2025/06/27

QUALICO2025 @ Masaryk University, Brno, Czech

On Linguistic Complexity: Form-Meaning Pairing Through Subword Token Polysemy

Takuto NAKAYAMA (Keio University, Tokyo, Japan)

Table of Contents

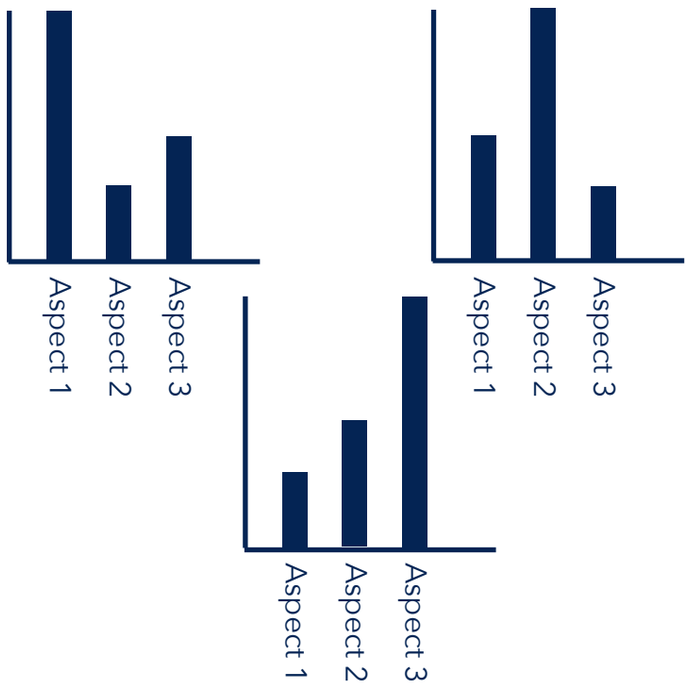
Slides →

GitHub →

1. Background & Purpose

Background: Equi-complexity

All languages are equally complex.



- more complex inflections
  →fewer word order rules
- more vocabulary
  →fewer inflectional rules

etc.

→Equi-complexity of Language

Background: Trade-offs=Equi-complexity

"Impressionistically it would seem that the total grammatical complexity of any language, counting both morphology and syntax, is about the same as that of any other." (Hockett, 1958, 180)

"No one should draw the conclusion from the paper that the Pirahã language is in any way 'primitives'. It has the most complex verbal morphology I am aware of. And a strikingly complex prosodic system."
(Everett, 2005, 62)

Background: Trade-offs≠Equi-complexity

"As long as it is impossible to quantify the overall complexity of a single language, it is also impossible to compare different languages with respect to that quantity." (Fenk-Oczlon and Fenk, 2014, 145)

"As to linguistic examples concerning the limited relevance of such trade-offs for the claim of an equal overall complexity, let us take our significant negative cross-linguistic correlation 'the fewer phonemes per syllable, the more syllables per word'. This correlation can be interpreted as a complexity trade-off between phonological complexity and morphological complexity. But it does not at all indicate an equal overall complexity" (Fenk-Oczlon and Fenk, 2014, 151)

Background: Semantics is less focused

Little research has focused on semantics

- Formal aspects have been focused more than semantic ones.
- Semantics is undispensable to see an overall linguistic complexity.

> "Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. **These semantic aspects of communication are irrelevant to the engineering problem.** The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design." (Shannon, 1948, 1)

Background: Summary

- As long as there is no measurement for an overall linguistic complexity, we can't say anything about the equi-complexity hypothesis.

  → **Focusing on every micro-level feature to depict an overall linguistic complexity.**

- There has been little research focusing on semantic aspects concerning in linguistic complexity yet.

