European Journal of Statistics and Probability
Vol.10, No.2, pp., 14-28, 2022
Print ISSN: 2055-0154(Print),
Online ISSN 2055-0162(Online)

Morphosyntactic Complexity in Old Japanese

Wenchao Li
Zhejiang University

Citation: Li W. (2022) Morphosyntactic Complexity in Old Japanese, European Journal of Statistics and Probability, Vol.10, No.2, pp., 14-28

ABSTRACT: *Old Japanese (592–794 AD) had a uniquely complex writing system: variant Chinese; classical Chinese; man'yōgana; senmyoo gaki. This study takes a mathematical linguistic approach, employing word length and dependency distance as metrics of the lexical and syntactic complexity of Old Japanese. We find that the distribution of Japanese dependency directions is balanced, indicating that Japanese is neither a strongly head-initial nor strongly head-final language. Neither an advcl relation nor a cc relation are detected, suggesting that syntactic structure in Old Japanese is simpler than Modern Japanese. Among all the dependency relations, 46.3 per cent were of an adjacent relationship, rendered by case, mark, and det (with DD = 1), while nsubj, advmod, obl, and acl were long-distanced and presented a diverse range, with nsubj, for example, ranging from 1 to 29. Mean dependency distance and frequency fit a power law function ($y = ax^b$) well. Among texts, Senmyōgaki bears a relatively short mean word length, while Kojiki presents the longest word length. The mean word length-frequency distributions of Bussokusekika and Fudoki fit the Cohen-binomial model and Senmyō fits the Palm-Poisson model. The distribution of mean word length and their frequencies supports Zipf's (1949) principle of least effort: shorter words tend to be more frequently used.*

KEYWORDS: Morphosyntactic, complexity, old, Japanese

INTRODUCTION

Modern Japanese is phonologically moraic (1a) and morphologically agglutinative (1b).

(1) a. **moraic**

L H L
| | |
ku ji ra

b. **agglutinative**

話-させ-られ-まし-た-か

speak (stem)-causative-passive voice-honorification-tense. past-question marker

Syntactically, Japanese is alleged to have canonical SOV word order, e.g. (2).

(2) Taroo-ga-sushi-o-tabeta
Taroo-NOM-sushi-ACC-eat-PAST
'Taroo ate sushi.'

Intriguingly, Japanese word order appears to be free. A benefactive event, 'Taroo gave an apple to Jiroo', can be encoded into six expressions, cf. (3):

(3) **Free word order**
a. Taroo ga Jiroo ni ringo o ageta. b. Taroo ga rongo o Jiroo ni ageta.
c. Jiroo ni Taroo ga ringo o ageta. d. Jiroo ni ringo o Taroo ga ageta.
e. Ringo o Taroo ga Jiroo ni ageta. f. Ringo o Jiroo ni Taroo ga ageta.

The syntactic feature of relatively free word order can be traced back to Old Japanese. Yanagida (2014) demonstrates that when the subject and object in a clause are both marked, their order is free. Examining Old Japanese further reveals a 'negation auxiliary-verb' order. (4) is extracted from *Man'yōshū*, in which the negation 不 '*zu*' independently antecedes the verb.

(4) Man'yōshū.1.16
山 乎 茂 入而 毛 **不取**
yama *wo* sige-mi irite *mo* tora-zu
mountain ACC grow thick-NMLZ come.GER FOC NEG.pick

The negation auxiliary was confirmed to antecede the verb in 1,438 tokens in the corpus, which demonstrates a non-agglutinative character. This morphological feature prompts us to consider whether Old Japanese presents a different dependency direction.

Old Japanese is a dead language spoken during the Asuka (592–710 AD) and Nara periods (710–794 AD). Chinese characters were borrowed to represent vernacular Japanese in writing. Three writing systems co-existed: (a) Mixed Chinese-Japanese script, termed *hentai-kanbun* 'variant Chinese', which refers to a script that combines Chinese and a phonetic transcription of Japanese. A representative work is *Kojiki*, the

