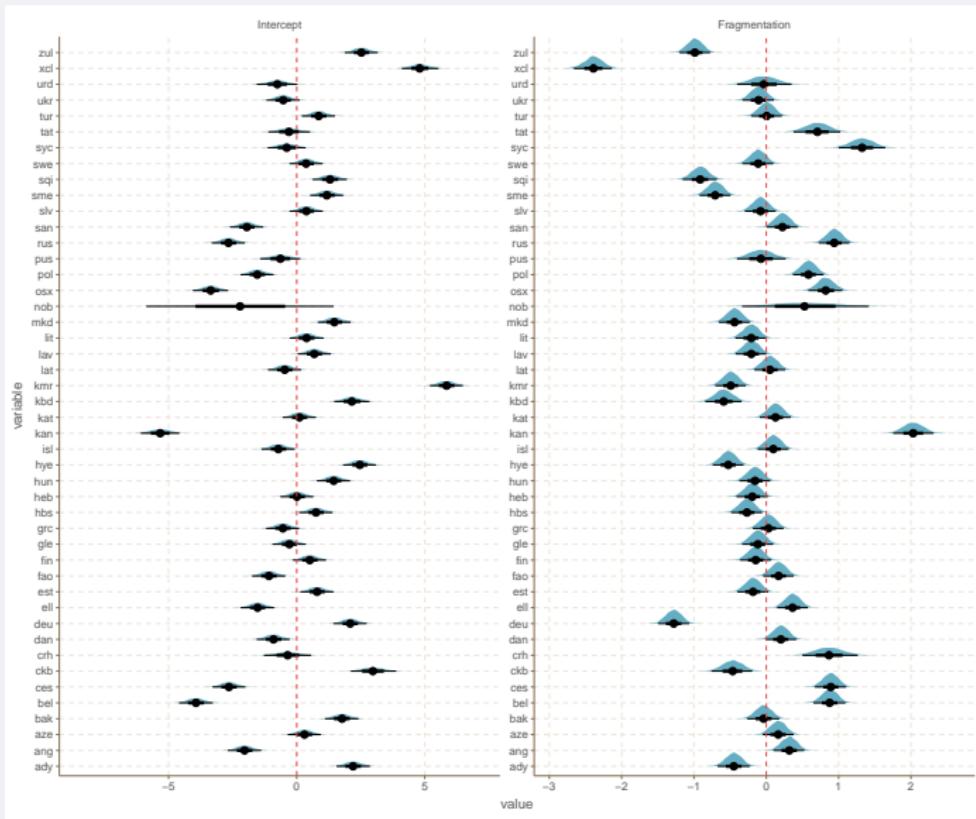
We find that:

- We can predict most inflection classes very well
- Small sample sizes mildly underestimate E-complexity
- Small sample sizes *seriously* overestimate I-complexity
- Most languages have relatively low I- and E-complexity
- **Number of cells does not correlate with I-complexity**
- **High fragmentation correlates higher I-complexity**
- Automated morpheme-based approaches do not correlate in any way with either E- or I-complexity (as defined here)

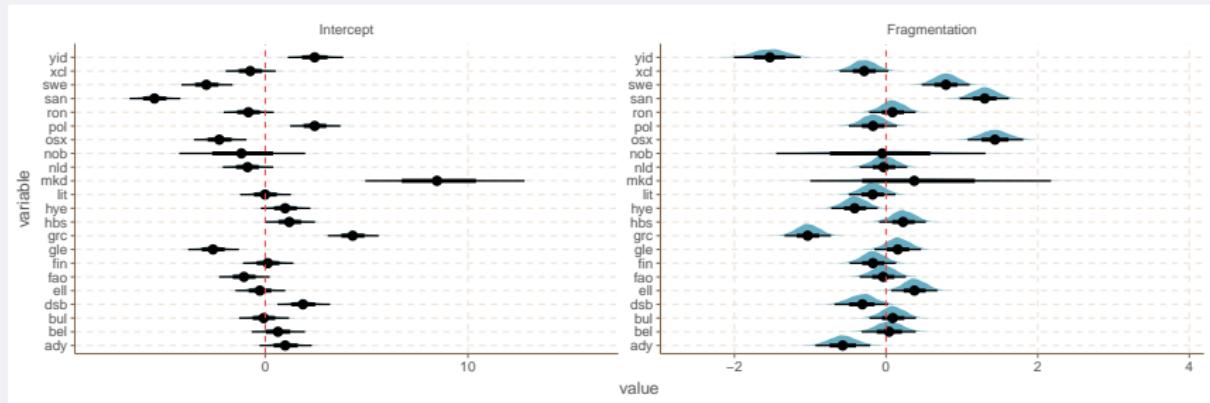
Thank you!