  → **Focusing on semantics**

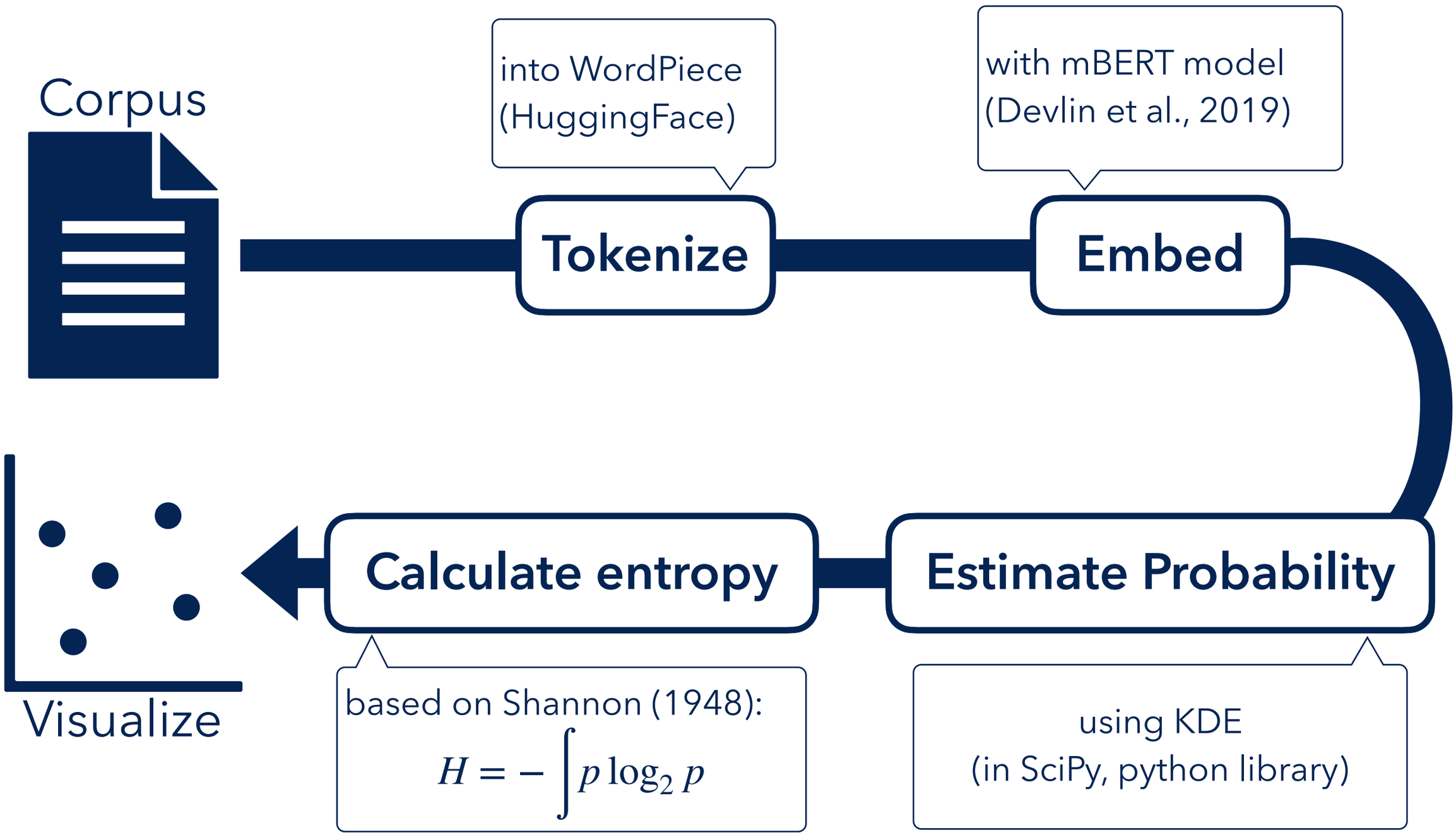
Purposes

1. to compare the degree of polysemy across languages by focusing on form-meaning pairings as a measure of linguistic complexity,

2. to introduce a method in which the number of meanings a form corresponds to is automatically estimated.

2. Methodology

Methodology



Methodology: 1. Tokenization

WordPiece (HuggingFace)

[…, w##, ##o, ##r, ##d, …] ← 1. split into the characters. (the slot referred with ##)

…, **wo##**, ##or, ##rd, … ← 2. merge adjacent pair and compute a score below:

$$\text{Score} = \frac{\text{frequency}_{\text{pair}}}{\text{frequency}_i \times \text{frequency}_j}$$

$\frac{2}{3}$	$\frac{1}{6}$	$\frac{1}{6}$

[…, w##, **wo##**, ##o, ##r, ##d, …] ← 3. The pair of the highest score is merged.

The process will go on until reaching the desired vocabulary size.