oldest extant chronicle in Japan. (b) Classical Chinese. A representative work is *Nihonshoki*, the second oldest book of classical Japanese. (c) *Man'yōgana,* which borrows Chinese characters in three ways, i.e. solely borrowing semantic value, solely borrowing phonological value, and borrowing both semantic and phonological values. A representative work is *Man'yōshū*, the oldest collection of Japanese poetry. The complexity of the writing system of Old Japanese raises the question of how lexically and syntactically Old Japanese is characterised. Moreover, the affiliation of Japanese has long been a matter of debate among linguists both in and outside Japan. For example, Japanese has been argued to originate from the Korean language or Polynesian languages (Ohno 1957); it has also been argued to be a member of the Altaic family (Hattori 1959) or Dravidian family (Ohno 1981), a mixture of the Tungus and Austronesian language families (Matsumoto 2007), or as having developed in the West Liao River region in the Early Neolithic and dispersed to the Korean peninsula and to the Japanese islands in the Late Neolithic and Bronze Age (Robbeets et al. 2021). A diachronic examination of morphosyntax in early Japanese might shed light on the affiliation of the language.

To this end, a mathematical analysis is conducted, taking word length and dependency distance as metrics of the lexical and syntactic complexity of Old Japanese. The paper is organised as follows. Section 2 outlines the framework and methodology (incl. the corpora and syntactic parser), and Section 3 addresses syntactic and lexical complexity. Section 4 concludes the paper.

FRAMEWORK AND METHODOLOGY

Dependency grammar

Dependency grammar was initially advanced by Tesnière (1959) and developed by Hudson (1990, 2007). Liu (2008) employs dependency distance (DD) as an insightful metric to measure syntactic complexity. The examples below, drawn from *Man'yōshū* 1.2., were written in *man'yōgana*:

(5) *Man'yōshū*.1.2.

海原波　　unapara *pa*

加萬目立多都　*kamame* tati-*tatu.*

The annotation was:

(6)

Sentence number	dependent			head			Dependency type
	Order number	Word	Pos	Order number	Word	Pos	
S1	1	unapara	Noun	4	tati-tatu	Compound verb	nsubj
S1	2	pa	Case	1	unapara	Noun	case

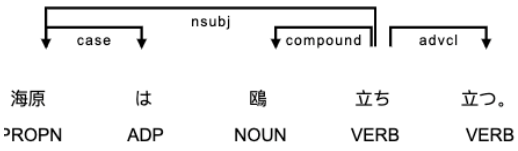


Figure 1. Dependency relation and direction

There are four dependency relations in (6). Among them, the verb 立ち is the GOVERNOR, and all elements are connected via a 'governor-dependent' relationship. DD refers to the linear distance between a word and its governor. The calculation of DD is aligned with Liu, Hudson, and Feng's (2009) insights. The words in a sentence are assigned in a string (i.e. W_1 ...W_i ...W_n). Regarding any dependency relationship between words W_a and W_b, if W_a is a governor and W_b is its dependent, then the DD between the two words is |governor − dependent| (the absolute value). The MDD of the whole sentence would then be:

$$MDD = \frac{1}{n}\sum_{i=1}^{n} |DD_i|$$

Calculation

MDD and word length (MWL) are calculated in the following steps:

Step 1: Draw raw data from the corpora

Step 2: Parse each sentence via the GiNZA v4 Parser (National Institute for Japanese Language and Linguistics, and Megagon Labs)

Step 3: Calculate the MDD from the parsed outputs

Step 4: Produce a computer programme to calculate word length

Data

This study's central goal is to explore the lexical and syntactic complexity of Old Japanese. To this end, distributions of dependency distance and word length are employed as metrics. Data are drawn from myth, chronicle, poetry, and imperial edicts. A detailed list of materials is provided in Table 1.

Table 1. Materials

Historical period	Materials	Genres
Nara Period (710-794)	Kojiki 712	Myth
	Fudoki 713	Poetry
	Nihonshoki 720	Chronicle
	Bussokuseki poetry 753	Poetry
	Man'yōshū 759	Poetry
	Shoku Nihongi senmyō 797	Imperial edict

RESULTS AND DISCUSSIONS

Syntactic complexity

Dependency relation types and the mean dependency distance of texts written in four writing styles (variant Chinese; classical Chinese; *man'yōgana*; *senmyoo gaki*) are presented in Table 2.