- Ackerman, Farrell and Robert Malouf (2013). "Morphological Organization: The Low Conditional Entropy Conjecture". In: *Language* 89.3, pp. 429–464.
- (2016). "Word and Pattern Morphology: An Information-Theoretic Approach". In: *Word Structure* 9.2, pp. 125–131.
- Beniamine, Sacha and Matías Guzmán Naranjo (2021). "Multiple Alignments of Inflectional Paradigms". In: *"Multiple Alignments of Inflectional Paradigms"*. "Multiple Alignments of Inflectional Paradigms". Vol. 4.
- Bonami, Olivier and Sacha Beniamine (2016). "Joint Predictiveness in Inflectional Paradigms". In: *Word Structure* 9.2, pp. 156–182.
- (2021). "Leaving the stem by itself". In: *All Things Morphology: Its independence and its interfaces*. Ed. by Sedigheh Moradi et al. Vol. 353. John Benjamins Publishing Company, pp. 81–98.
- Bürkner, Paul-Christian (2017). "Brms: An R Package for Bayesian Multilevel Models Using Stan". In: *Journal of Statistical Software* 80.1, pp. 1–28.
- (2018). "Advanced Bayesian Multilevel Modeling with the R Package Brms". In: *The R Journal* 10.1, pp. 395–411.
- Carpenter, Bob et al. (2017). "Stan: A Probabilistic Programming Language". In: *Journal of Statistical Software, Articles* 76.1, pp. 1–32.
- Cotterell, Ryan et al. (2019). "On the Complexity and Typology of Inflectional Morphological Systems". In: *Transactions of the Association for Computational Linguistics* 7, pp. 327–342.
- Guzmán Naranjo, Matías (2019). *Analogical Classification in Formal Grammar. Empirically Oriented Theoretical Morphology and Syntax*. Berlin: Language Science Press.
- (2020). "Analogy, Complexity and Predictability in the Russian Nominal Inflection System". In: *Morphology* 30, pp. 219–262.
- Kirov, Christo et al. (2018). "UniMorph 2.0: Universal Morphology". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. Paris, France: European Language Resources Association (ELRA).
- Lupyan, Gary and Rick Dale (Jan. 2010). "Language Structure Is Partly Determined by Social Structure". In: *PLOS ONE* 5.1, pp. 1–10.

