Original Article

Is Language Diachronically Equally Complex? : An Information Theory Approach to Linguistic Complexity

Takuto Nakayama

Faculty of Health Sciences, Kyorin University (part-time).

5-4-1 Shimorenjaku, Mitaka, Tokyo. 181-0013.

email : tnakayama.a5ling@gmail.com

Abstract

The equi-complexity axiom that no language on the Earth is simpler or more complex than others has been an unchallenged dogma in linguistics. Only a few linguists have examined whether this axiom is true until the end of the 20th century. Nowadays, technological development, (e.g., computers), accelerates the movement in which linguists attempt to clarify the eui-complexity axiom. However, few studies focus on a diachronic perspective of linguistic complexity. This study tackles the diachronic perspective of linguistic complexity and provides suggestions regarding whether equi-complexity is a general axiom of language. This study focuses on the entropies of Shannon's information theory of each character in Modern English and Middle English documents. As a result, Modern English and Middle English seem to have similar linguistic complexity. It is suggested that, although linguistic structures (e.g., vocabulary and grammar) change over time, linguistic complexity does not change dramatically.

Keywords : *linguistic complexity, information theory, diachronic perspective*

1 Introduction

Linguists have generally accepted that no language on the Earth is simpler or more complex than others ; if it has very complex aspects, such as syntactical or morphological structures, it always has other aspects that are much simpler, a phenomenon called equi-complexity. This statement had been an unchallenged dogma in linguistic research ; however, some linguists have examined whether this statement is really true using statistics or computational methodologies invented since the beginning of the 20th century. For example, using computer simulations, Reali et al. (2018) found that compared to smaller groups of people, the languages of larger groups have simpler linguistic struc-tures, whereas their vocabulary systems are more complex, which showed equi-complexity seemed to be sat-isfactory. However, only a few studies have focused on the diachronic

point of view of linguistic complexity.

Most research examining equi-complexity has focused on the differences between language varieties close to each other, but not on diachronic changes in linguistic complexity.

This study tackles the diachronic perspective of linguistic complexity and provides suggestions regarding whether equi-complexity is a general axiom of language. If the linguistic complexity of a language is stable, despite changes in individual aspects of the language over time, such as the grammatical features and vocabulary, the degree of complexity of the language must be kept by any entity, which could be one of the most basic characteristics of language. By contrast, if the complexity of a language varies over time, it can provide a counter-example against equi-complexity, posing a new question: How and why does linguistic complexity change over time? A diachronic view of linguistic complexity is likely to be a hot topic, not only for linguistic complexity itself but also for general features of languages.

2 Related Research on Linguistic Complexity

2.1 Equi-complexity Axiom

During the 20th century, linguists believed that every language was equally complex, a concept referred to as the equi-complexity axiom. This states that "in 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). For example, even if language A has a simpler syntactic structure than language B, the morphology of language A is more complex than that of B, which balances the complexities of the languages as a whole. Equi-complexity has been an unchallenged dogma "almost acquiring (more implicitly than explicitly, though) the status of an established truth" (Kortmann and Schlöter, 2020).

Edward Sapir was the first researcher to mention this idea in his book, in which he states, "[b]oth 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, 104). Even in the present day, many linguists agree with the axiom to some extent. Daniel Everett, who studies Pirahã people living in Amazon and their language claimed, in his paper, that "[n]o 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). Further, McMahon and McMahon (2013) pointed out that "not having words for vaccine or satellite television does not make one language less structurally

complex than others, just as not having the things these words refer to does not make a human group less highly evolved, reflecting instead simply the nature of its society" (7).

However, since the end of the 20th century, a movement of linguists has been attempting to examine whether the axiom is really true. This has been accelerated by the development of computer technology, which enables us to manipulate a massive amount of data larger than manually possible (Kortmann and Schlöter, 2020). Until such a technological revolution, the equi-complexity of language could not be examined. For instance, some linguists have used computer simulations to illustrate how linguistic complexities behave in communities with different populations (e.g., Reali et al., 2018; Vogt, 2007), which will be introduced in the next section. Equi-complexity axioms are now being revisited as one of the hottest topics in linguistic research.