Table 2. Dependency relations in the Nara period (710–794)

Dependency relation	Dependency distance	Mean dependency distance	Percentage
case	1	1	36%
compound	1, 2, 3	2	26%
mark	1, 2	1.5	8%
nmod	1, 2, 3	2	6%
nsubj	1, 2, 3, 4, 6, 13	4.8	5%
obl	2, 3, 4, 7, 9	5	4%
advmod	1, 2, 3, 5	2.7	4%
obj	2, 3, 5	3.3	3%
acl	1, 3, 6	3.3	3%
aux	1, 2	1.5	3%
det	1	1	1%

Further results in Tables 3 and 4 demonstrate that the distribution of Japanese dependency directions appears to be balanced, suggesting that Japanese is neither a strongly head-initial nor strongly head-final language. Crucially, neither the *advcl* relation nor the *cc* relation are detected, suggesting that the syntactic structure in Old Japanese is simpler than Modern Japanese. Among the 11 dependency relations, *compound* had the second largest attribution (26%), and its DD ranged from 1 to 3. This might be linked to the arbitrary combinationality of multiple verbs in Old Japanese, where multiple verbs are combined via verb serialising. The following nine combinations were confirmed: Tran. + Tran. + Unacc.; Tran. + Tran. + Tran.; Tran. + Unacc. + Unacc.; Tran. + Unacc. + Unacc.; Tran. + Tran. + Unerg. + Unacc.; Unacc. + Unacc.; Unacc. + Unerg.; Tran. + Tran.; Unerg. + Unerg.

Table 3. MDD, dependency direction, and dependency relations in Old Japanese

Genres	Head initial	Head final	MDD	Dependency relation types
Myth Kojiki 712	0.46	0.54	2.28	11
Chronicle Nihonshoki 720	0.40	0.60	3.0	9
Poetry Fudoki 713	0.46	0.54	1.26	10
Poetry Bussokuseki 753	0.44	0.56	1.62	9
Poetry Man'yōshū 759	0.46	0.54	1.19	10
Imperial Shoku Nihongi Senmyō 797	0.41	0.59	1.20	12

Table 4. Dependency direction in Old Japanese

Genres	case	mark	nsubj	obj	acl	aux	compound	nmod	advmod	acl	det	obl
Kojiki	1	0	0	-2	0	1, 2	-1	-2	-1	0	-1	-4, -7, -9
Nihonshoki	1	1, 2	-3, -6, -13	0	0	1, 2	-1, -23	-1, -2, -3	-1, -2, -3	0	-1	
Manyooshuu	1	1	-3, -5	-5	-1, -3	1	-1	-2, -3	0	-1, -3	-1	-7, -9
Senmyoo	1, 2	1	-29	-3	-1, -3, -6	1	-1, -2, -3	-2, -3	-5	-6	0	0

A closer picture addressing dependency distances and their frequencies in *Kojiki*, *Nihonshoki*, *Man'yōshū*, and *Shoku Nihongi Senmyō* are presented in Tables 5–8.

Table 5. Dependency distance and frequency in Kojiki

Dependency relation	**Dependency distance**	**Frequency Percentage**
case	1	0.36
compound	-1	0.26
mark	0	0.08
nmod	-2	0.06
nsubj	-2, -1	0.05
obl	-7, -9, -4	0.04
advmod	-1	0.04
obj	-2	0.03
acl	0	0.03
aux	1, 2	0.03
det	-1	0.01

Table 6. Dependency distance and frequency in Nihonshoki

Dependency relation	**Dependency distance**	**Frequency (percentage) of dependency relation**
case	1	25
mark	1, 2	7
advmod	-3, -1, -2	7
aux	1, 2	5
nsubj	-6, -13， -3	3
compound	-1, -2	6
cop	1	1
det (determiner)	-1	1
nmod (noun modifier)	-1, -2, -3	2

Table 7. Dependency distance and frequency in Man'yōshū

Dependency relation	**Dependency distance**	**Frequency (percentage) of dependency relation**
case	1	5
mark	1	6
obl	-7, -9	2
obj	-5	1
nsubj	-3, -5	2
acl	-3, -1	2
aux	1	1
compound	-1	6
det (determiner)	-1	1
nmod (noun modifier)	-3, -2	3

Table 8. Dependency distance and frequency in Shoku Nihongi Senmyō

Dependency relation	**Dependency distance**	**Frequency (percentage) of dependency relation**
case	1，2	21
mark	1	2
nsubj	-4, -1	2
obj	-3	1
obl	-2, -3	2
acl	-1, -3, -6	4
aux	1	3
nsubj	-29	1
compound	-1，-2，-3	29
nmod (noun modifier)	-2, -3	6
advmod	-5	1

Among all the dependency relations in this study, 46.3 per cent were an adjacent relationship, rendered by *case*, *mark*, and *det* (with DD = 1); aux, compound, and mark (with DD no longer than 2). This finding aligns with that of Buch-Kromann (2005). Conversely, *nsubj*, *advmod*, *obl*, and *acl* were long-distanced and presented diversity, with *nsubj* ranging from 1 to 29, *advmod* from 1 to 5, *obl* from 4 to 9, and *acl* from 1 to 6. The MDD of *nsubj* relation merits discussion. As first asserted by Li and Liu (2018: 255), *nsubj* and *obj* in English and Chinese obey a dependency distance minimization law. The *nsubj* relationship in Japanese, however, presents a different picture, confirming the assumption that Japanese is a topic-prominent language (Li and Thompson 1976), further analysed in (8).

