The Equi-complexity vs. Typology: Measurement of Overall Linguistic Complexity and Typological Categories

Takuto NAKAYAMA (Keio University, Tokyo, Japan)
e-mail: tnakayama.a5ling@gmail.com

1. Purpose

- To demonstrate the similarity of the overall complexities of three major typological categories: agglutinative, fusional, and isolating languages, while:

i) Considering multiple linguistic facets
ii) Dealing with the facets in the same way

2. Background

- It has been believed that no language is simpler/more complex than others; all are equally complex (cf. Hacket, 1958).
- We have not corroborated this belief yet, but some suggest that this belief is true (e.g., Bentz et al. 2022).
- Are agglutinative, fusional, and isolating languages equally complex, or variable in their linguistic complexities?

3. Methodology (Nakayama, 2023)

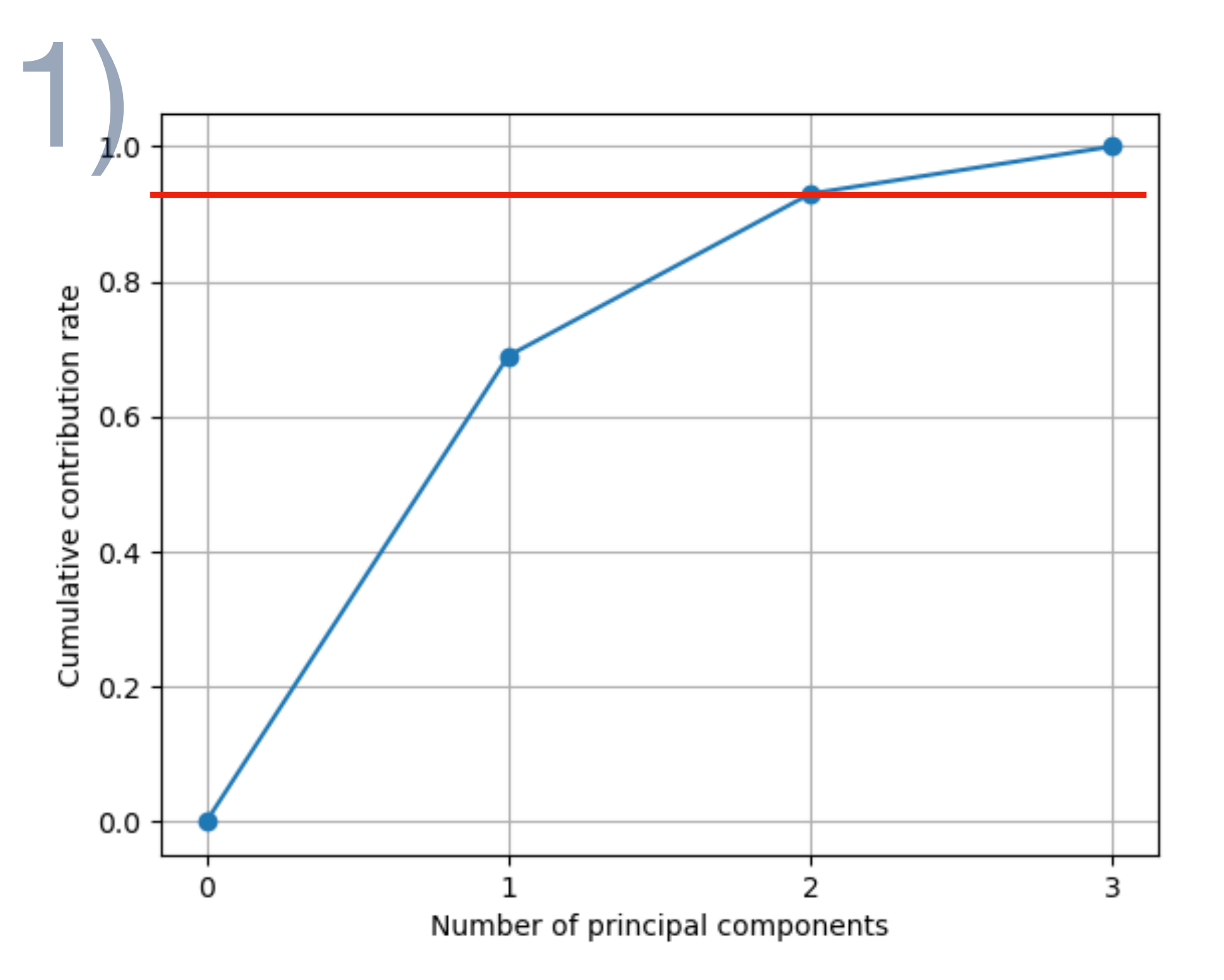
1. Calculate the entropy H (Shannon, 1948) of an n-element sequence from the ith to jth element of a document, the length of which is l:

