What Quantifying Word Order Freedom Reveals about Dependency Corpora

(a work in progress)

Maja Buljan

LTG Seminar

2021-10-25

Outline

Introduction

- Futrell et al., 2015.
  Quantifying Word Order Freedom in Dependency Corpora
- syntactic annotation and numeric measures for linguistic analysis

- body of work in similar vein
  - Culbertson & Smolensky, 2012. *A Bayesian Model of Biases in Artificial Language Learning: The Case of a Word-Order Universal*
  - Gulordava et al., 2015. *Dependency Length Minimization Effects in Short Spans: A Large-Scale Analysis of Adjective Placement in Complex Noun Phrases*
  - Gulordava & Merlo, 2016. *Multi-lingual Dependency Parsing Evaluation: a Large-scale Analysis of Word Order Properties using Artificial Data*
  - Gulordava, 2018. *Word order variation and dependency length minimisation: a cross-linguistic computational approach*
  - Levshina, 2019. *Token-Based Typology and Word Order Entropy: A Study Based on Universal Dependencies*

Motivation / RQ

- WOF and other measures; cross-lingual analysis, domains and sources
- but: does the measure reflect **language**, or **annotation**?
- **spoiler:** harder than expected

Background

- quantify linear order of words through unordered dependency graph
- using **conditional entropy**

$$H(X|C) = \sum_{c \in C} p_C(c) \sum_{x \in X} p_{x|c}(x|c) \log p_{X|C}(x|c)$$

- X – dependent variable
- C – conditioning variable

Background

Parameters TBD:

1. estimating H from joint counts of X and C,
2. information contained in X
3. information contained in C

Trade-offs:

- avoiding **data sparsity**
- retaining linguistic **interpretability**

Background

- computing entropy only on **local subtrees**
  (given content-head dependency; UD ✓)
- X: **linear order** of words
- C: unordered local subtree
  (dependency **relation type**, **POS tags** of heads and dependents)

Measures:

1. Relation Order Entropy (ROE)
2. Subject-Object Relation Order Entropy (SOE)
3. Head Direction Entropy (HDE)

Background

Universal Dependencies v1 → v2; changes in annotation guidelines

- **segmentation** and word-internal spaces
- NSUBJPASS **subsumed** under NSUBJ, with subtype option (NSUBJ:PASS)
- oblique nominals at clause level: NMOD → OBL
- coordination: **immediately succeeding** conjunct
- special relation for elliptical constructions
- **MWE**: revisions for particular subtypes
- and more!

Experimental Setup

- "parallel" treebanks; UDv1.4 and UDv2.8 (last/latest release)
- **matching** challenges:
  - tokenization, lemmatization, abbreviation
  - MWEs, complex names, etc.
  - no sentence ID standard before UDv2.0
  - sporadic expansion of treebanks
- 31 of 34 languages from original study
  (incomplete annotations for Bengali, Japanese, Telugu)
- compare **full** treebanks, and 10x **random samples** (10k sentences)