(7) Zoo wa hana ga nagai
elephant TOP nose NOM long
'The elephants have long noses'.
(Mikami 1960)

In (7), *wa* is a topic marker, indicating the topic *zoo* 'elephant', and *ga* is a nominative case particle indicating the subject *hana* 'nose'. In a subject clause, the nominative case particle could be replaced by the topic marker to emphasise the subject (8a) or mark a contrast (8b).

(8) a. ringo ga suki da → b. ringo wa suki da
apple NOM like COP apple TOP like COP
'I like apple' 'apple, I like'

Moreover, the nominative case *ga* indicates new information, while the topic marker *wa* indicates old information. When the subject is marked by a topic marker, several grammatical items can be inserted between the subject and predicate (e.g. oblique cases, adverbs, and cleft clauses), which increases the syntactic complexity.

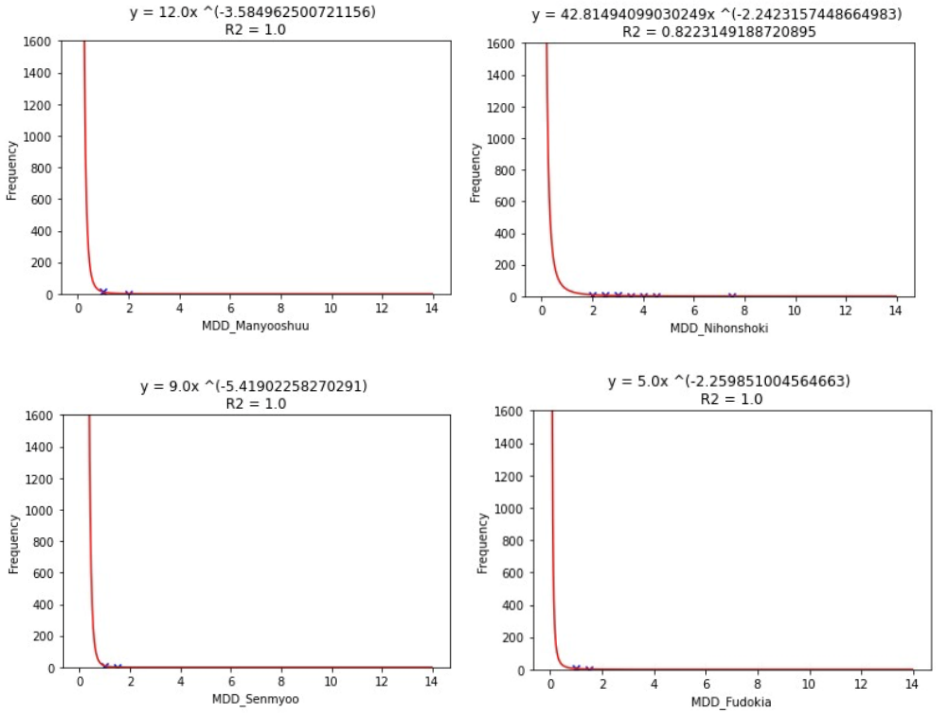
MDD-frequency regularity

We further aim to explore the regularity in the distributions and frequencies of MDD. A computer program is produced to fit the power law function ($y = ax^b$). The fit of the

data reveals that, apart from *Kojiki*, MDD and frequency demonstrate a close fit to a power law function, with 0.8967 as the lowest value of the determination coefficient R^2 and 1.0000 as the highest ($R^2 > 0.90$, very good; $R^2 > 0.80$, good; $R^2 > 0.75$, acceptable; $R^2 < 0.75$, unacceptable). These findings are summarised in Table 9 and Figure 2.

