Keio University

CALAMVS GLADIO FORTIOR
1858

10.09.2024

The 21st International Congress of Linguists @ Poznań, Poland

Linguistic Complexity Through Form-Meaning Pairings:

An Information Theoretical Approach to Equi-Complexity of Language

Takuto NAKAYAMA (Keio University, Japan)

Outline

1. Background & Purpose
2. Previous Studies
3. Dataset & Settings
4. Results
5. Conclusion

Slides→

GitHub→

1.Background & Purpose

Background

Equi-complexity of language:

All languages are equally complex.

[I]n a comparison of any two languages higher and lower degrees of complexity in different sub-domains of morphological and syntactic structure will ultimately balance each other out.

(Kortmann and Schlöter, 2020)

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)

Purpose

This research aims to measure an overall complexity of a language:

1. by measuring the unpredictability of how many meanings a given form represents, and
2. basing on a linguistic unit independent of any specific language.

Why Form-Meaning Pairing?

- Currently, there is no consensus on how to measure overall linguistic complexity across languages.
- To address this, this research views language as a sequence of linguistic units that refer to something, which at least appears to be a common feature across all languages.
- In this context, a straightforward definition of linguistic complexity is the number of meanings that one form refers to.

2.Previous Studies

Previous Studies

"Both simple and complex types of language of an indefinite number of varieties may be found spoken at any desired level of cultural advance. When it comes to linguistic form, Plato walks with the Macedonian swineherd, Confucius with the head-hunting savage of Assam." (Sapir, 1921, 268–269)

Edward Sapir

Charles Hockett

"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)

Previous Studies

Bentz et al., 2023

- Participants of the workshop evaluated languages based on various aspects.
- The evaluations were merged into vectors representing the overall complexity.
- Comparing these vectors, no significant differences were found.

→ Suggested that equi-complexity might be valid.

Koplenig et al., 2023

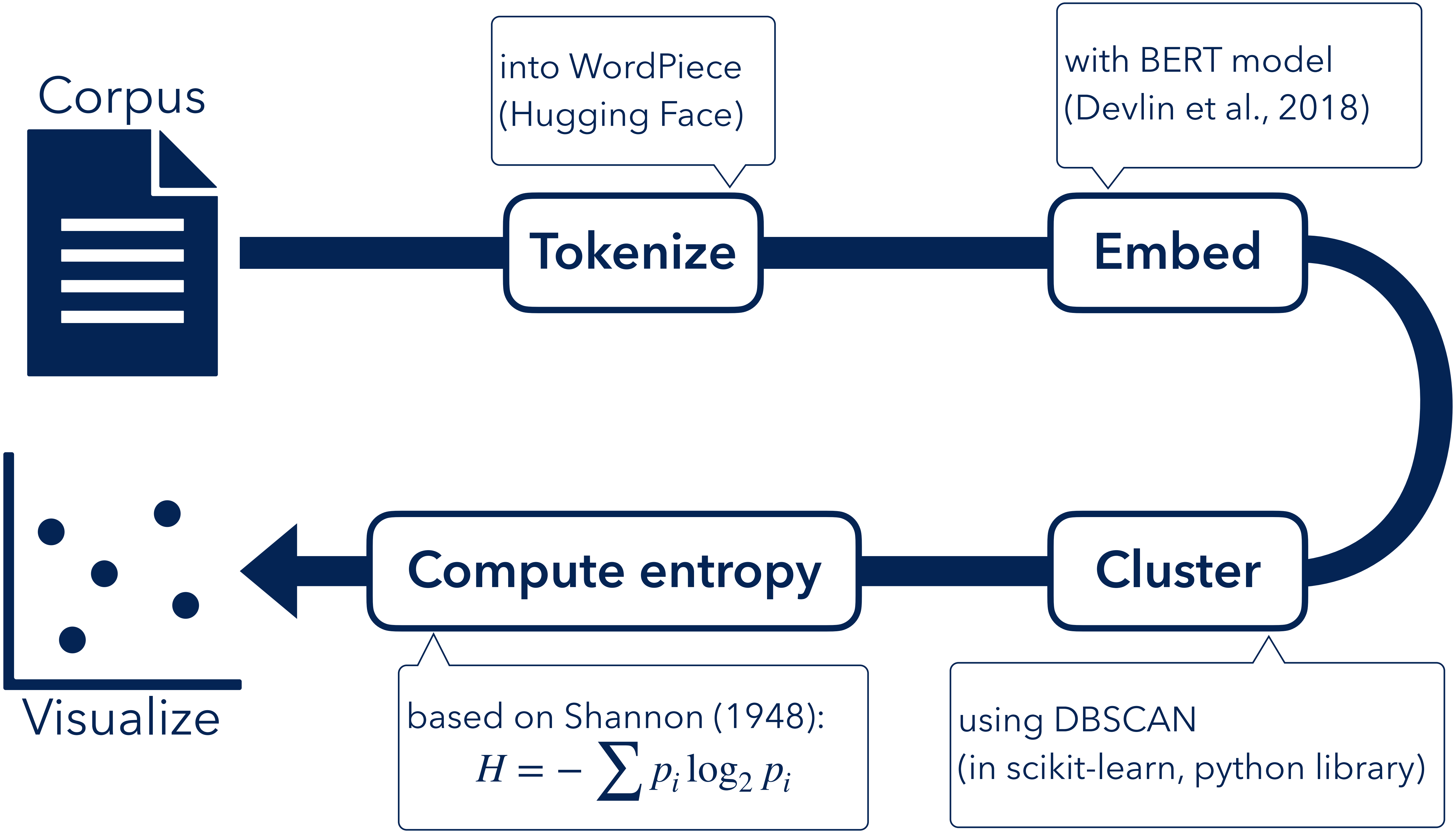
- Involved analyzing multiple corpora to assess the complexities of languages.
- Found that differences in complexity within one corpus were also found in others.
- Differences in linguistic complexity among languages are somehow significant.