Experimental Setup

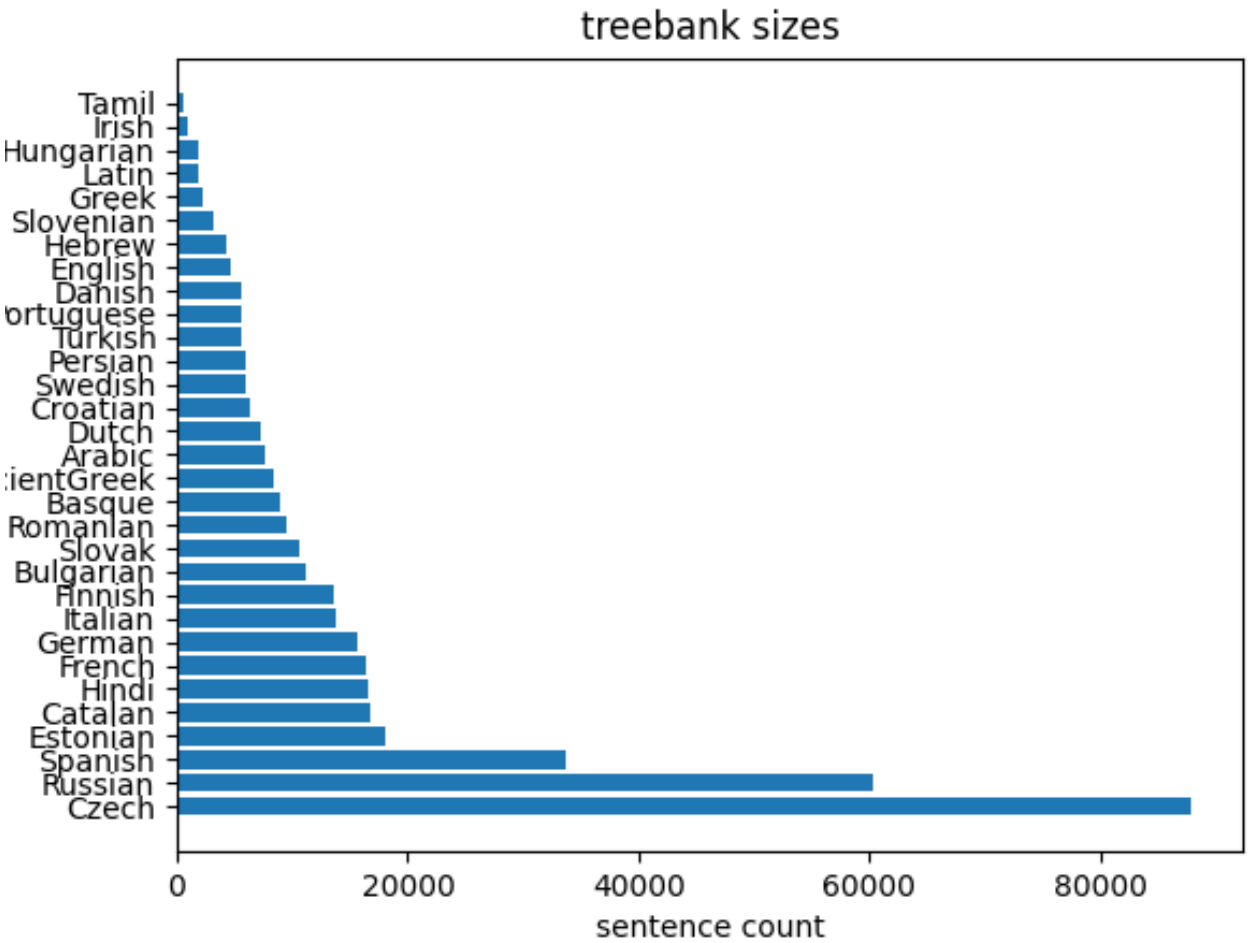
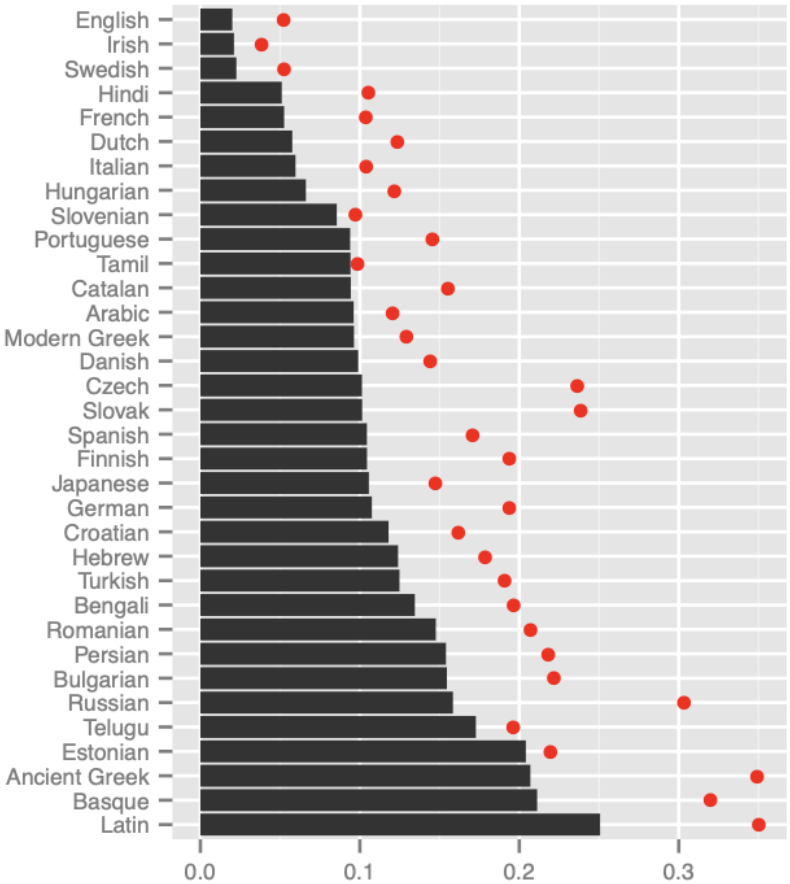