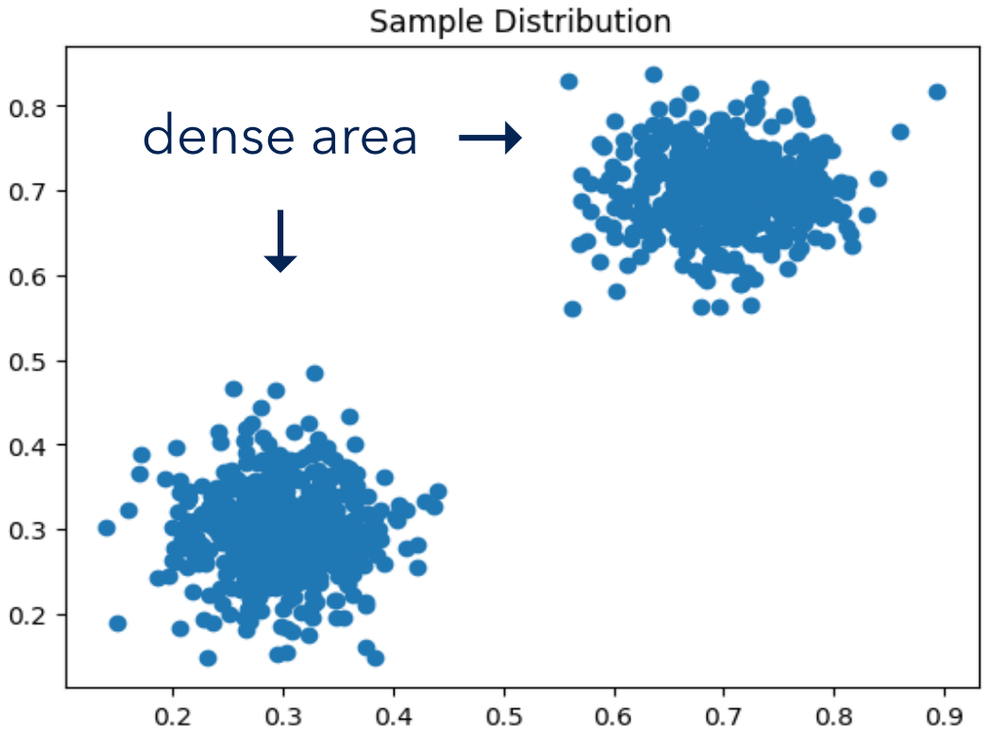
Methodology: 2. Embedding

multilingual BERT (Devline, et al., 2019)

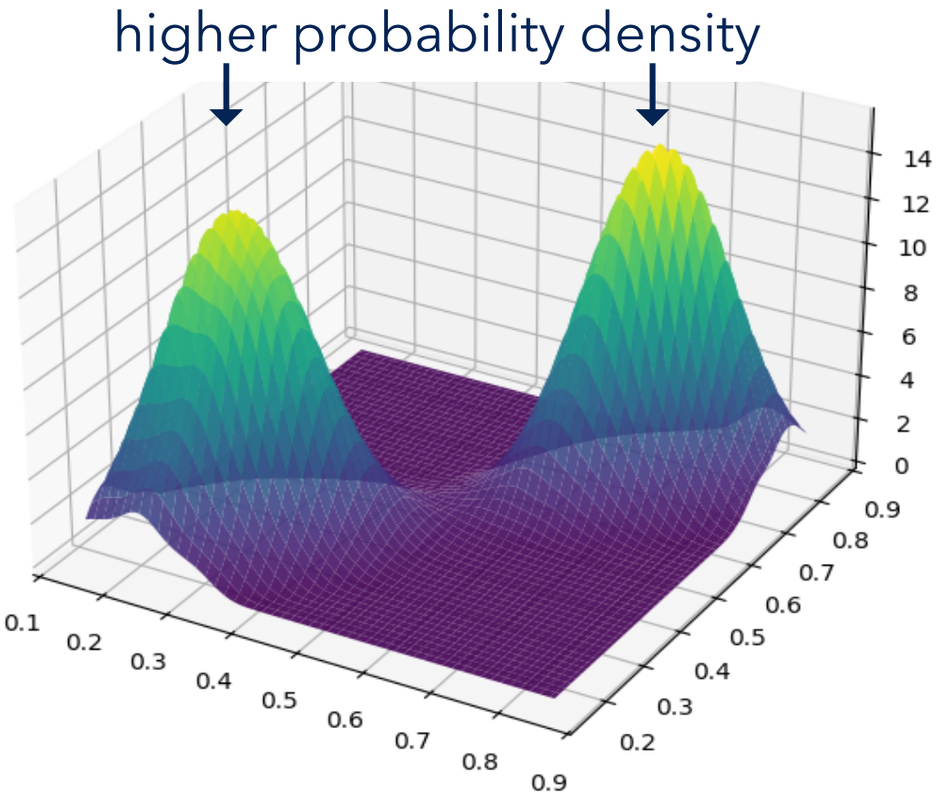
- the model, bert-based-multilingual-cased,
- got contextualized embeddings for each subword,
- used the last layers as embeddings (768 dimensions),
- classified the embeddings based on each subword,
- reduced the dimensions with tSNE (768→2 dimensions).