→ Challenged the equi-complexity.

The debate of equi-complexity of language has not yet ended.

3.Dataset & Settings

Methodology



Corpus

Tokenize

into WordPiece
(Hugging Face)

Embed

with BERT model
(Devlin et al., 2018)

Cluster

using DBSCAN
(in scikit-learn, python library)

Compute entropy

based on Shannon (1948):

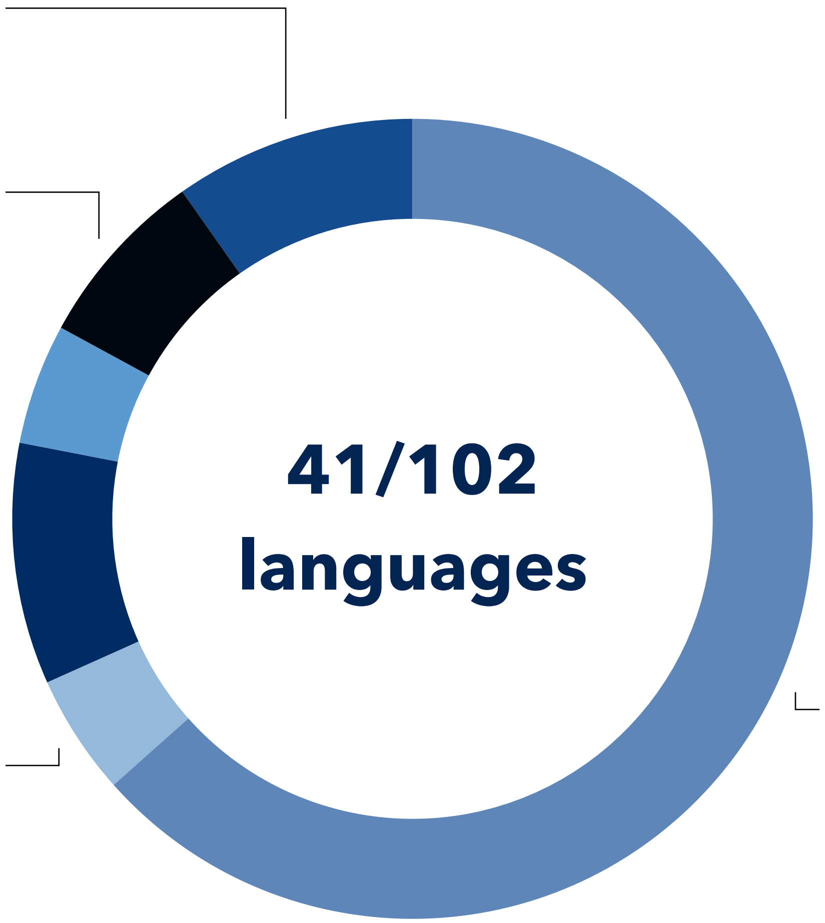
$$H = -\sum p_i \log_2 p_i$$

Visualize

Dataset

Multilingual Bible Parallel Corpus

(Christodouloupoulos and Steedman, 2015)



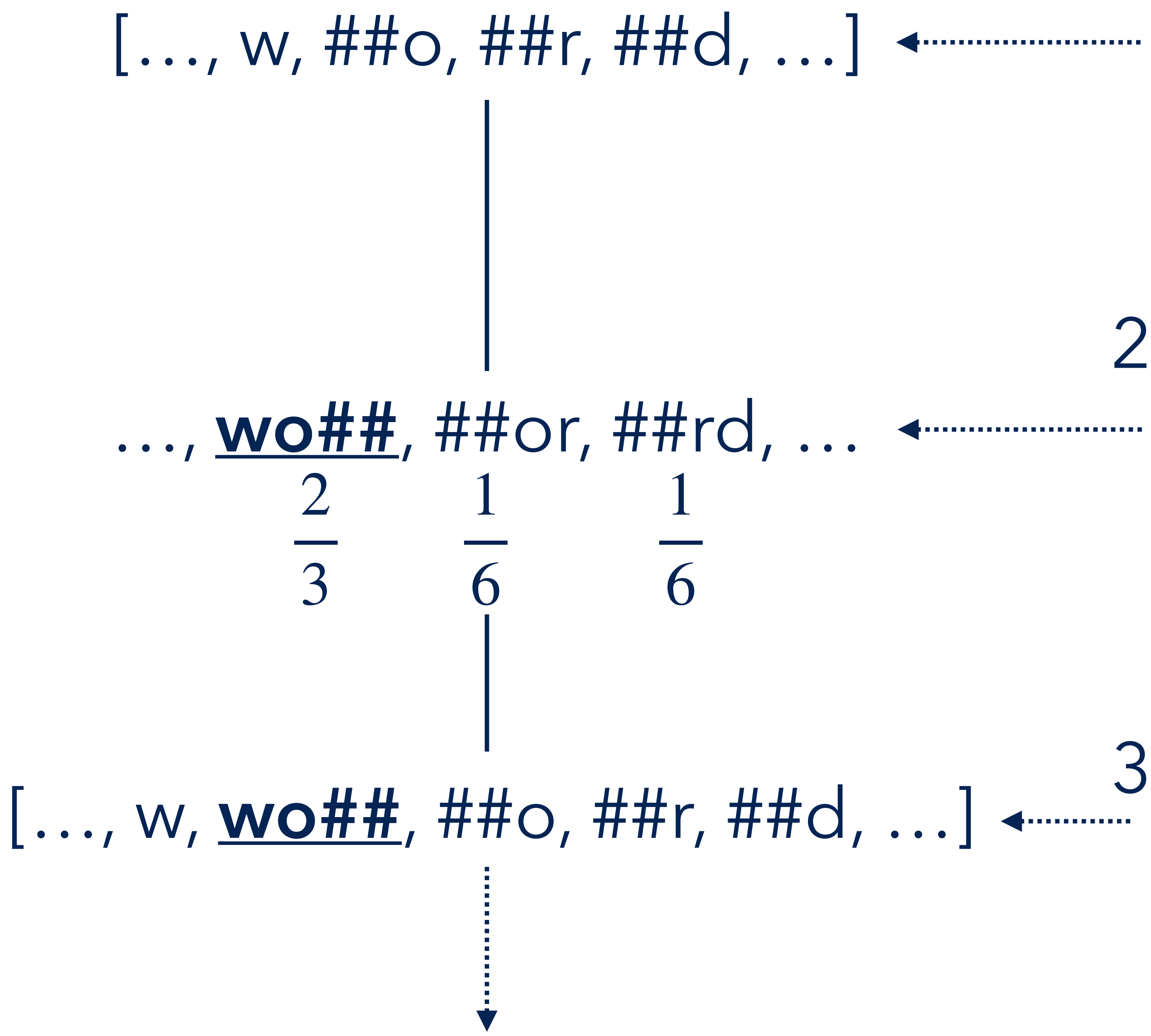
Used data that both:

- consist of the Old and New Testament, and
- the BERT model pre-trained for.