Figure: Treebank sizes per language, in number of sentences

Findings

1. original study vs. rerun
2. full treebanks vs. random samples
3. UD1 vs. UD2

Findings



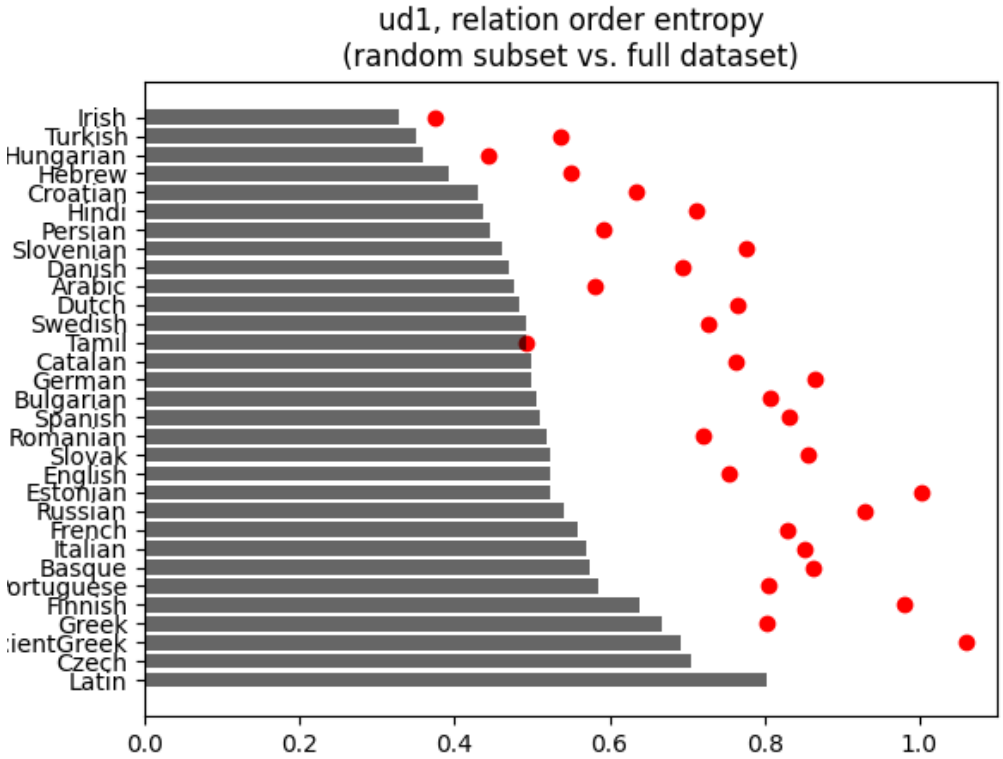
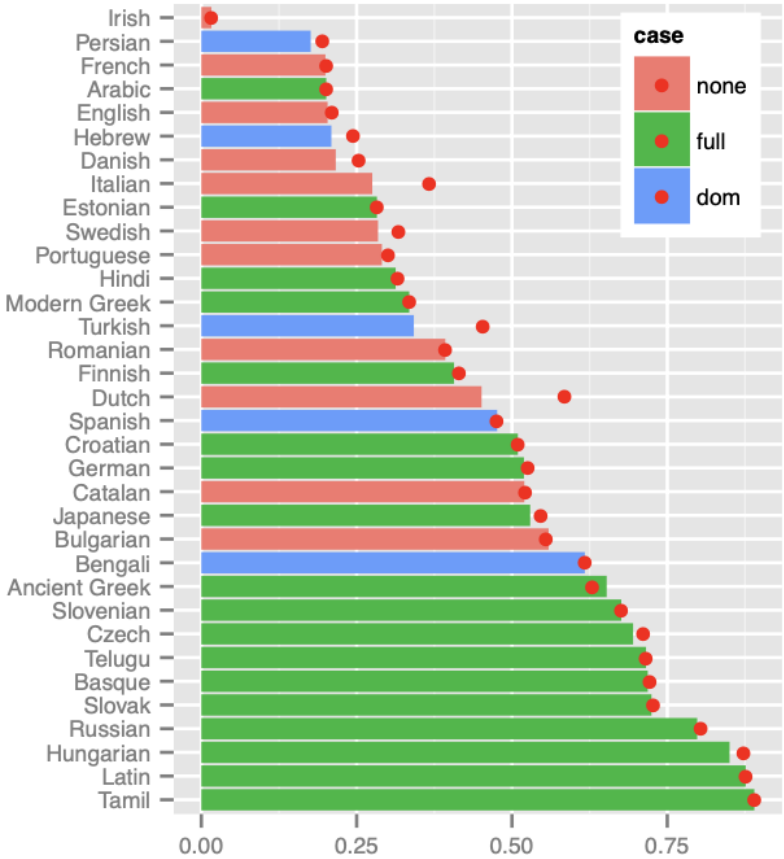