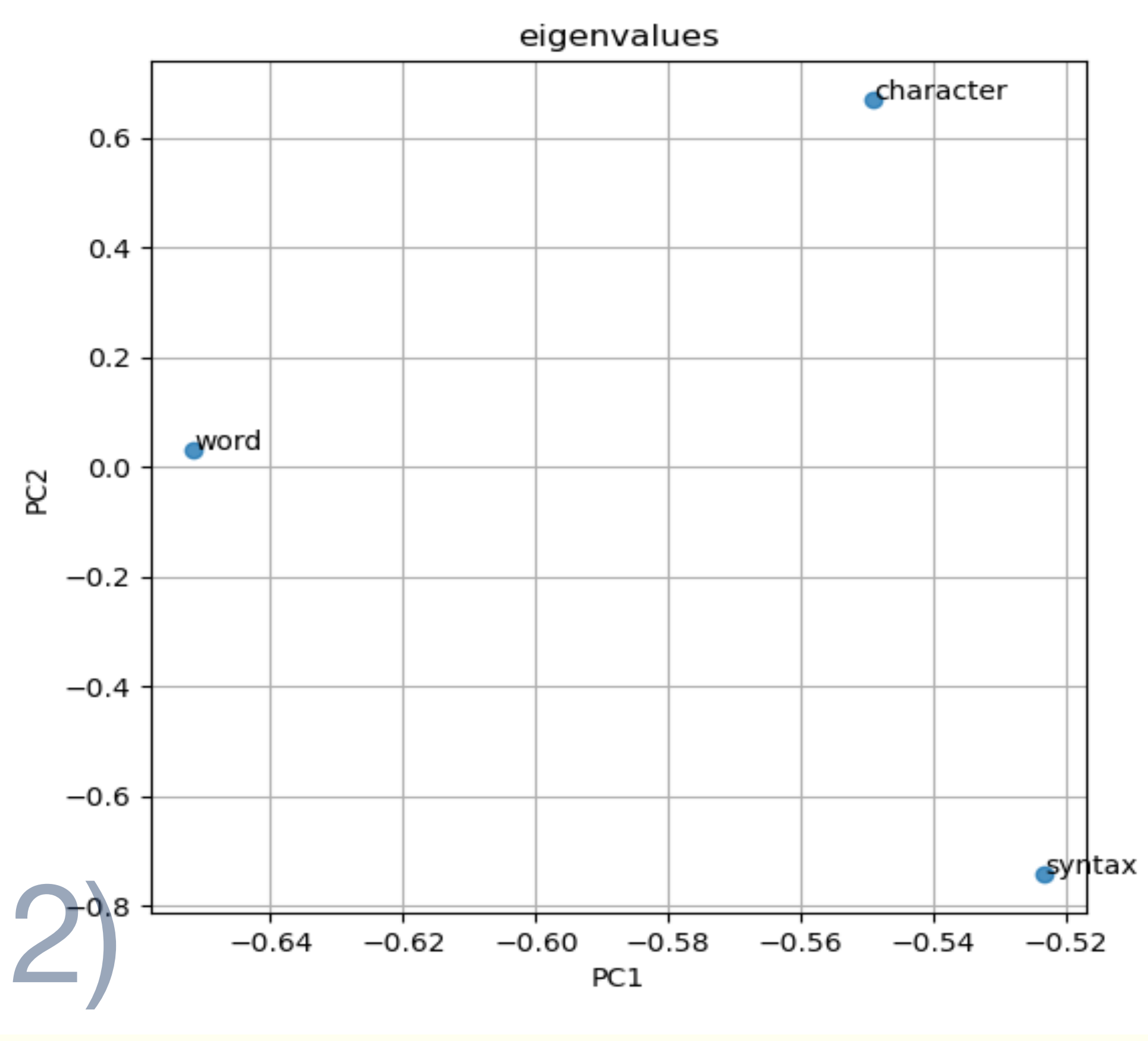
$$H(x_{ij}) = \sum_{i=1,j=n}^{l-n+1,l} \frac{1}{n} p(x_{ij}) \log_2 p(x_{ij}) \ (1 \leq n \leq M) \ ,$$