Given data: 41 languages

GitHub→

WordPiece (Hugging Face)



[…, w, ##o, ##r, ##d, …]

…, **wo##**, ##or, ##rd, …

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

[…, w, **wo##**, ##o, ##r, ##d, …]

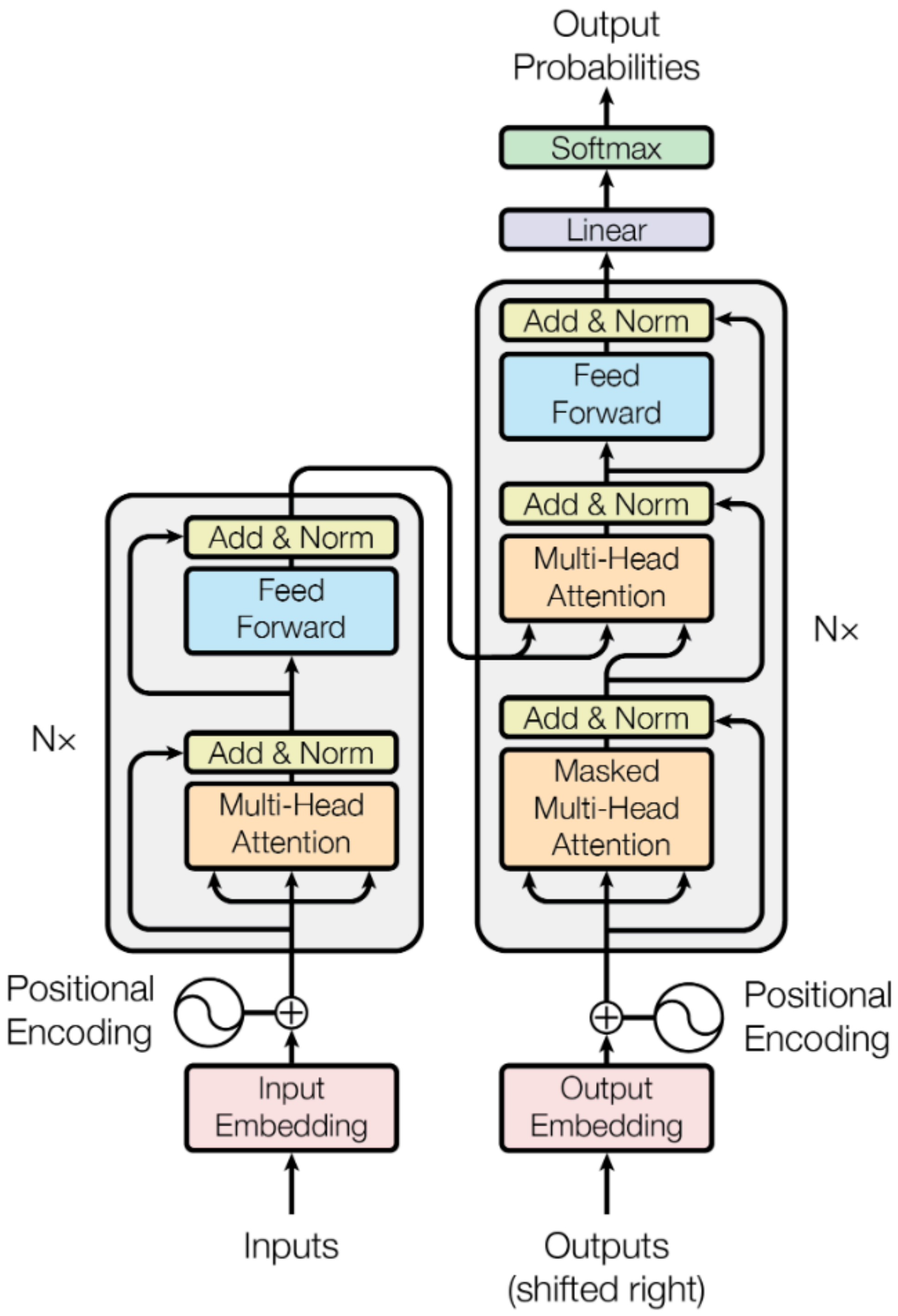
1. split into the characters. (the initial without special marks; the rest with ## referring to a slot)
2. merge adjacent pair and compute a score below:

$$\text{Score} = \frac{\text{frequency}_{\text{pair}}}{\text{frequency}_{\text{i}} \times \text{frequency}_{\text{j}}}$$

3. The pair of the highest score is merged.

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

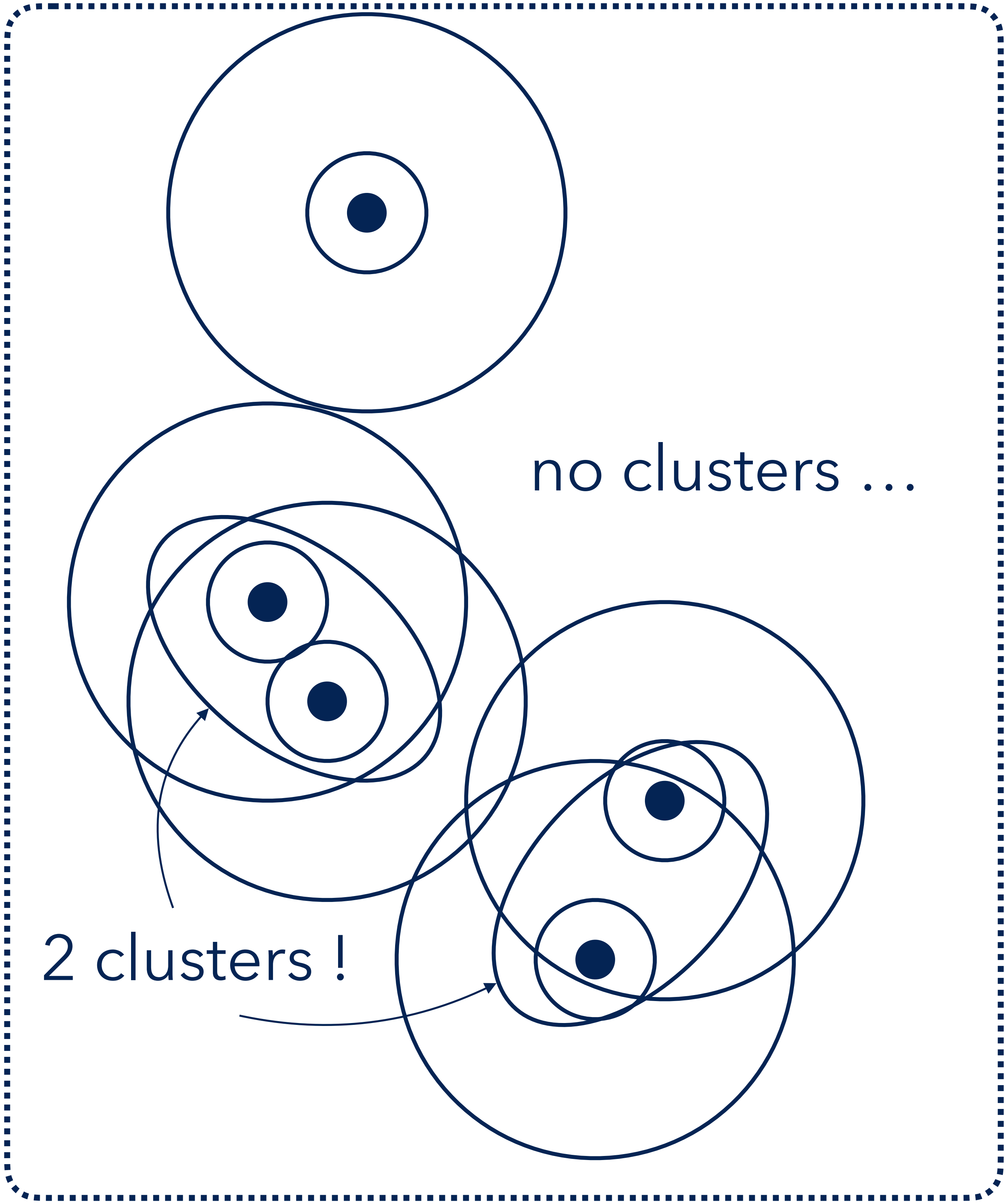
BERT



(Vaswani et al., 2017)

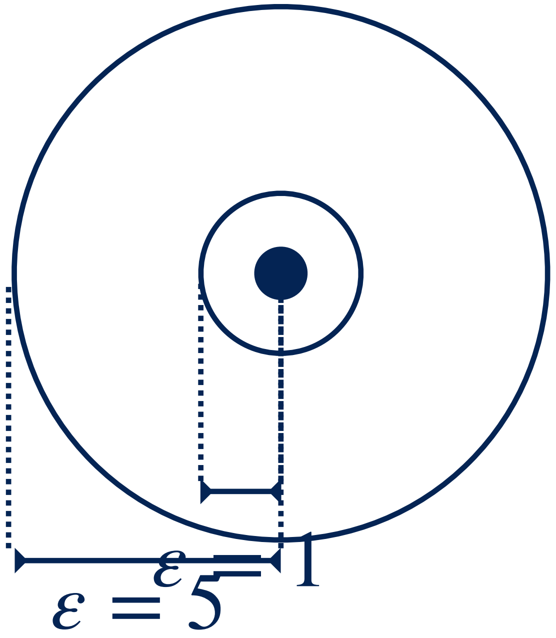