Figure: ROE; random sample vs. full treebank

Findings



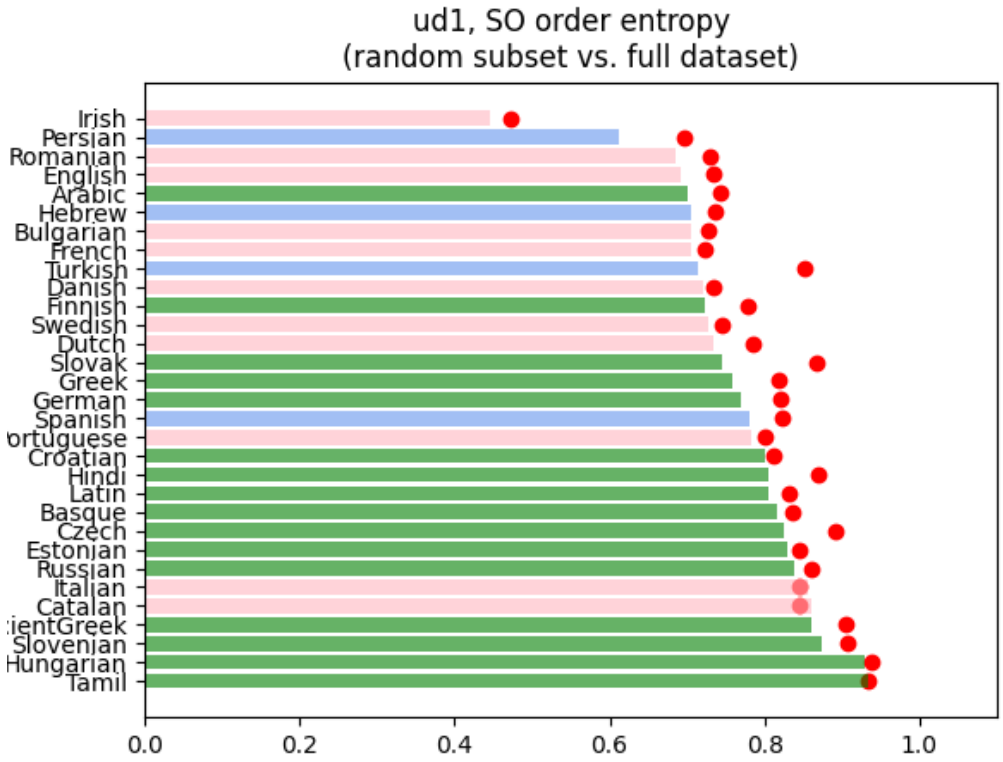


Figure: SOE; random sample vs. full treebank

Findings

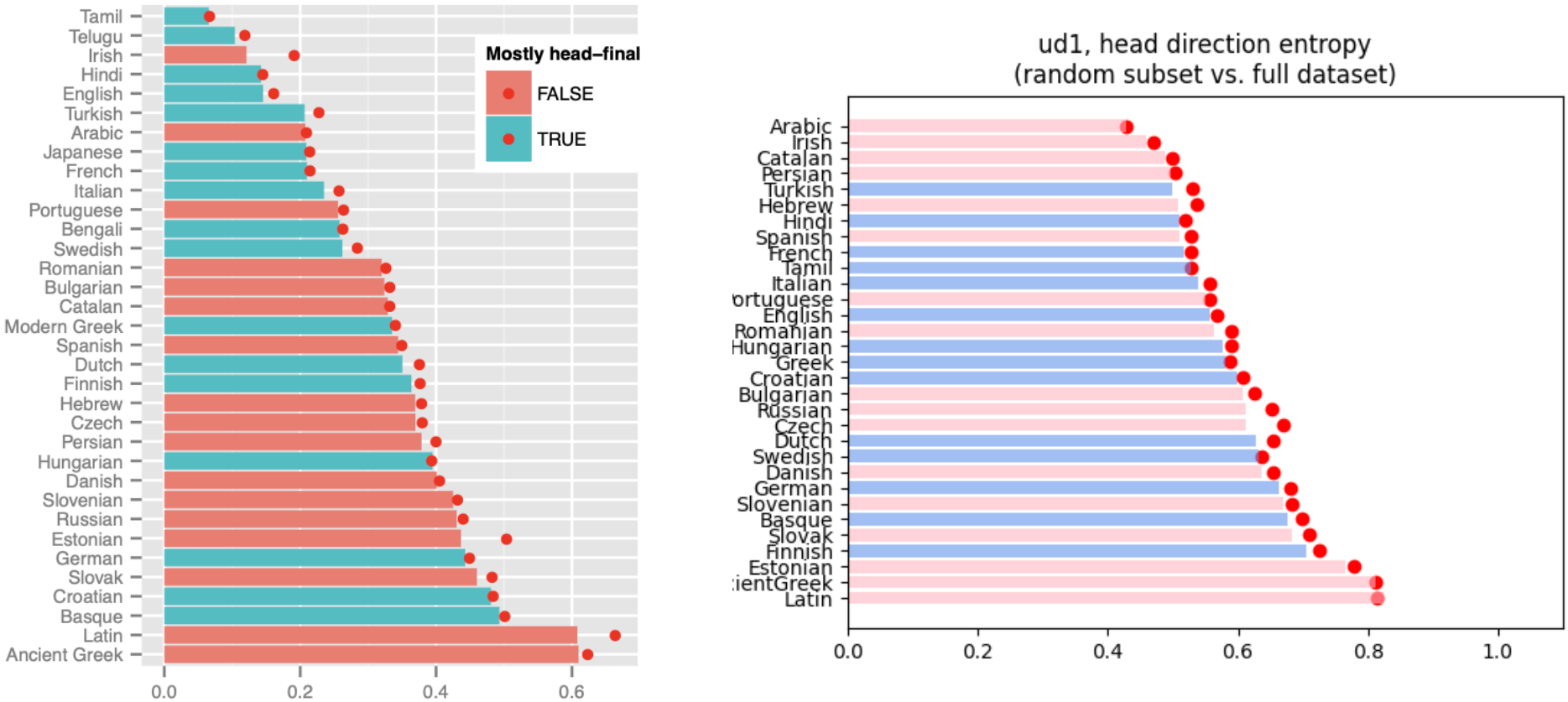


Figure: HDE; random sample vs. full treebank

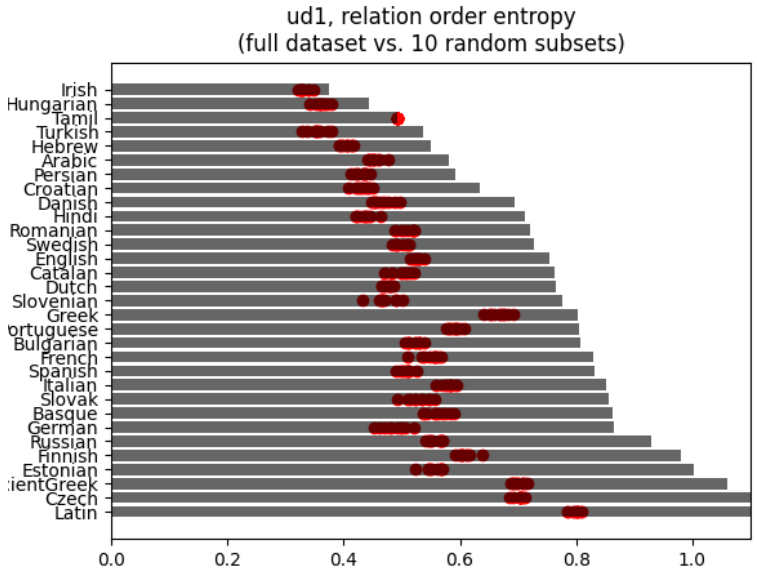
Findings

	full tb	sample
ROE	.328	.075
SOE	.103	.092
HDE	.099	.039

Table: Spearman ρ, original study vs. rerun (UD1 only)

	ori	ud1	ud2
ROE	.258	.244	.134
SOE	.576	.094	.260
HDE	.419	.339	.211

Table: Spearman ρ (averaged), full treebank vs. random sample



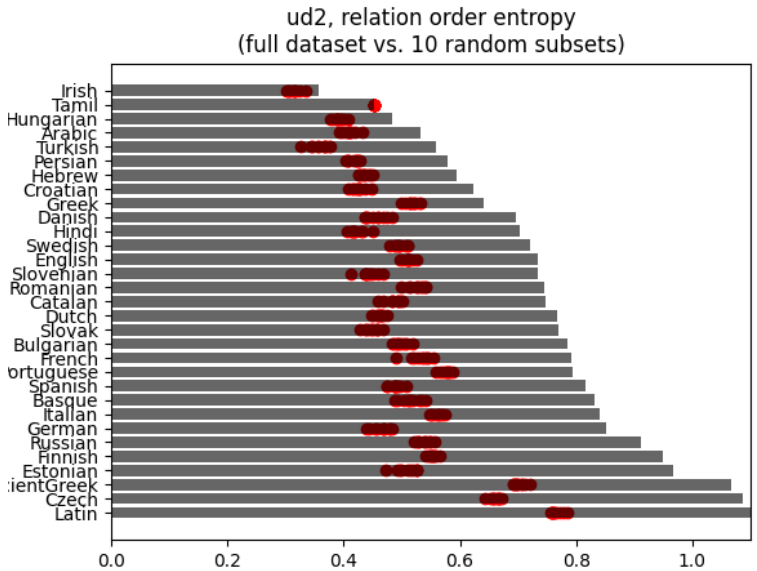


Figure: ROE; UD1 vs. UD2 (full treebank, 10 random samples)

