Are All Languages Equally Complex?: Information Theory-Based Method to Measure the Overall Complexity of a Language

Takuto NAKAYAMA (Keio University, Tokyo, Japan)

1. Purpose

- To propose a measurement for an "overall" linguistic complexity while:

i) Considering multiple linguistic facets

ii) Dealing with the facets in the same way

3. Setting

- This study deals with four facets – character, word, syntax, and semantics.
  - Syntax: POS tags by NLTK in Python.
  - Semantics: Tags by PyMUSAS in Python (Piao et al., 2016).
- It compares two versions of the New Testament (cf. Ehret and Szmrecsanyí, 2016):
  - Simple (not edited)
  - Complex (10% of the characters are randomly substituted)

2. Methodology

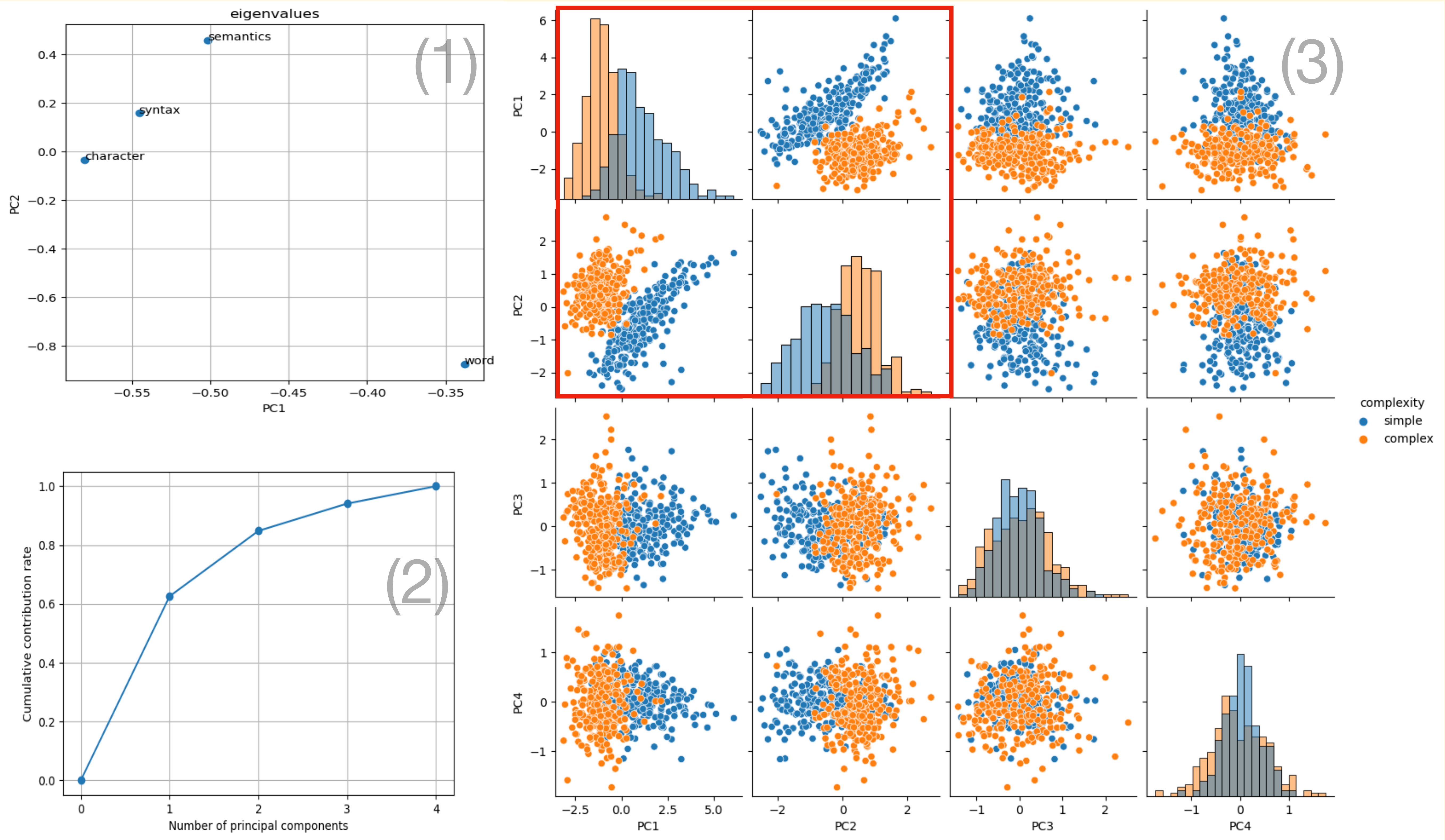
1. Calculating the entropy H (Shannon, 1948) of an n-element sequence from the ith to jth character of a document, the length of which is l:

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

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

2. Determining 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. Applying 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. Conducting principal component analysis for the vectors from Step 3.
5. Observing the scatter plots of the principal components.

4. Results



(1) **Eigenvalues**
All eigenvalues are negative for the 1st principal component (PC1); only "word" was negative for the 2nd one (PC2).

(2) **Contribution rate**
The PC1 and PC2 explain more than 80% of all information in the results.

(3) **Pair plot of PC**
Each point refers to each chapter of the New Testament. PC1 and PC2 (highlighted with a red square) are the focus.

5. Discussion

- PC1 seems to represent an overall complexity: positive means simpler and negative means more complex, because negative PC1 refers to that exponents are closer to 0 (=complex), while positive PC1 refers to those farther from 0 (=simple).
- Most of orange dots (complex texts) are negative for PC1; about half of the blue dots (simple texts) are positive.
- The other half of blue dots on negative are also negative on PC2, which means their word complexities are higher than orange dots.

6. Conclusion

- This method can provide visualization in which the text is simpler or more complex than the other.
- This method can deal with any number of facets users want to consider with the unified process, as long as the facets can be described in a sequence; thus, it could be useful for cross-linguistic research.
- A limitation of this method is that it cannot provide a micro viewpoint of each facets. For example, regarding syntax, this method cannot deal with dependent structure, which Bentz et al. (2022) focused on.

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*, *0*(0). https://doi.org/10.1515/lingvan-2021-0054.

Deutcher, G. (n.d.). "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.

Ehret, K., & Szmrecsanyi, B. (n.d.). An informationtheoretic approach to assess linguistic complexity. In R. Baechler & G. Seiler (Eds.), *Complexity, isolation, and variation* (Vol. 57, pp. 71–94).

Piao, S. S., Bianchi, F., Dayrell, C., D'egidio, A., & Rayson, P. (2015). Development of the multilingual semantic annotation system. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* (pp. 1268–1274).

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

Contact: tnakayama.a5ling@gmail.com