Methodology: 3. Estimation of Probability

Kernel Density Estimation (Rosenblatt, 1956; Panzen, 1962)



scatter plot
(Raw embeddings)



probability distribution
estimated from the embeddings

Methodology: 4. Calculation of Entropy

Entropy (Shannon, 1948)

An entropy H_i for subword i is given by:

$$H_i = -\int p(x)\ \log_2 p(x)\, dx\ ,$$

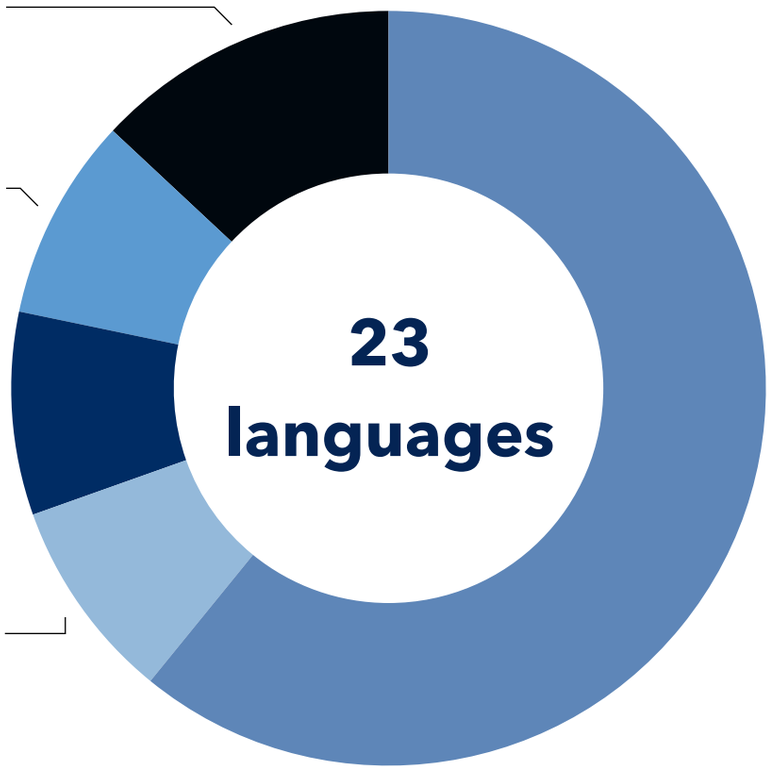
where $p(x)$ refers to the probability at a certain area x within a embedding space.

Then, the entropy H for a language is given by:

$$H = \frac{1}{n}\sum_{i=1}^{n} H_i\ ,$$

where there are n subwords in the texts.

Dataset



Dryer and Haspelmath (2013)