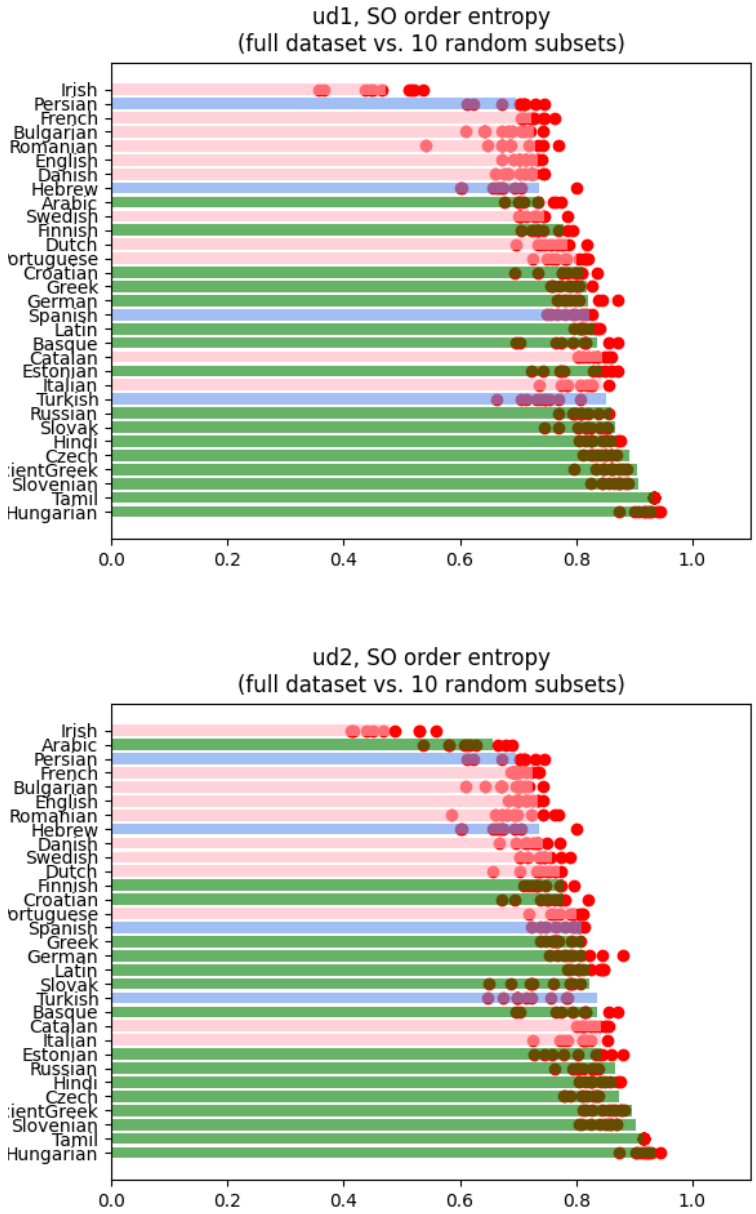


Figure: SOE; UD1 vs. UD2 (full treebank, 10 random samples)

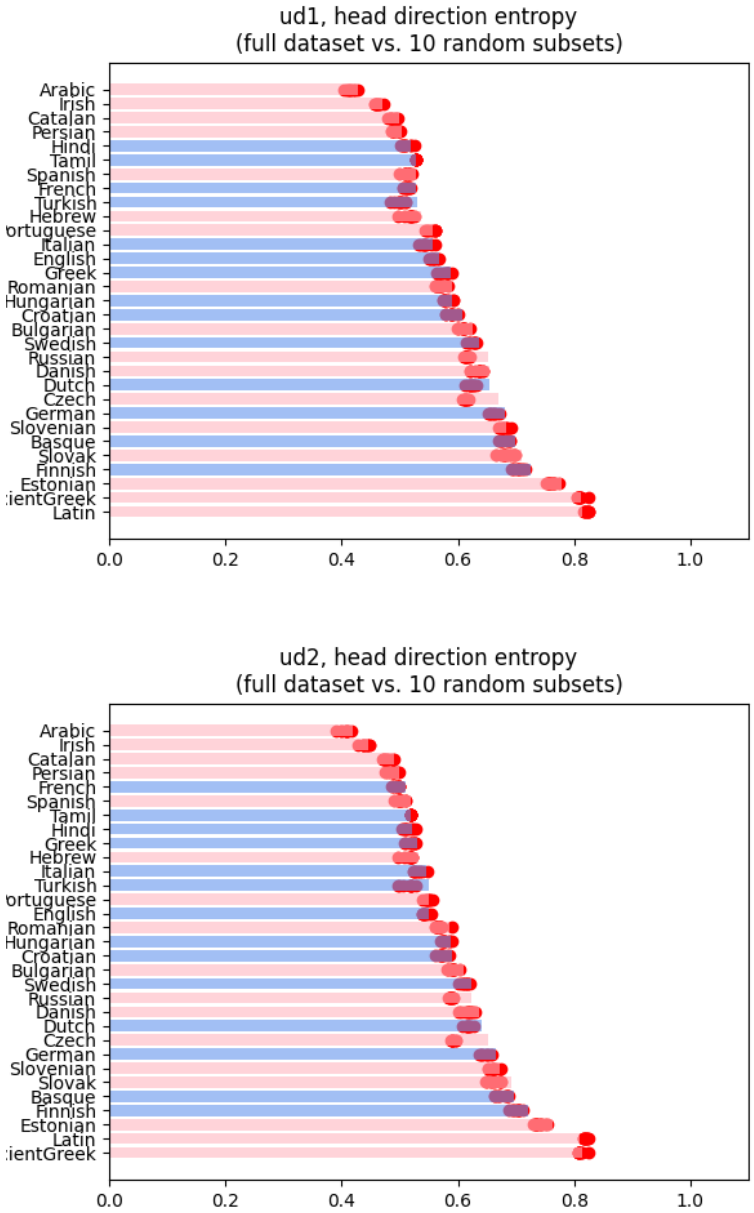


Figure: HDE; UD1 vs. UD2 (full treebank, 10 random samples)