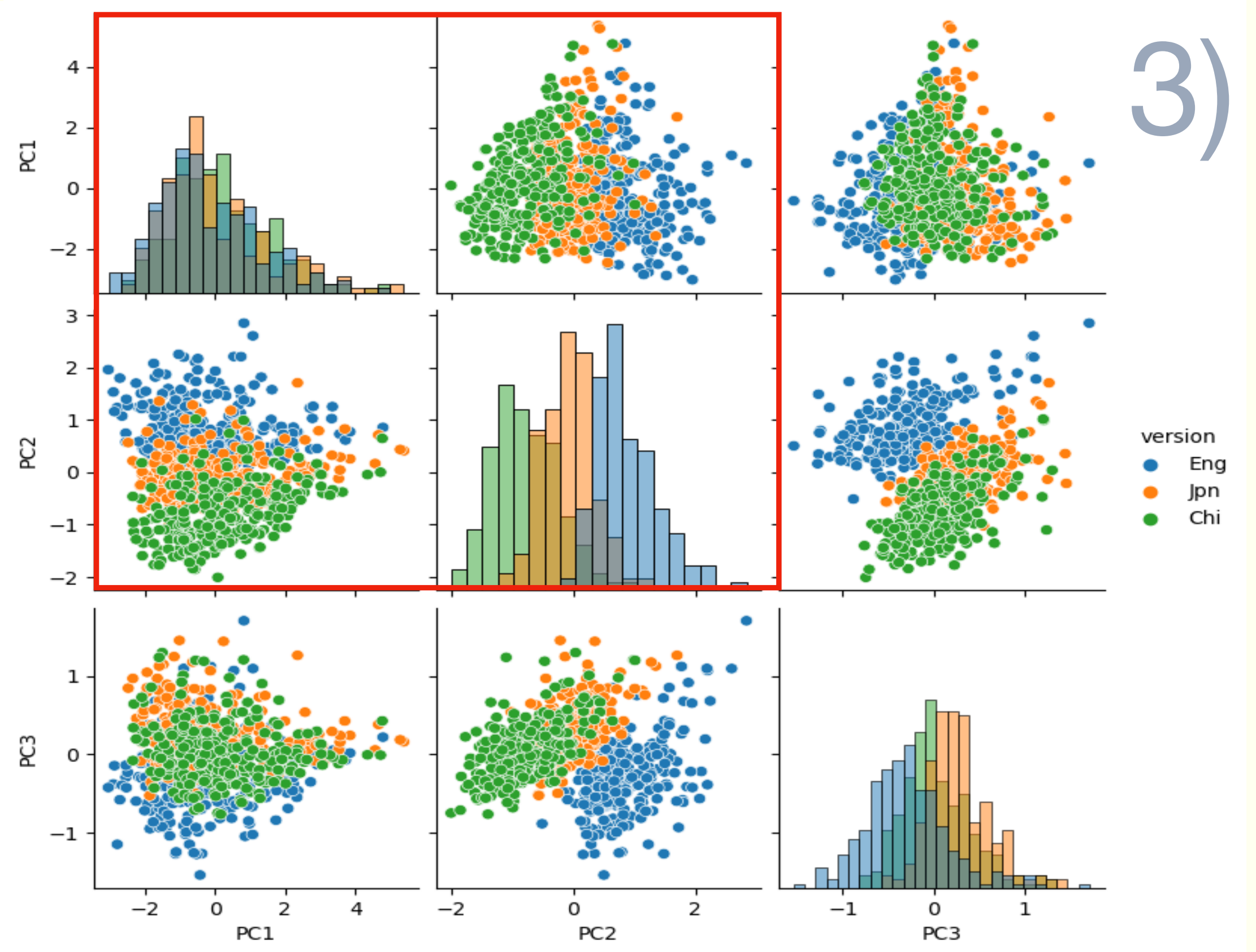
in which M refers to a number great enough to make all the strings different from each other.