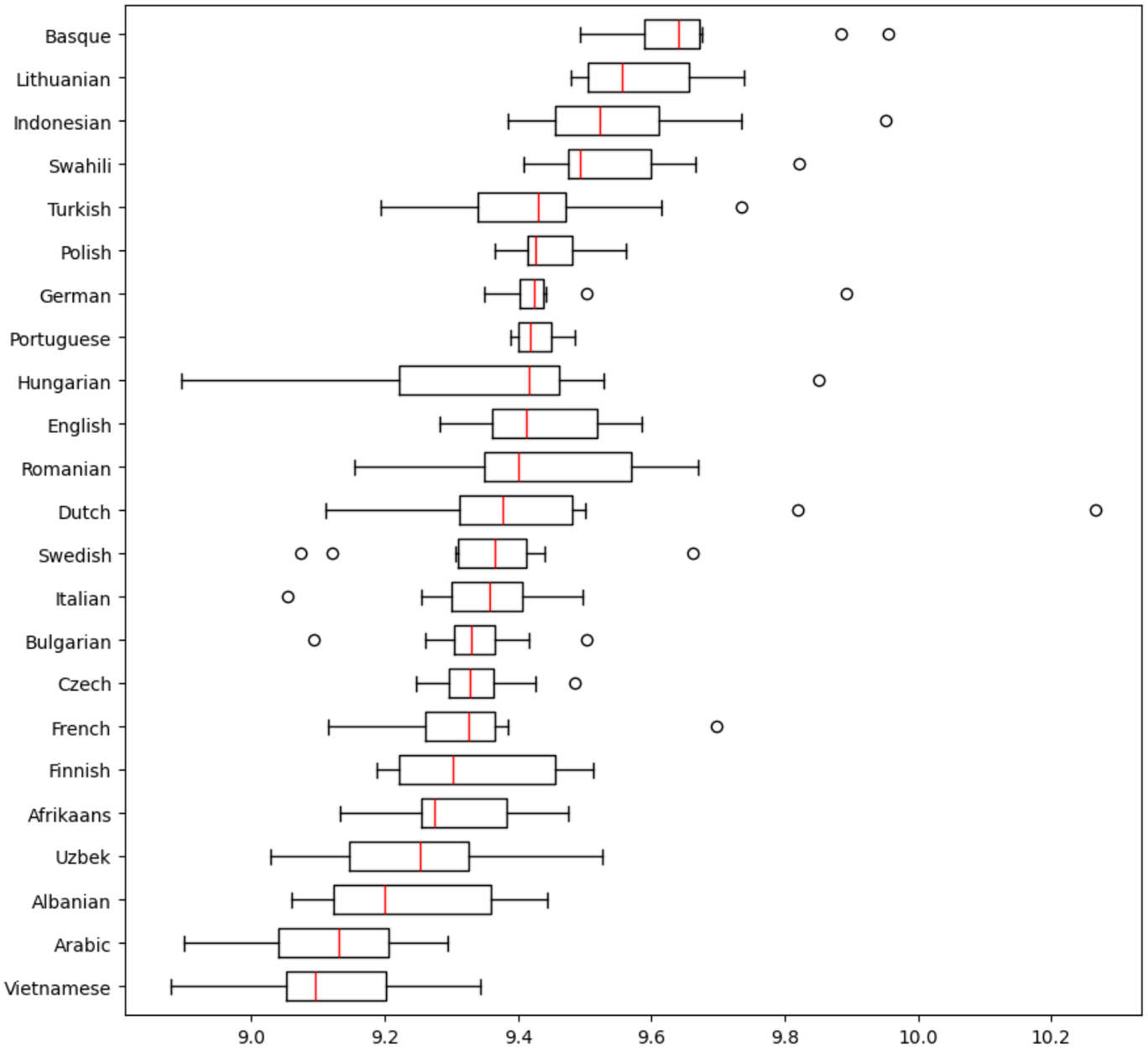
100 wikipedia articles
×
10 times
for each language

<Language List>

Afrikaans, Albanian, Arabic, Basque, Bulgarian, Czech, Dutch, English, Finnish, French, German, Hungarian, Indonesian, Italian, Lithuanian, Polish, Portuguese, Romanian, Swahili, Swedish, Turkish, Uzbek, Vietnamese