2. 2 Linguistic Complexity and Community Size

One of the major determinant factors of linguistic complexity is the size of the community. Milroy and Milroy (1985) claimed that the network structure of the community is significant for how the language of the community changes. They asserted that there are more innovative changes in a loose-tied society than in a tight-tied society in which conservative linguistic usages are more likely to be maintained. This idea was inherited from such a research area as sociolinguistic typology. Trudgill (2011) asserted that "the issue at hand of this sociolinguistic typology is whether it is possible to suggest that certain linguistic features are more commonly associated with certain types of society or social structure than others" (xvi). This research field focuses on the correlation between the typology of society in which a certain language is spoken and the complexity of the language or its varieties (Shibuya, 2022).

Based on this research, recent linguists have used a statistical approach to examine this topic. Lupyan and Dale (2010) found that in larger groups languages have simpler inflectional morphology than those spo-ken in smaller groups by statistically examining more than 2,000 languages in the world. They hypothesized that "language structures are subjected to different evolutionary pressures in different social environments. ... [L]anguage structures appear to adapt to the environment ... in which they are being learned and used" (Lupyan and Dale, 2012, 1). Some studies (e.g., Reali et al., 2018; Vogt, 2007, etc.) have also approached this topic using computational methods, while other strands of research (e.g., Raviv et al., 2019) have used experimental methods. Reali et al. (2018) used computer simulations to demonstrate that languages of larger groups have simpler linguistic structures, but more complex lexicons than those of smaller groups. Raviv et al. (2019), through experimental methods, found that in a larger group

the language developed faster and more consistently than in a smaller group.

However, the characteristics of linguistic complexity introduced above have yet to offer a sufficient explanation for certain aspects. For example, they have not explained why a language is conservative and less learnable because it might be impossible to avoid innovation by newcomers even in a closed small community. Newcomers, such as newborn babies or foreign people, who can innovate the language, are consistently added to even closed societies. As Kirby et al. (2008) reported, the accumulation of the generations may lead a language to a more learnable form in experiments in which participants try to learn a nonce language that consists of strings of alphabets as its forms and sets of arrows and shapes as its meanings from the performance the prior participants had done as the previous generations. Thus, the observation that the size of the community affects the complexity of its language requires more detailed research.

3 Methodologies

3. 1 Shannon's Information Theory

3. 1. 1 Mathematical Measurement of Complexity

In previous studies, linguistic complexity was measured by observing the extent of complexity of each aspect of language, such as the number of constituents or varieties of phonemes. For example, the chart below is a set of indices of complexity.

1. Epistemic modes A. Formulaic complexity a. Descriptive complexity: length of the account that must be given to provide an adequate description of a given system. b. Generative complexity: length of the set of instructions that must be given to provide a recipe for producing a given system. c. Computational complexity: amount of time and effort involved in resolving a problem.
2. Ontological modes A. Compositional complexity a. Constitutional complexity: number of constituent elements (e.g., in terms of the number of phonemes, morphemes, word, or clauses) b. Taxonomic complexity (or heterogeneity): variety of constituent elements, that is, the number of different kinds of components (e.g., tense-aspect distinctions, clause types). B. Structural complexity a. Organizational complexity: variety of ways of arranging components in different modes of interrelationship (e.g., phonotactic restrictions, variety of distinct word orders). b. Hierarchical complexity: elaborateness of subordination relationships in the modes of inclusion and subsumption (e.g., recursion, intermediate levels in lexical-semantic hierarchies).
3. Functional complexity A. Operational complexity: variety of modes of operation or types of functioning (e.g., cost-related differences concerning the production and comprehension of utterances). B. Nomic complexity: elaborateness and intricacy of the laws governing a phenomenon (e.g., anatomical and neurological constraints on speech production; memory restrictions).