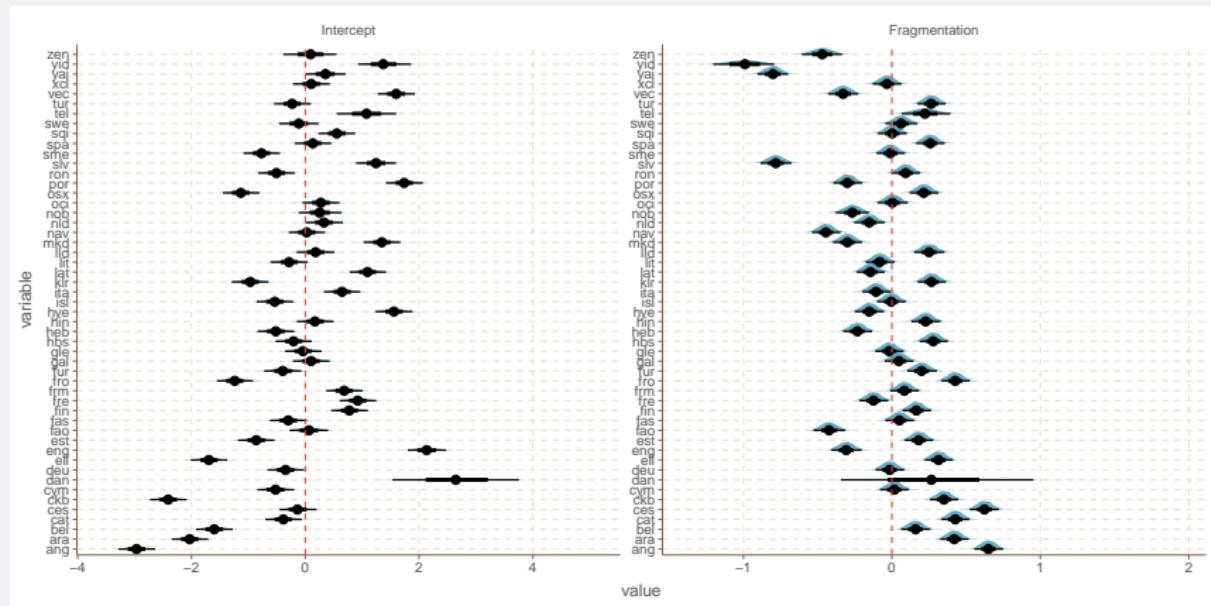
Nouns: coefficients group effects



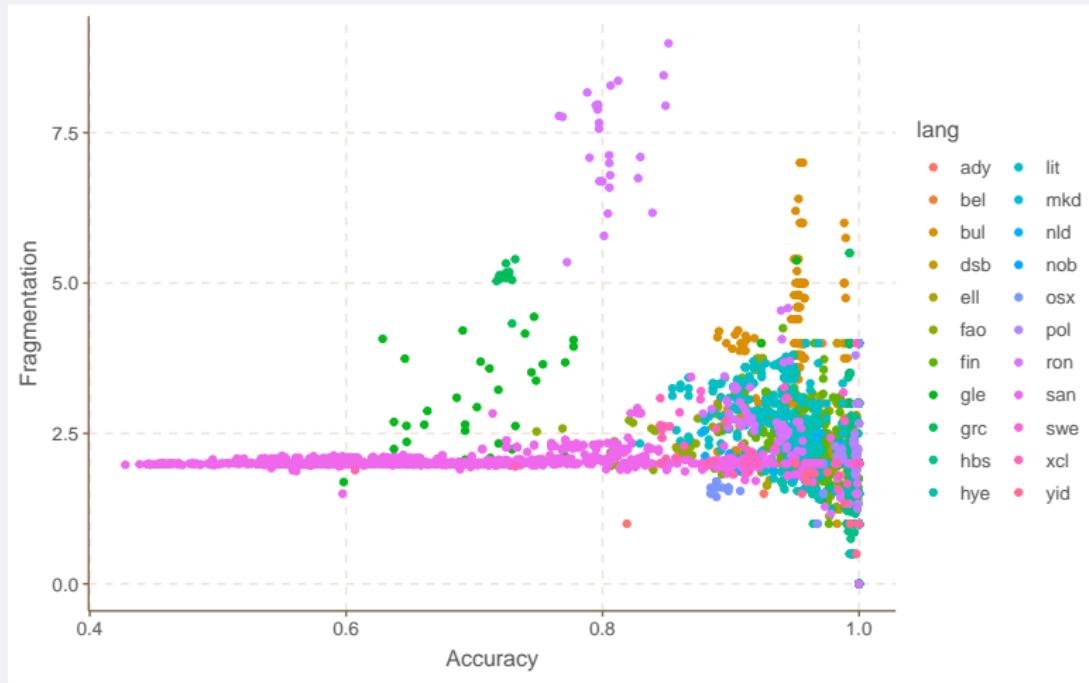
Adjectives: coefficients group effects



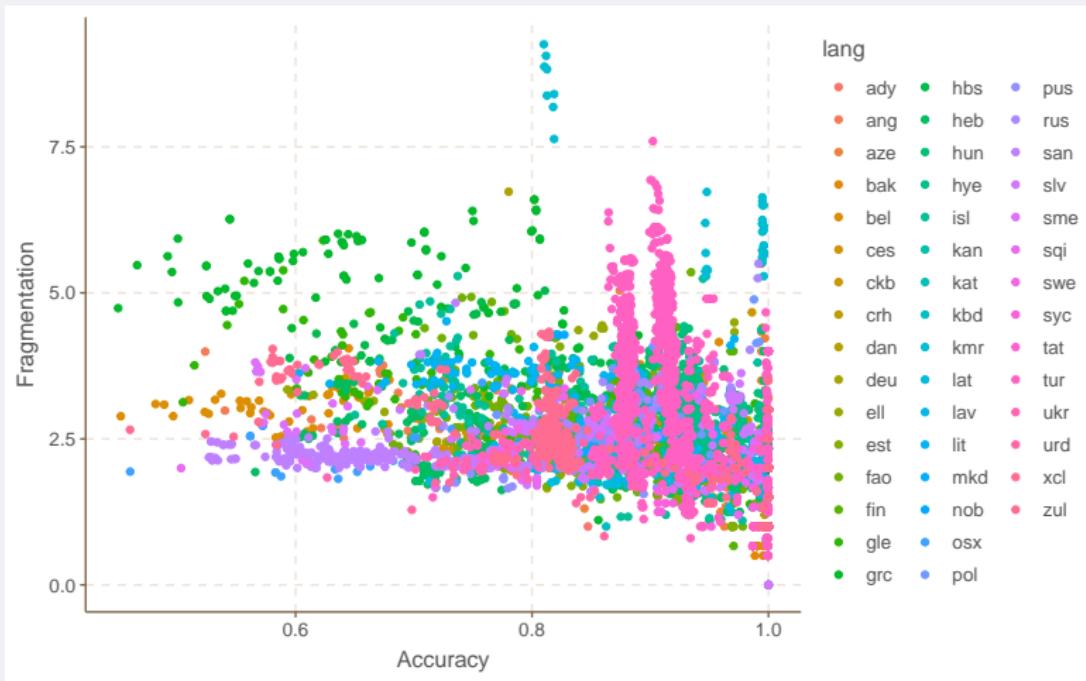
Verbs: coefficients group effects



Accuracy vs fragmentation adjectives



Accuracy vs fragmentation nouns



Accuracy vs fragmentation verbs

