UNIVERSITÄT LEIPZIG

erc European Research Council

Communicative efficiency and syntactic predictability:
A cross-linguistic study based on the UD corpora

Natalia Levshina

Leipzig University

1. Communicative efficiency

- Language users are ‘engineers’ constantly optimizing communication, trying to make it maximally efficient.
- Economy: predictable information can be expressed by less linguistic material than surprising information (cf. neo-Gricean pragmatics).
- Economy can be observed ‘online’ (e.g. omission of segments, loss of articulatory detail, contextual ellipsis) and ‘offline’ (e.g. the omission becomes conventionalized).
- ‘Offline’ economy effects in word length:
  - Zipf’s (1935) Law of Abbreviation: frequent words are shorter, rare words are longer.
  - Piatandosi et al. (2011): ngram-based average predictability is correlated with word length even more strongly than simple frequency.

2. Research hypothesis

- Syntactic predictability is predictability of a word given its 'partner' in a syntactic dependency.

- e.g. **the** genuine antique mahogany **table**

  P(the|table) > P(table|the)

  the is more predictable than *table*

- Average syntactic predictability of a word is expected to be negatively correlated with word length.

3. UD corpora 2.0

- Arabic
- Chinese
- English
- Finnish
- German
- Hindi
- Persian
- Russian
- Spanish

4. Average Syntactic Information Content (ASIC)

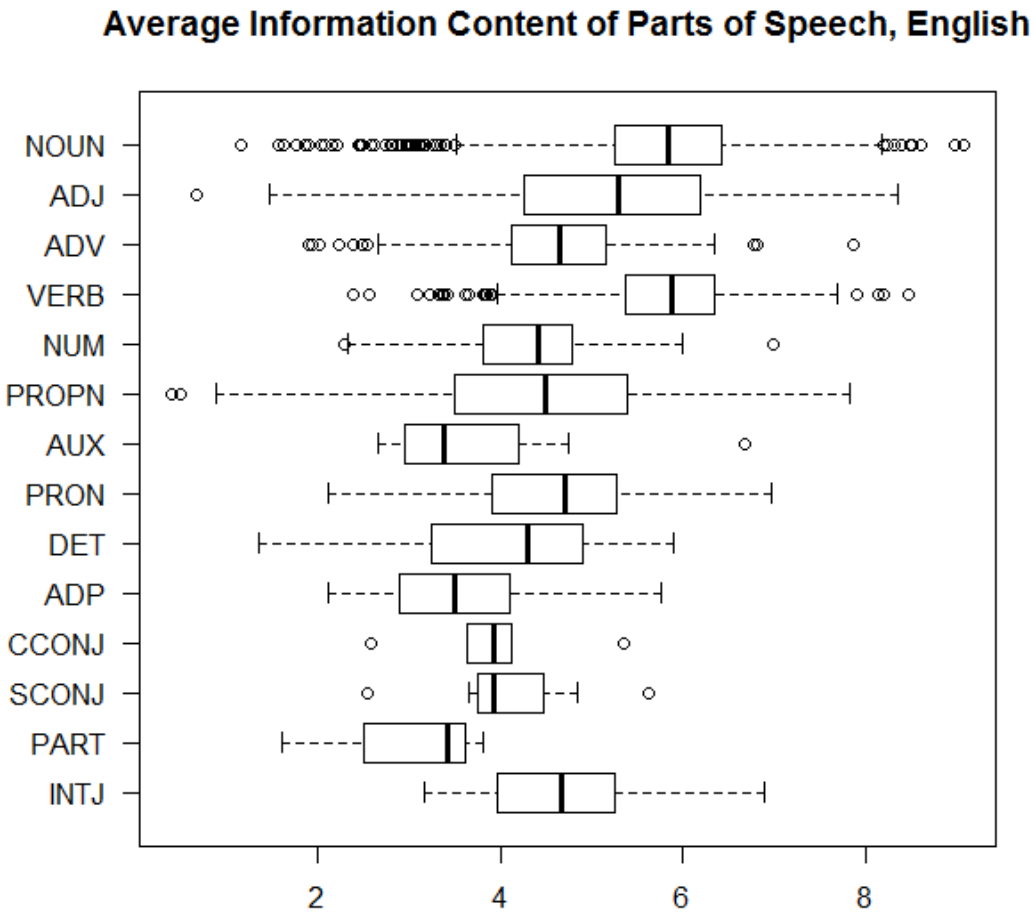
- Dependency triplets, e.g.

  the/DET det table/NOUN

  woman/NOUN nsubj walks/VERB

- Conditional probabilities:
  - P (the|table) = Frequency of triplet divided by frequency of table/NOUN
  - P (table|the) = Frequency of triplet divided by frequency of the/DET
- ASIC scores are negative log-transformed average conditional probabilities.
- ASIC is the reverse of predictability. I expect words with high ASIC to be longer, and words with low ASIC to be shorter (positive correlation).