Figure 1 Baechler and Seiler (2016 : 4)

However, observing language based on a chart of indices is likely to be subjective and biased. This article, therefore, will exploit a mathematical method to measure complexity using Shannon's information theory. Shannon's information theory (cf. Shannon, 1948) focuses on what information is, how to mathematically measure information, and what kind of mathematical structure information has (Amari, 1970). Shannon's work attempts to make information transmission more efficient.

In order to introduce Shannon's information theory, a few technical terms must first be introduced: information value and entropy. An information value I is a value where the information about something that occurs at the probability p happens, which is defined as follows:

$$I = -\log_2 p \tag{1}$$

Equation (1) states that the less frequently the phenomenon occurs, the more valuable is the information that the phenomenon will occur.

Entropy H is the expected value of a sequence of information values. Hence, a phenomenon X_i $(1<=i<=n)$ occurs at the probability p_i, the entropy $H(X)$ is:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{2}$$

In the case of language, a phenomenon corresponds to a linguistic unit, such as a character or a word. Here, we refer to a phenomenon as a character, and a sequence of phenomena as a sentence. In fact, materials larger than a sentence, such as a novel or a document, also can be examined using this equation. Thus, the term, sentence, refers not only to one sentence, from one period to the next period, but also to larger units.

This study used entropy to evaluate the complexities of languages. However, the definition of entropy offered above considers that a character appears alone, independent from the others, which is far from a sentence in the real world. In real usage, a character in a sentence may be predictable from another character in the sentence, which affects and reduces the information value of the character. Further, from the definition above, we are not able to distinguish the difference between the entropy of a sentence and that of a string of characters that consists of each alphabet used at exactly the same time as the sentence. In the next section, advanced definitions of entropy are introduced.

3. 1. 2 Mathematical Definition of Entropy

In order to consider the correlation between other characters, Equation (2) requires the input of the probability of multiple characters. Considering the correlation between a character and one right after the character, the probability that a string of characters, $x_ix_j (1 \leqq i < j \leqq n)$ occurs is required to calculate the entropy that includes the correlation of two characters.

Thus, the equation for the correlation between two characters can be generalized as follows:

$$H(x^n) = -\sum_{i=1,j=n}^{1-n+1,l} p(x_{ij}) \log_2 p(x_{ij}) \tag{3}$$

in which n refers to the number of a sequence of characters in question and l to the length of the sentence. x_{ij} refers to a string from the ith character to the jth one. i and j always satisfies $(1 \leqq i \leqq l - n + 1)$ and $(n \leqq j \leqq l)$, respectively. $j - i$ is always $n - 1$, and p_{ij} corresponds to a probability at which a sequence from the ith character to jth character of the sentence. Theoretically, the number of characters in a sentence produced can reach at some time the infinite as n in Equation (3) is increasing. Thus, a theoretical entropy

$$\begin{aligned} H &= \lim_{n \to \infty} H(X^n) \\ &= \lim_{n \to \infty} -\sum_{i=1,j=n}^{l-n+1,l} \frac{p(x_{ij}) \log_2 p(x_{ij})}{n} \end{aligned} \tag{4}$$

Given that this is only theoretical because we cannot observe the infinity number of characters this study uses a more practical equation based on Equation (3).

3. 2 Dataset

This study used four documents, including two written in Modern English and two in Middle English: *The Adventures of Sherlock Holmes, The Origin of Species, Sir Gawain and the Green Knight, and The Canterbury Tales.* All the documents are from Project Gutenberg. The information on these documents is summarized in the chart below.