- A neural language model that is based on Transformers by HuggingFace.
- Context-based embeddings.
- In the present research, bert-based-multilingual-cased (Devlin et al., 2017) is used.
- The last layers: 768 layers.

DBSCAN



no clusters …

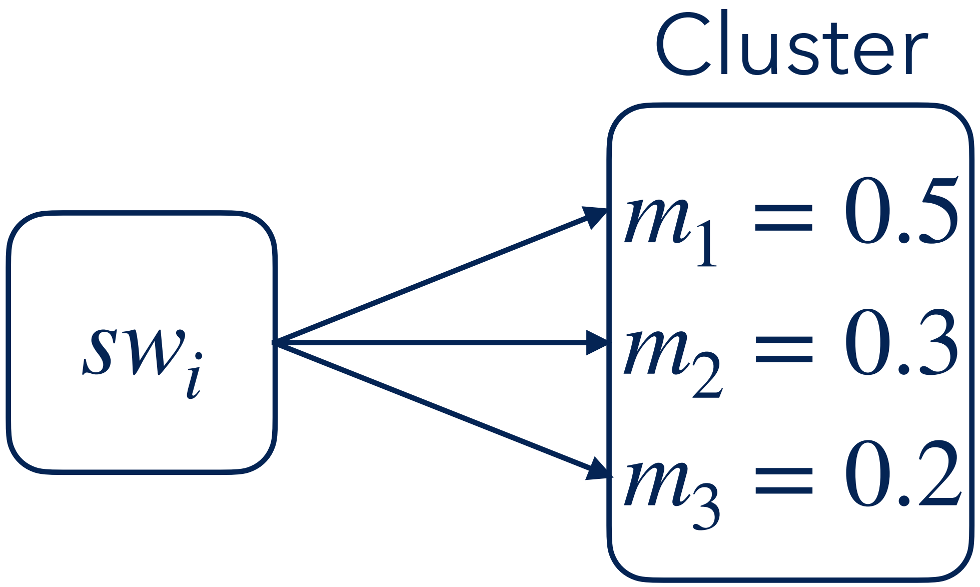
2 clusters !