Table 9. Fitting the Power Law function to MDD of different genres in Old Japanese

Genres	**MDD**	***a***	***b***	**R^2**	**Fitting results**
Manyooshuu	1.19	12.0	-3.58	1.0	$y = 12x^{-3.58}$
Nihonshoki	3	42.8	-2.24	0.8223	$y = 42.8x^{-2.24}$
Senmyōgaki	1.2	9.0	-3.58	1.0	$y = 9x^{-3.58}$
Fudoki	1.26	5.0	-2.25	1.0	$y = 5x^{-2.25}$
Bussokusekika	1.62	4.0	-5.41	1.0	$y = 4x^{-5.41}$



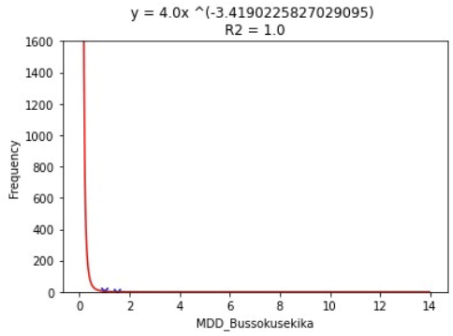
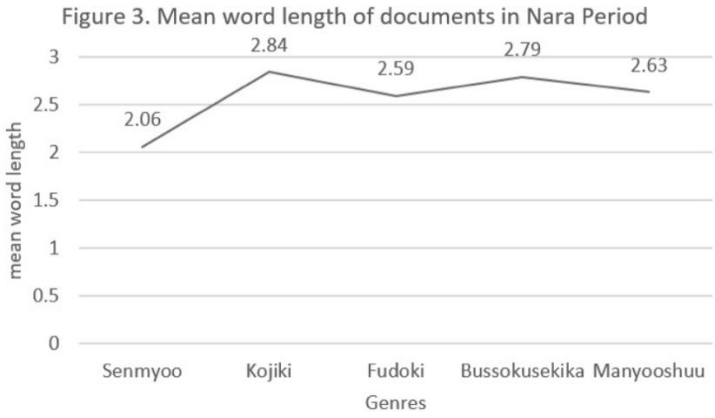


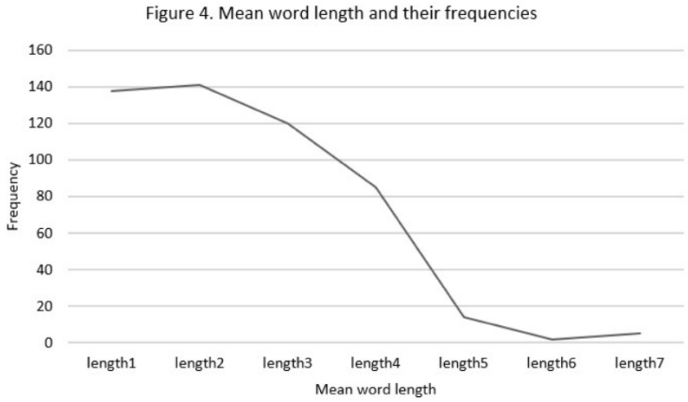
Figure 2. Fitting results of power law function

Lexical complexity

Having highlighted syntactic complexity in Old Japanese, this section seeks to examine lexical complexity by examining the word length of documents written in four writing styles: variant Chinese; classical Chinese; *man'yōgana*; *senmyōgaki*. Figure 3 shows the mean word length of each document.



The mean word length of Old Japanese documents ranges from 1 to 7. It seems that the senmyōgaki bears the relatively short mean word length while Kojiki presents the longest word length. The two poetries, Fudoki and Man'yōshū display similar length. The distribution of mean word length from 1 to 7, and their frequencies are presented in Figure 4, which supports Zipf's (1949) principle of least effort: shorter words tend to be more frequently used.



A further analysis through Altmann-fitter is carried out to explore the parameters that might indicate a trend in MWL-frequency distribution. The finding reveals that Bussokusekika and Fudoki can be demonstrated by Cohen-binomial model, and Senmyō is fitted to the Palm-Poisson model.

Table 10. Fiting Altmann-fitter to MWL-frequency distribution

Genres	**Fitting results**				
Bussokusekika	Cohen-binomial (n,p,α)	n	p	α	R^2
		6.0000	0.2748	0.2764	0.9926
Fudoki	Cohen-binomial (n,p,α)	6.0000	0.2748	0.2764	0.9926
Senmyō	Palm-Poisson (a; R = x-max)	a	R		R^2
		0.1532	7		0.9950

Figure 5 demonstrates the fitting result of Senmyō.