Types of Alphabet refers to how many types of the alphabet are used in a document, including punctuational symbols, such as hyphens and periods. The Middle English documents also include a character that is not used in current English, þand ë, for example. The documents were chosen because they were written in approximately the same year in each period, and read by many people, which suggests that ways of writing of these documents were widely accepted. In addition,

	The Adventures of Sherlock Holmes	The Origin of Species
Year	1892	1859
Author	Arthur C. Doyle	Charles R. Darwin
Size	561,643 characters	950,569 characters
Types of Alphabet	87	91
Abbreviation	Holmes	Species
	Sir Gawain and the Green Knight	**The Canterbury Tales**
Year	around 1400	1370–1400
Author	Geoffrey Chaucer	Unknown
Size	111,534 characters	1,021,833 characters
Types of Alphabet	81	91
Abbreviation	Gawain	Canterbury

Table 1 The Overview of the Dataset

this study aimed to observe genre-independent features. Thus, in the case of Modern English, a novel and a scientific book were chosen, whereas both documents in Middle English are stories written in verse because there were few literature genres at that time. In the following, the four documents are described using the abbreviations in the chart above.

3. 3 Procedure

The procedure follows three steps: abstracting patterns of characters from a document, calculating the probability of an occurrence of a character, and calculating the entropy. In Step 1, all the patterns in a document that consisted of a certain number of characters being set in advance were abstracted. For example, from the sentence, "abcdefg", six patterns of two characters can be abstracted as "ab," "bc,", "cd," "de," "ef," and "fg," or the five patterns of three characters: "abc," "bcd," "cde," "def," and "efg". In fact, we can make $52^2 (= 2704)$ patterns in the case of two alphabets, including both upper and lower cases. However, the probabilities of patterns that do not appear in the document are 0, which does not have any effect on the calculation of entropy. Thus, in this study, patterns that did not exist in the document were ignored. In Step 2, the probabilities of occurrence of each pattern were calculated with how many times each pattern occurs divided by the number of all the patterns in the document. For instance, in the case of the sentence, "abcdefg" and the six patterns of two characters, the probability of "ab" is $1/6 \fallingdotseq 0.167$. The general form of the probability of a pattern consisting of n characters can be described as follows:

$$\frac{(\text{occurrence of the pattern})}{(\text{length of the whole sentence}) - n + 1} \tag{5}$$

In Step 3, finally, entropy is given from Equation (3) and the results of the previous steps.

The three steps were applied to the case of 1-character to 100-character patterns. In order to see how different or how close the linguistic complexities of the documents or the language in a certain period were, this study observed how the entropies of the documents behaved as the number of characters of the patterns increased. We also observed how the differences between the averaged entropies of Modern and Middle English behave with an increasing number of characters.

4 Results

4. 1 Modern English

The results of the analysis of Modern English documents are shown in Figure 2-5. Figure 2 shows the entropies of Holmes, and Figure 3 shows the ones of Species. Both are plotted together in Figure 4, in which the two are drawn with a dotted curve for easy identification. Figure 5 shows the absolute values of the difference in the entropy of Holmes and the one of Species at each number of characters in the pattern. As shown by the figures, the entropies of the two documents are quite similar in almost all numbers of characters. Observing the differences between those entropies, they had the largest gap in the 6-character patterns, whereas they were closest in the 9-character patterns, both of which were not quite large. The ratio of the average differences between the entropies to the average of the entropies was nearly 0.026, which could mean that both documents have almost the same linguistic complexity.

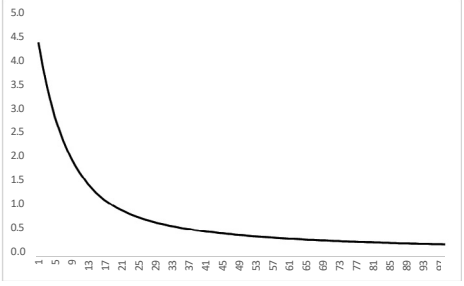

Figure 2 Holmes

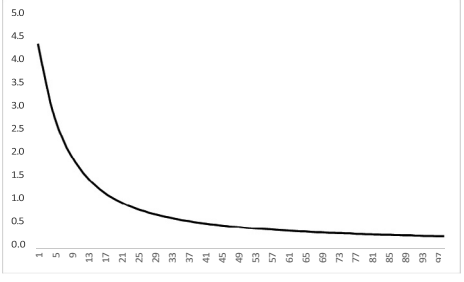

Figure 3 Species

4. 2 Middle English

Figure 6-9 show the results of Middle English document analysis. Figure 2 shows the entropies of Gawain, and Figure 3 shows those of Canterbury, which are plotted together in Figure 8. Similar to