Findings

	full tb	sample
ROE	.135	.156
SOE	.204	.112
HDE	.449	.293

Table: Spearman ρ (averaged), UD1 vs UD2

And in case you're wondering...

	full tb	sample	
ROE	-0.10	.000	ud1
	-0.19	-0.02	ud2
SOE	.198	-0.28	ud1
	.180	-0.20	ud2
HDE	.026	-0.14	ud1
	.414	-0.16	ud2

Table: Spearman ρ, treebank size rankings vs. WOE

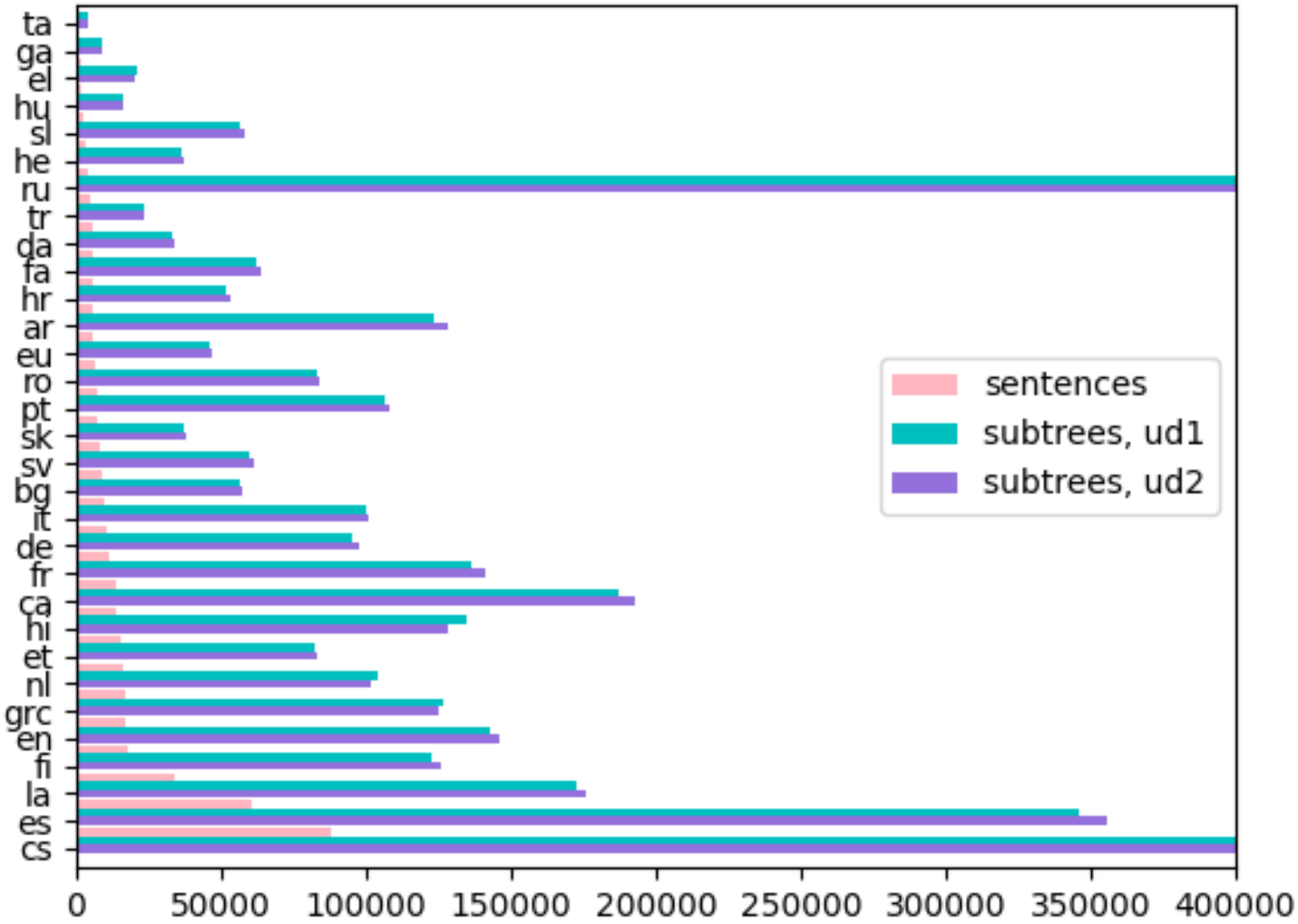