5. ASIC by POS

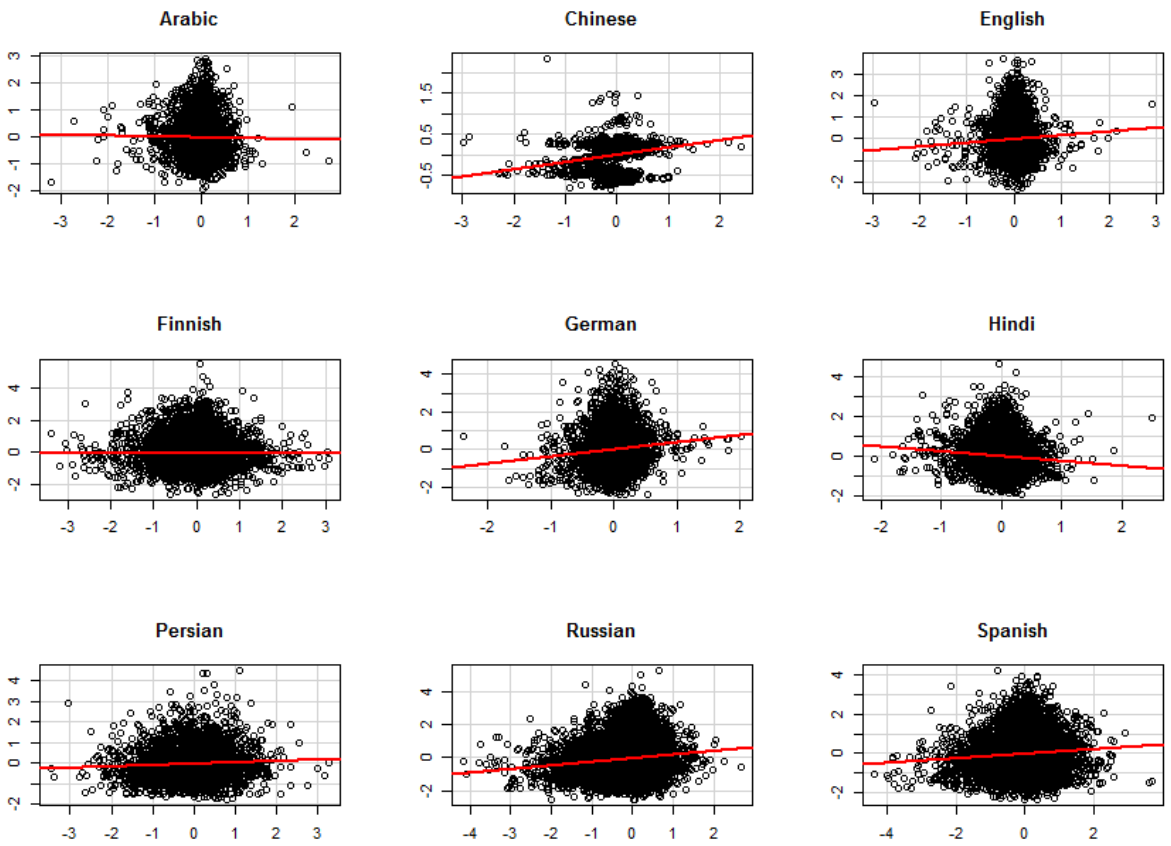

Average Information Content of Parts of Speech, English
NOUN
ADJ
ADV
VERB
NUM
PROPN
AUX
PRON
DET
ADP
CCONJ
SCONJ
PART
INTJ
2
4
6
8
Average Syntactic Information Content

6. Correlations between ASIC and Word Length*

Language	Spearman correlation coefficient
Arabic	-0.07***
Chinese	0.29***
English	0.11***
Finnish	-0.05***
German	0.18***
Hindi	0.24***
Persian	0.11***
Russian	0.11***
Spanish	0.14***

*Word length was measured as UTF-8 string length.

7. Poisson GLM (Length ~ ASIC + log-frequency + ngram Info)



Partial regression plots (effect of ASIC while controlling for all other variables). R package *car* (Fox & Weisberg 2011). *Ngram Info* means average inf. content given 1 word on the left and average inf. content given 1 word on the right.

8. Conclusion and future work

- We find a significant positive effect of ASIC on word length in most languages, also when frequency and ngram-based predictability are controlled for.
- Next step: test on larger corpora.

8 1/2

百花齊放
Let a hundred UD
corpora bloom!