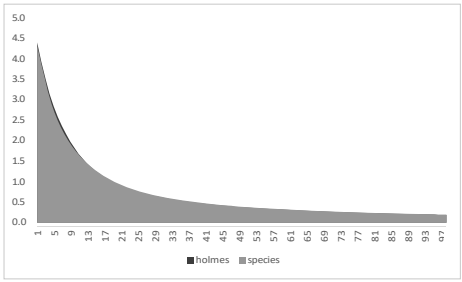

Figure 4 Modern English

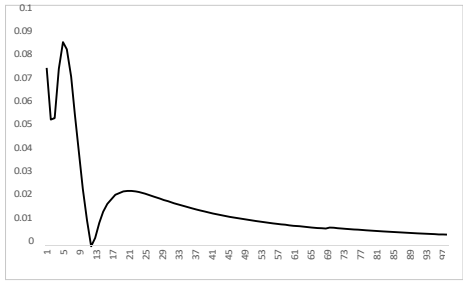

Figure 5 */H (holmes)−H (species)/*

Figure 5, Figure 9 shows the absolute values of differences in the entropies of the documents. Gawain and Canterbury have a relatively large gap around the 10-character pattern. This was observed from the differences of the entropies, Figure 8, in which the curve becomes the highest at around 10 characters. Although the documents had slightly larger differences than the case of Modern English, both entropies decreased in a similar way as the number of characters increases.

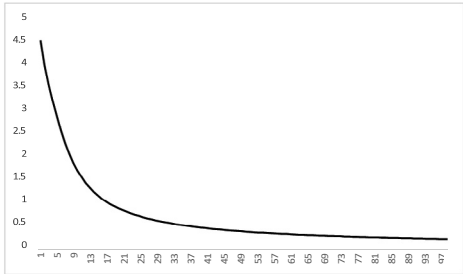

Figure 6 Gawain

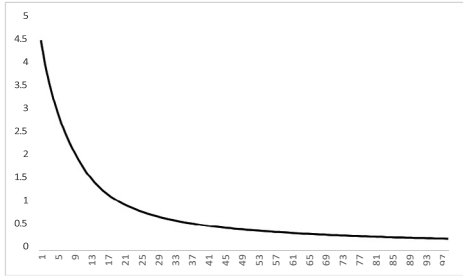

Figure 7 Canterbury

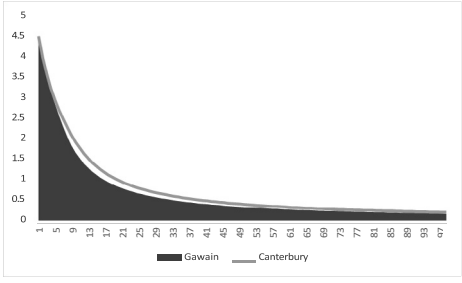

Figure 8 Middle English

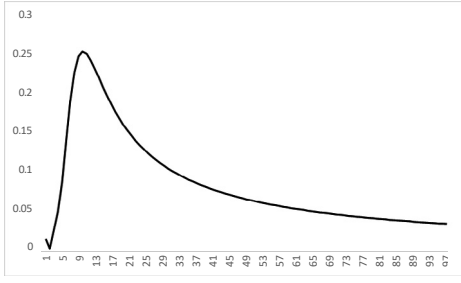

Figure 9 */H (gawain)−H (canterbury)/*

4. 3 General Analysis

Figure 10 shows the absolute values of differences between the average entropies of the Modern

English documents and of those of Middle English documents. Figure 11 describes the ratios of the values in Figure 10, based on the average entropies of Modern English and Middle English documents. As Figure 10 shows, when the number of characters in the pattern is 9, the difference between the entropy of Modern English and Middle English documents is smallest. As the difference becomes large, the entropy starts decreasing as the number of characters increases. This tendency is also observed in Figure 11, in which the ratio of the difference of the entropy to the average entropy is the smallest at the 9−character pattern. Further, as the number of characters increases, the ratio is maintained at a certain value, around 0.06.