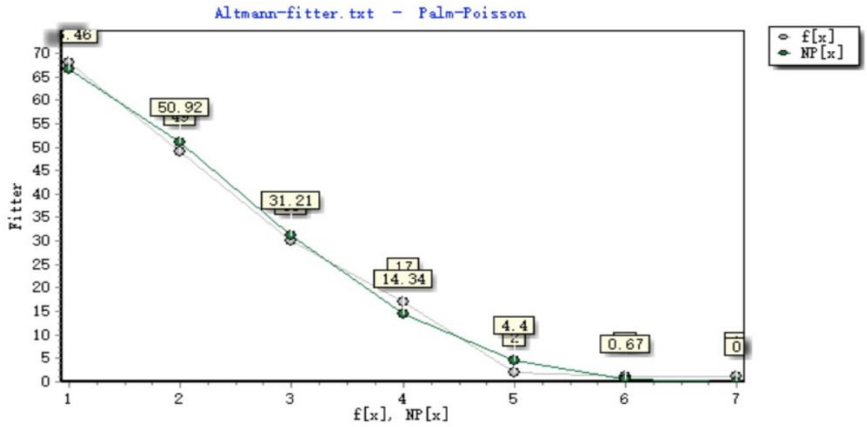


Figure 5. Fitting Altmann-Fitter to the MWL of *Senmyō*

CONCLUSION

The previous section presented an overall picture of syntactic complexity from Old Japanese to the present day. This study's findings have revealed that the syntactic structure in Old Japanese appears to be simpler and to have short DD. As confirmed by the null tokens of cc relation and advcl relation, Old Japanese is mostly conveyed by single clause. The distribution of MDD and their frequency in Nihonshoki, Fudoki, Bussokusekika and *Man'yōshū* can be well demonstrated by the power law function ($y=ax^b$). Altmann-Fitter was used to fit the dynamic MDD data of the four writing styles: variant Chinese; classical Chinese; man'yōgana; senmyoo gaki. In terms of lexical complexity, the mean word length ranges from 1 to 7. Among the four writing styles, the senmyōgaki bears the relatively short mean word length while Kojiki presents the longest word length. The two poetries, Fudoki and Man'yōshū display similar length. The relation between mean word length and frequencies supports Zipf's (1949) principle of least effort: shorter words tend to be more frequently used.

References

Hattori, Shiro. (1959). *Nihongo no Keitoo [Japanese genealogy]*. Iwanami Press.

Hudson, Richard. (1990). *English Word Grammar*. Oxford: Blackwell.

Hudson, Richard. (2007). *Language networks: The new word grammar*. Oxford: OxfordUniversity Press.

Li, Wenwen and Liu, Haitao. 2018. Hanying zhubinyu jufa jiliang tezhengde duibiyanjiu [A quantitative linguistic study on Chinese and English subject and object]. In Liu Haitao (ed.). *Research progress on quantitative linguistic studies*, Hangzhou: Zhejiang University Press, pp. 244-267.

Liu, Haitao. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2): 159-191.

Liu, Haitao, Hudson, Richard, Feng, Zhiwei. (2009). Using a Chinese treebank to measure dependency distance. *Corpus Linguistics and Linguistic Theory* 5 (2), 161-174.

Matsumoto, Katsumi. (2007). *Sekaigengo no naka no nihongo: nihongo keetooron no aratana chihee [Japanese among cross linguistics: a new view on Japanese genealogy]*. Sanseedo Press.

Ohno, Susumu. (1957). *Nihongo no kigen [The origin of Japanese language]*. Iwanami Press.

Ohno, Susumu. (1981). *Nihongo to Tamirugo [Japanese and Tamil]*. Shinchoo Press.

Robbeets, M., Bouckaert, R., Conte, M. *et al.* (2021). Triangulation supports agricultural spread of the Transeurasian languages. *Nature* 599**,** 616–621.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris: Klinksieck.

Yanagida Yuuko. (2014). Gengoruikeiron kara mita Joodai Nihongo no shugohyooji/ mokutekigohyooji: "*ga*" to "*o*" to *zero* hyooji ni tsuite [A typological investigation to the subject marker and object marker in Old Japanese]. *Nihongogaku* 33 (14): 124-137.

Yan Jianwei and Liu, Haitao. (2021). Morphology and word order in Slavic languages: Insights from annotated cor- pora. *Voprosy Jazykoznanija*, 4: 131–159.