3. Result

Result: Average entropy



Median

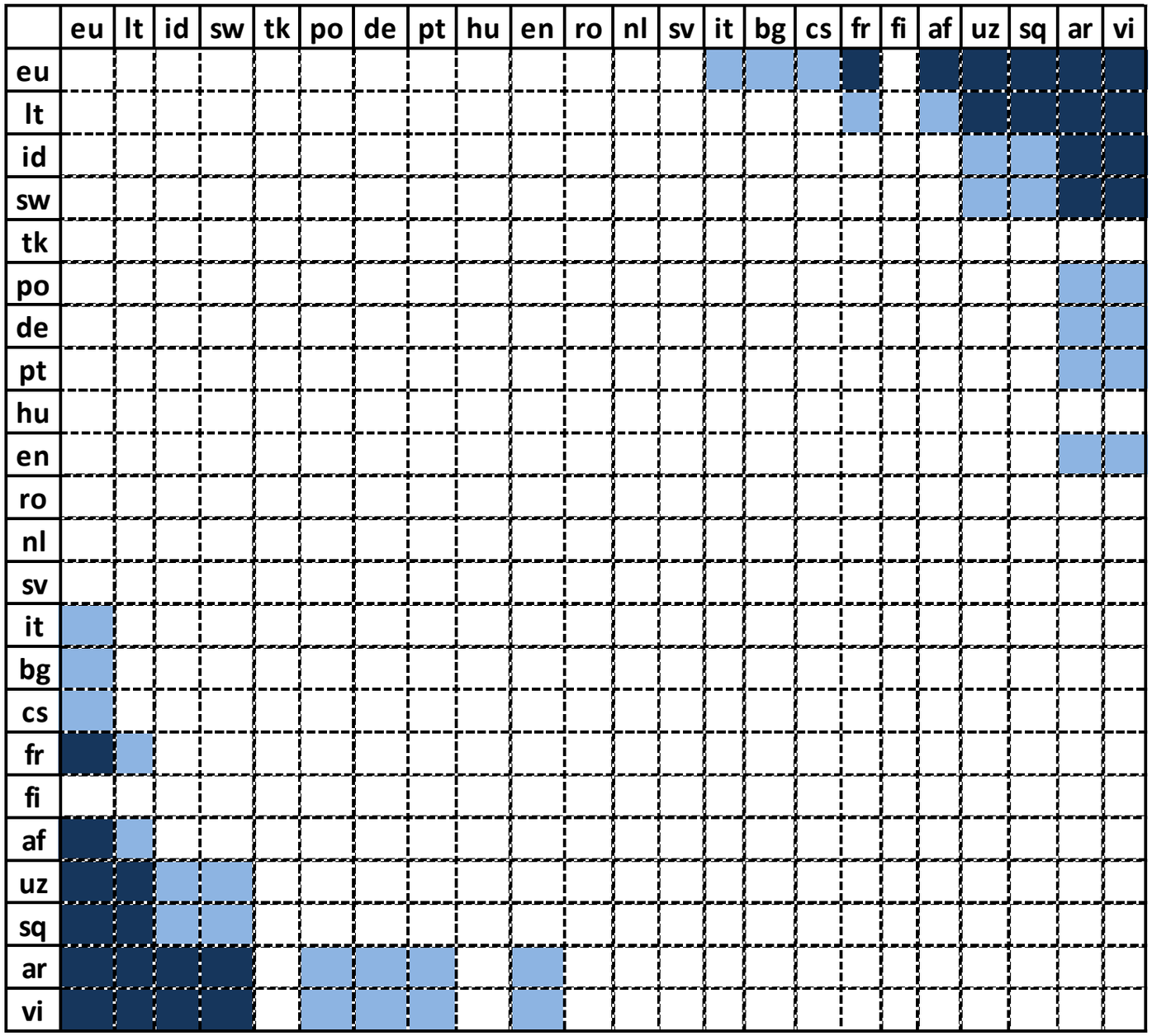
- max: Basque ≈ 9.642
- min: Vietnamese ≈ 9.097

Each data point

- max: Dutch ≈ 10.266
- min: Hungarian ≈ 8.951

→The range is not narrow, but not totally random either

Result: Dunn's test



: $p < 0.05$

: $p < 0.01$

- There are significant differences between distribution of the enntropies.
- Languages have a loose similarity in their polysemy of the vocabulary.

4. Conclusion

Conclusion & Limitations

Conclusion

- Polysemy of vocabulary is not the same among languages, but not totally at random neither.
  →at least, all languages seem not to be equally complex.
- Languages share a loose similarity in their polysemy of the vocabulary.
  →This trend might make us see linguistic complexity similar

Limitation

- The size of the data is not enough.
- The relationship between polysemy of vocabulary and overall linguistic complexity is not clear yet.

References

Acknowledgement: This work was supported by JSPS KAKENHI Grant Number JP24KJ1938.

Everett, D. L. (2005). Cultural Constraints on Grammar and Cognition in Pirahã: Another Look at the Design Features of Human Language. *Current Anthropology*, *46*(4), 621–646. https://doi.org/10.1086/431525

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). B*ERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv [cs.CL]. http://arxiv.org/abs/1810.04805

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *Wals online (v2020.4)* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13950591

Fenk-Oczlon, G., & Fenk, A. (2014). Complexity trade-offss do not prove the equal complexity hypothesis. *Poznań Studies in Contemporary Linguistics*, *50*(2), 145–155. https://doi.org/10.1515/psicl-2014-0010

Hockett, C. F. (1958). A Course in Modern Linguistics. Macmillan.

HuggingFace. (n.d.). *WordPiece tokenization*. https://huggingface.co/learn/nlp-course/en/chapter6/6

Parzen, E (1962) "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics* 33(3), 1065-1076.

Rosenblatt, M (1956) "Remarks on Some Nonparametric Estimates of a Density Function," The Annals of Mathematical Statistics 27(3), 832-837.

Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x