Figure: Subtree count

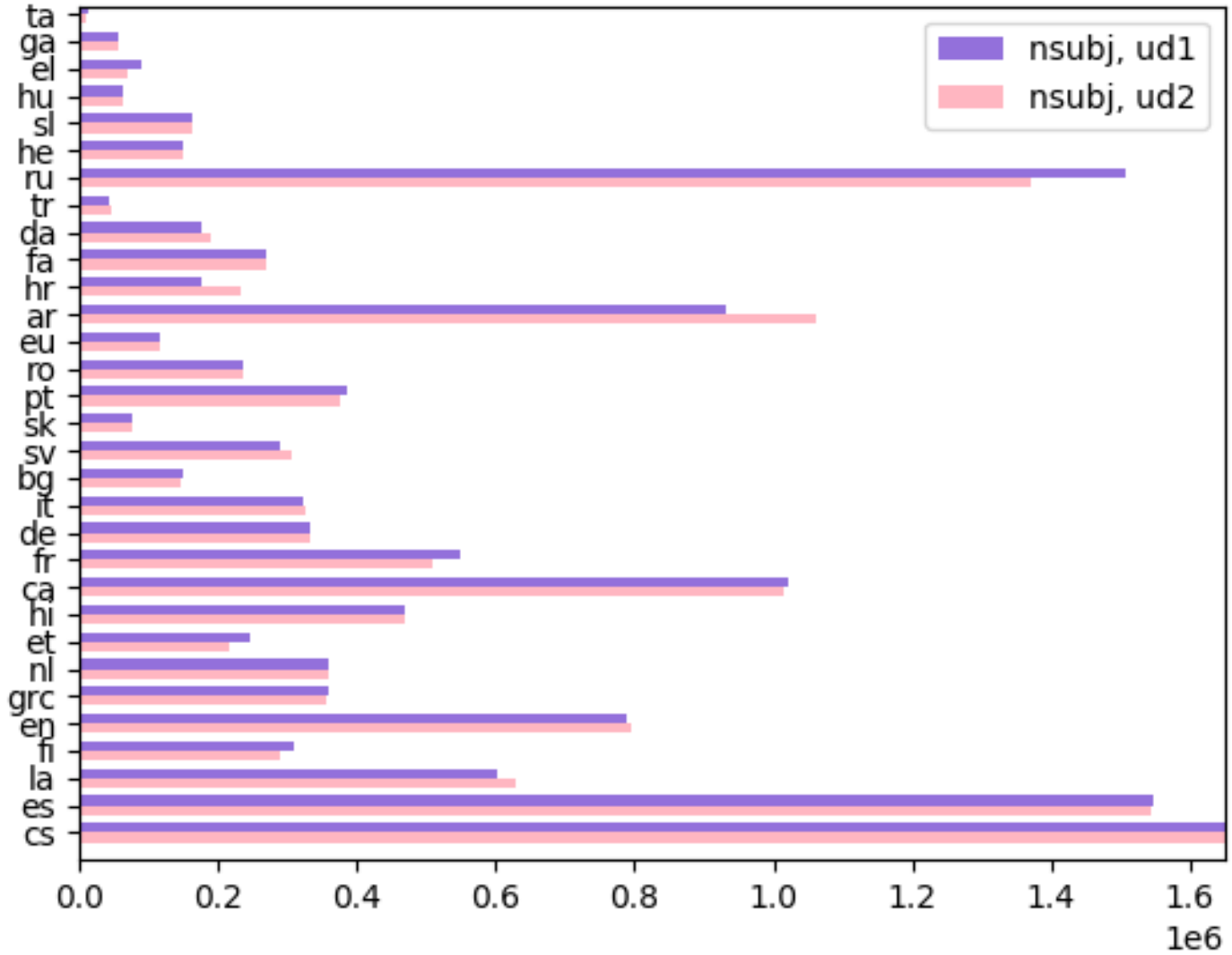


Figure: NSUBJ relation heads

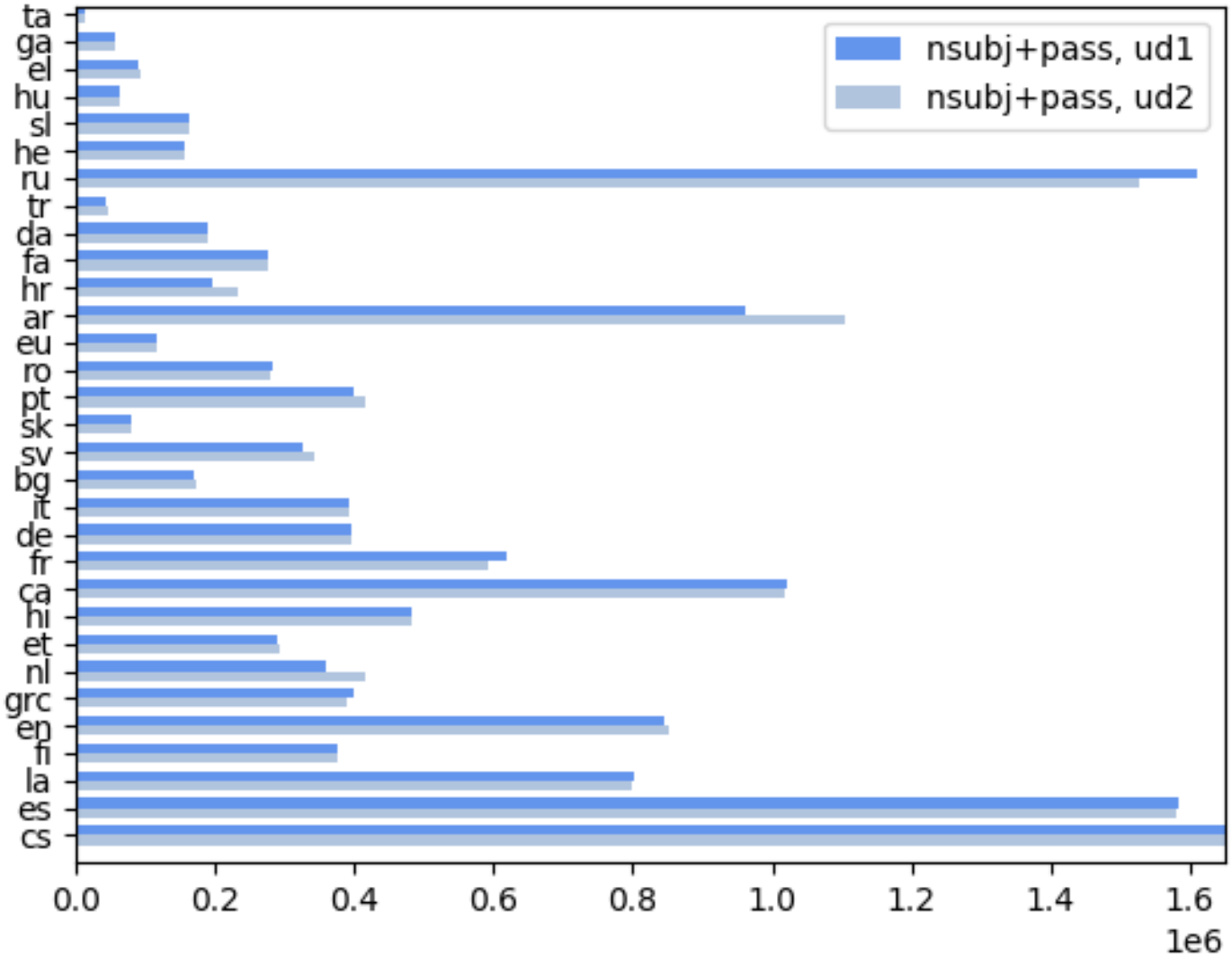


Figure: NSUBJ relation heads, incl. variations of PASS

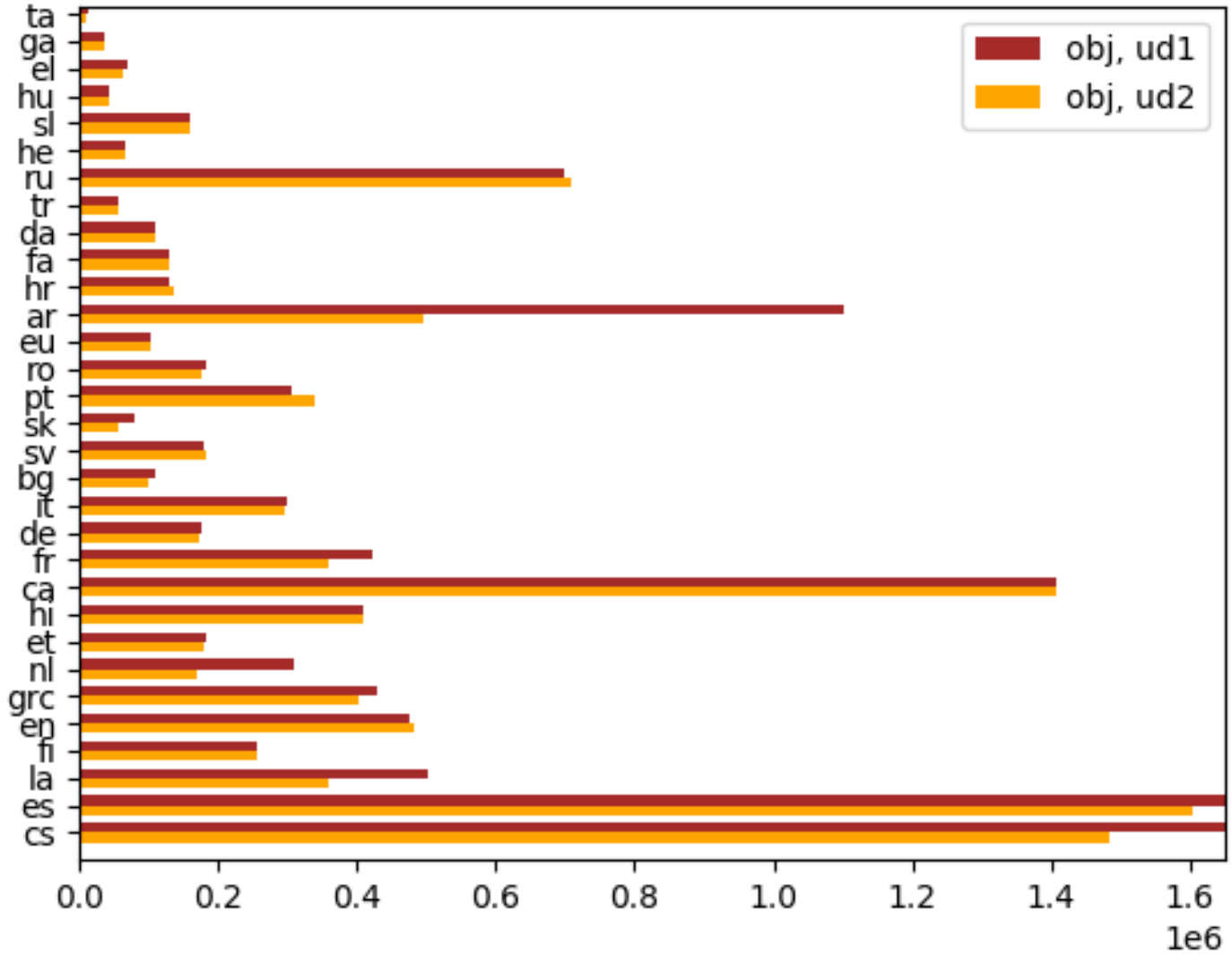


Figure: (D)OBJ relation heads

Conclusion

Can we draw reliable conclusions from dependency corpora?

- hard to find a **robust definition** for measure (subtrees!)
- changes in annotation $\rightarrow$ changes in big picture
- **random sampling** selectively unreliable (measure complexity)
- ...not to mention local treebank variations

And then

- enhanced dependencies
- manually annotated vs. automatically parsed
- and that's just sticking to one treebank environment!

Thanks!

Questions? Comments?