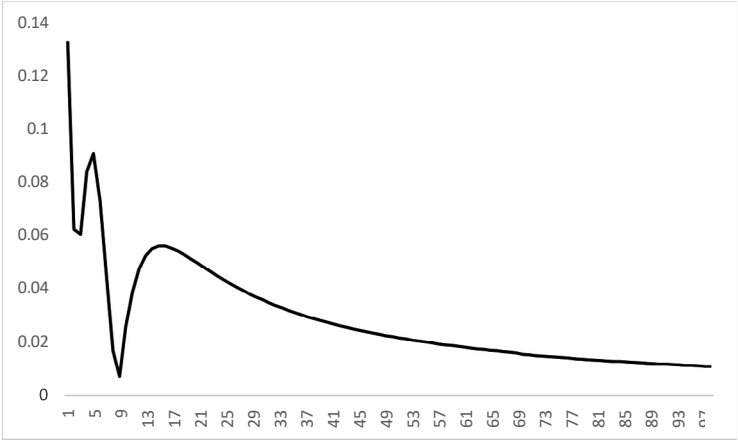

Figure 10 $/H(ModE) - H(ME)/$

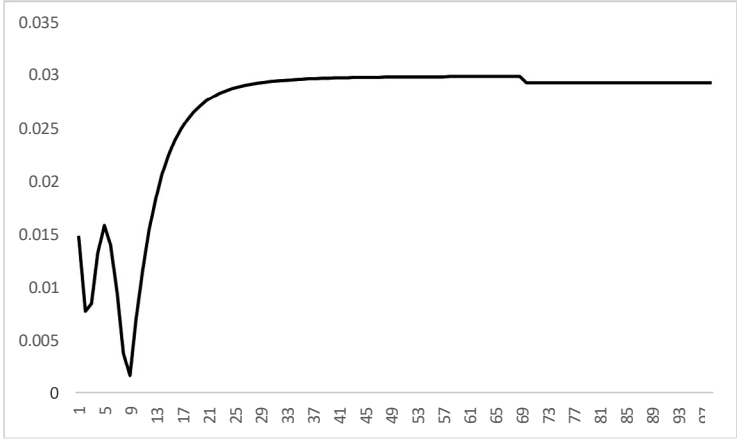

Figure 11 $\dfrac{/H(ModE) - H(ME)/}{H(ModE) + H(ME)}$

5 Discussion

The results show that linguistic complexity does not seem to vary over time. Observing the differences in the entropies of Modern English and Middle English, the differences are quite close with patterns of every length of characters. This can be seen in Figure 10, in which the differences generally decrease as the number of characters increases. The difference in the 1-character pattern is the largest, probably because the distribution of alphabets is different between Modern English and Middle English, which might be caused by the difference in grammar and vocabulary. Such a difference seems to be balanced as the pattern in question becomes longer. However, the entropy of the documents does not simply get closer to each other as the number of characters increases but reaches the closest condition at a certain number of characters that is relatively small. In this case, it is a 9-character pattern. The entropy of all the documents simply decreases as the number of characters increases. In fact, as the length of the characters of the pattern increases, fewer matches can be found in the patterns in the document, which make the entropy decline. However, as Figures 10 and 11 show, the lowest entropy is observed with the 9-character pattern, at least from 1 to 100-character pattern. This suggests that both Modern English and Middle English have a similar distribution of 9-character patterns in their documents. Although a concrete conclusion cannot be provided from only this research, this finding might indicate that even when vocabulary and grammar change, the distribution of patterns consisting of a certain number of characters never changes, and this is one of the factors that keeps a language equally complex over time.