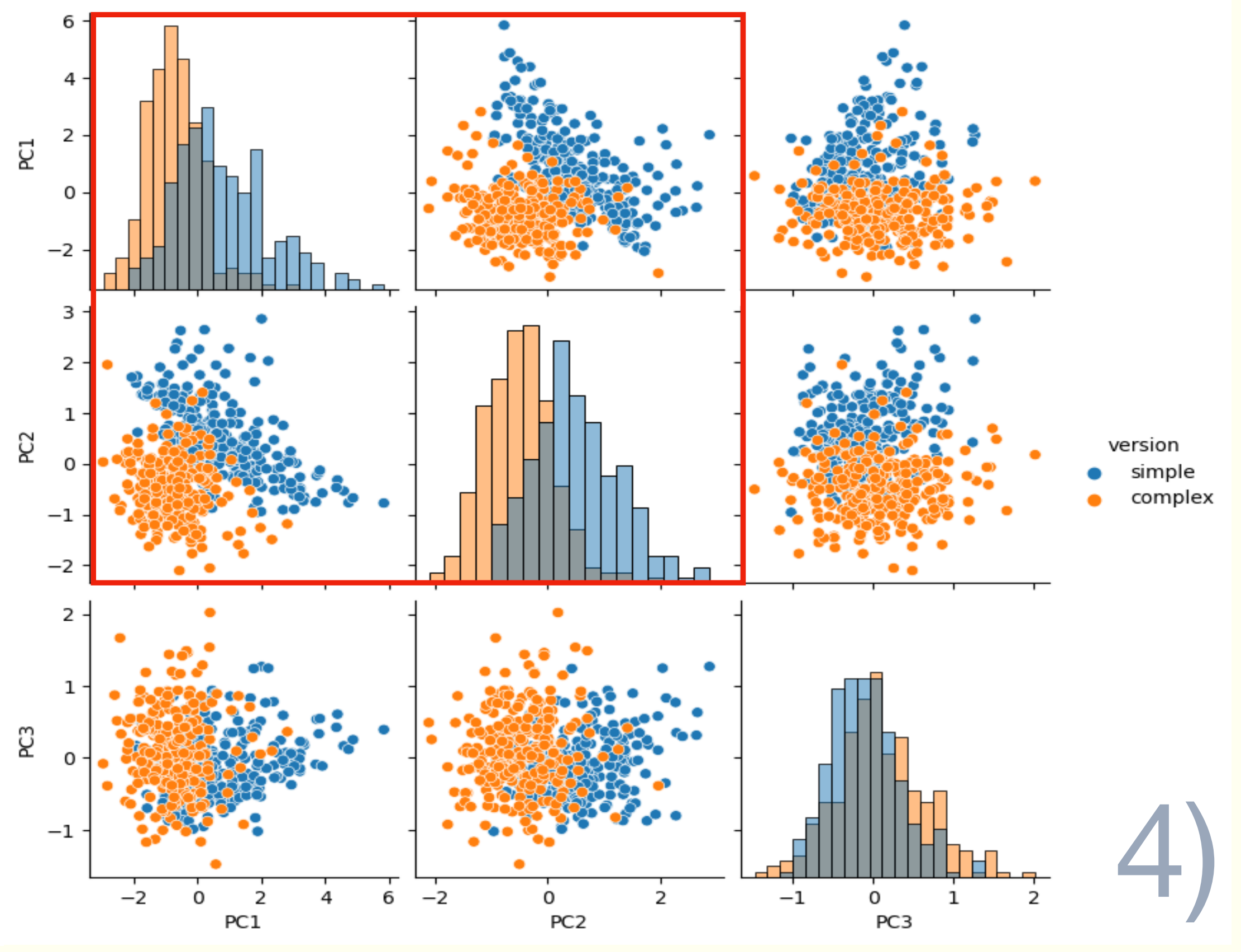
2. Determine the exponent from the power regression of Step 1 as a feature value of the entropy set in which the entropies decrease as n increases.
3. Apply Steps 1 and 2 to multiple facets of each document, such as characters and words, which gives a vector with multiple exponents (cf. Deutscher, 2009).
4. Conduct principal component analysis for the vectors from Step 3.
5. Observe the scatter plots of the principal components.

4. Settings & Results









Dataset

- English, Japanese, and Chinese text of the New Testament
- Demonstrate on three facets: character, word, and syntax
- Syntax: POS tags, tagged by NLTK in Python

1) Cumulative contribution rate
- PC1 + PC2 > 90% of the whole result

2) Eigenvalues
- PC1→an overall complexity
- PC2→PC2 = individual complexities

3) Pair plot 1
- All languages similarly scatter on PC1; they form a slight stripe on PC2

4) Pair plot 2
- Two clusters appear when the dataset is arranged (Nakayama, 2023)

5. Discussion

- All languages have a similar variation on PC1.
→ The overall complexities of languages are similar.
- English texts have a positive eigenvalue on PC2.
→ Character complexity ≥ Syntactic complexity.
- Chinese texts have a negative eigenvalue on PC2.
→ Character complexity ≤ Syntactic complexity.

6. Conclusion and Caveats

- Languages have at least a similar overall complexity, while individual facets have different degrees of complexity.
- The sequence of each facet is not exclusive but includes information about the others
(e.g., character strings does not only represent character complexity itself but also morphological and syntactic complexity).

Acknowledgment: This work is supported by JST SPRING, Grant Number JPMJSP2123.

References:
Bentz, C., Gutierrez-Vasques, X., Sozinova, O., & Samardžić, T. (2022). Complexity trade-offs and equi-complexity in natural languages: A meta-analysis. *Linguistics Vanguard.* https://doi.org/10.1515/lingvan-2021-0054.
Deutcher, G. (2009.). "Overall complexity": A wild goose chase? In G. Sampson, D. Gil, & P. Trudgill (Eds.), *Language complexity as an evolving variable* (pp. 243–251). Oxford University Press.
Nakayama, T. (2023 June 29). Are All Languages Equally Complex?: Information Theory-based Method to Measure the Overall Complexity of a Language [poster]. Quantitative Linguistics Conference 2023, Lausanne, Switzerland.
Hacket, C. B. (1958) *A course of modern linguistics.* Macmillan.
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