Shannon Entropy (Shannon, 1948)

$$H_{language} = -\frac{1}{N}\sum^{i}\sum^{j} p(sw_{ij})\log_2 p(sw_{ij}),$$

in which the language has N types of subwords and $p(sw_{ij})$ refers to the jth meaning of the ith subword in the language.



Cluster

sw_i

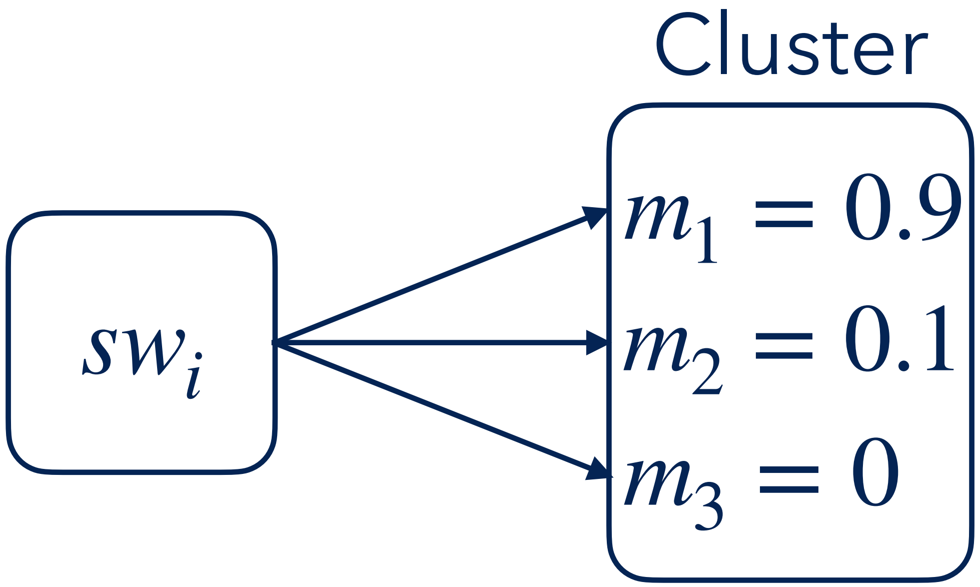
$m_1 = 0.5$

$m_2 = 0.3$

$m_3 = 0.2$

$H \approx 1.4854$

→higher complexity



Cluster

sw_i

$m_1 = 0.9$

$m_2 = 0.1$

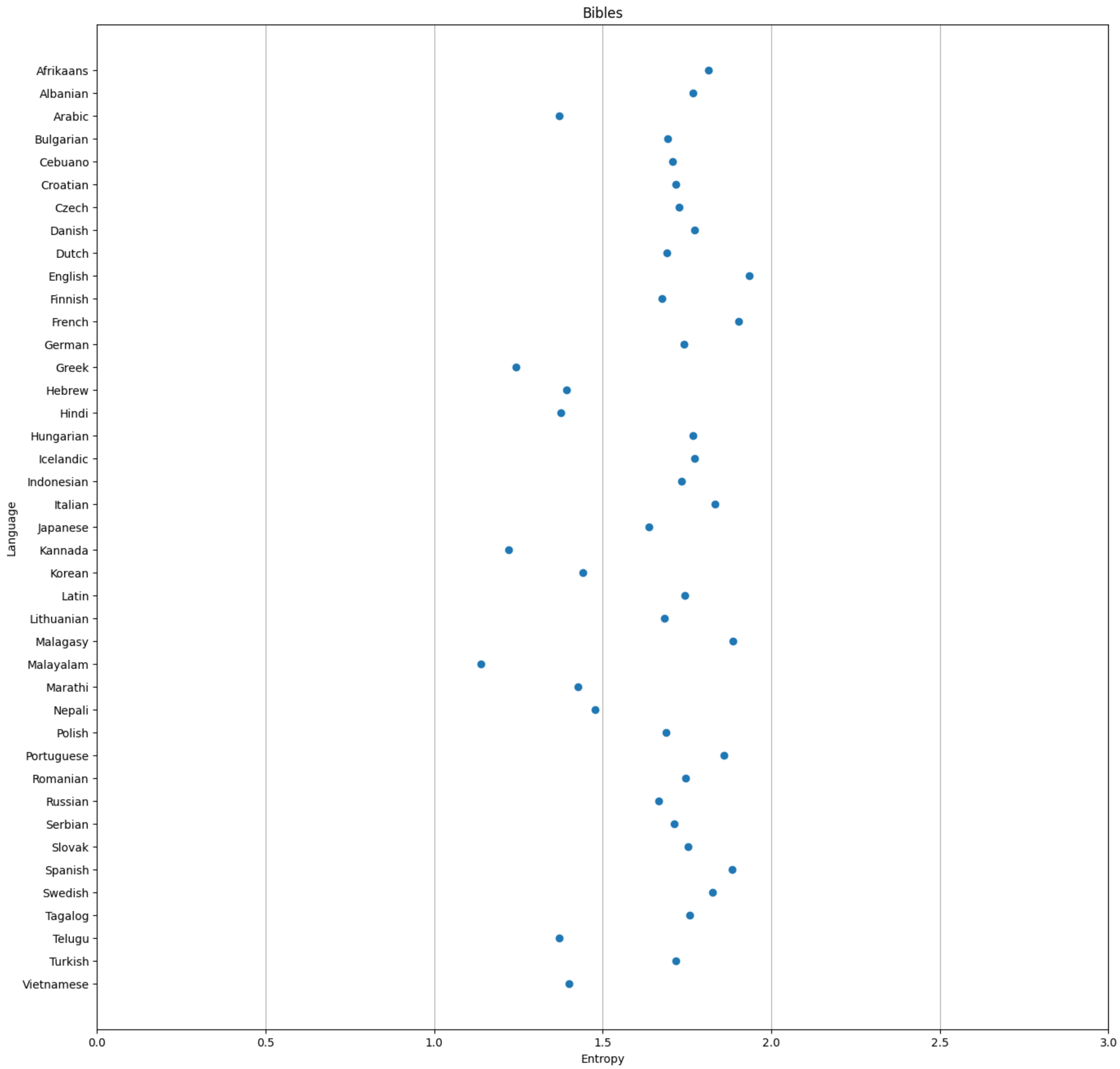
$m_3 = 0$

$H \approx 0.4689$

→lower complexity

4.Results

Results



- Indo-European languages and some others are placed from about 1.6 to 2.0

→3–4 meanings/sw

- The others are from 1.13 to 1.5

→2–3 meanings/sw

5.Conclusion

Conclusions & Limitations

Conclusions

- There seems to be a difference in linguistic complexity among languages, due to language families.
- Equi-complexity is a loose tendency among languages.

Limitations

- Further research should:
  - use data from balanced corpora, and
  - a larger number of languages (e.g., > 100).
- Only suggested “how many meanings” each subword might refer to, but not answered “why such number”.

References

Bentz, C., Gutierrez-Vasques, X., Sozinova, O., & Samardžić, T. (2023). Complexity trade-offs and equi-complexity in natural languages: a meta-analysis. *Linguistics Vanguard*, *9*, 9–25. https://www.degruyter.com/document/doi/10.1515/lingvan-2021-0054/html

Christodouloupoulos, C., & Steedman, M. (2015). A massively parallel corpus: the Bible in 100 languages. *Lang Resources & Evaluation*, *49*, 375–395. https://doi.org/10.1007/s10579-014-9287-y

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies, 1*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

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

Hockett, C. F. (1958). *A Course in modern linguistics*. Macmillan.

Hugging Face (n.d.). *WordPiece Tokenization*. https://huggingface.co/learn/nlp-course/en/chapter6/6#tokenization-algorithm

Koplenig, A., Wolfer, S., & Meyer, P. (2023). A large quantitative analysis of written language challenges the idea that all languages are equally complex. *Scientific Reports*, *13*(1), 15351. https://doi.org/10.1038/s41598-023-42327-3

Kortmann, B., & Schlöter, V. (2020, January 15). *Linguistic Complexity*. Oxford Bibliography. https://doi.org/10.1093/obo/9780199772810-0254

Sapir, E. (1921). *Language: An introduction to the study of research*. Harcourt, Brace & World Inc.

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

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