Figure 11 also shows that the ratio of the differences in the entropies of Modern English and Middle English to the average entropies stays at almost the same value, around 0.06, after more than a certain number of characters. This suggests that for patterns beyond a certain length of characters, the entropies of both Modern English and Middle English decrease at almost the same rate. This tendency could be caused by technical factors, one of which might be decreasing matches to a pattern with increasing the length of characters, but, with the smallest difference of the entropies. Thus, further research is necessary to draw a concrete conclusion regarding this issue.

6 Conclusion

This study examined whether linguistic complexity changes over time through an information theory approach. We can conclude that linguistic complexity does not appear to change

dramatically over time. The entropies of Modern English and Middle English behave, although not in exactly the same manner but in a quite similar way as the number of characters of the pattern increases. Moreover, the closeness of linguistic complexity over time can also be observed from the fact that the differences between the entropies of each period's English are quite small. The entropy of the documents does not simply get closer to each other as the number of characters increases but reaches the closest condition at a certain number of characters that is relatively small. The smallest difference between the entropies is observed in a pattern with a relatively small number of characters—a 9-character pattern in this study. L This suggests that although linguistic structures change, the distribution of patterns seems to stay in almost the same condition. This suggests that the equi-complexity of a language remains over time.

One of the limitations of this study is a shortage of data. Future studies include more documents, which will make it possible to cope with more precise statistical tests. An unsolved issue in this study, whether the distribution of patterns is the same over time on a certain number of character patterns, can be examined by massive data that could offer a better generalization of features. As a pilot study of equi-complexity from a diachronic point of view, the findings will be complemented by further research.

References

Amari, T. (1970). *Jouhourironn* [Information theory]. Daiamond Sha.

Baechler, E., & Seiler, G. (eds.) (2016). *Complexity, isolation, and variation*. Walter de Gruyter.

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

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory : An experimental approach to the origins of structure in human language. Proceedings of the National Academy of Sciences, 105(31), 10681-10686. https : //doi.org/10.1073/pnas.0707835105

Kortmann, B., & Schröter, V. (2020). Linguistic complexity. (https : //www.oxfordbibliographies. com/view/document/obo-9780199772810/obo-9780199772810-0254.xml).

Kortmann, B., & Szmrecsanyi, B. (Eds.). (2012). *Linguistic complexity : Second language acquisition, indigenization, contact*. Dd Gruyter. https : //doi.org/10.1515/9783110229226

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE*, 5(1), e8559. https : //doi.org/10.1371/journal.pone.0008559

McMahon, A., & McMahon, R. (2012). *Evolutionary linguistics* (1st ed.). Cambridge University

Press. https : //doi.org/10.1017/CBO9780511989391

Milroy, J., & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Cambridge University Press*, 21(1), 339–384.

Raviv, L., Meyer, A., & Lev–Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B* : Biological Sciences, 286(1907), 20191262. https : //doi.org/10.1098/rspb.2019.1262

Reali, F., Chater, N., & Christiansen, M. H. (2018). Simpler grammar, larger vocabulary : How population size affects language. *Proceedings of the Royal Society B* : Biological Sciences, 285 (1871), 20172586. https : //doi.org/10.1098/rspb.2017.2586

Sapir, E. (1921). *Language : An introduction to the study of speech.* Harcourt, Brace.

Shibuya, K. (2022). Gengo no fukuzatusei no genjo [The current circumstance of research on linguistic complexity]. *Handai Shakaigengogaku Kenkyunoto* [University of Osaka Sociolinguistics Research Notes], 18, 119–144.

Trudglii, P. (2011). *Sociolinguistic typology : Social determinants of linguistic complexity.* Oxford University Press.

Vogt, P. (2007). Group size effects on the emergence of compositional structures in language. In F. Almeida e Costa, L. M. Rocha, E. Costa, I. Harvey, & A. Coutinho (Eds.), *Advances in Artificial Life* (Vol. 4648, pp. 405–414). Springer Berlin Heidelberg. https : //doi.org/10.1007/978–3–540–74913–4_41