Morphology and word order in Slavic languages: Insights from annotated corpora

Yan Jianwei[a]
Liu Haitao[a, b, @]

[a]Zhejiang University, Hangzhou, China;
[b]Guangdong University of Foreign Studies, Guangzhou, China; htliu@163.com

Abstract: Slavic languages are generally assumed to possess rich morphological features with free syntactic word order. Exploring this complexity trade-off can help us better understand the relationship between morphology and syntax within natural languages. However, few quantitative investigations have been carried out into this relationship within Slavic languages. Based on 34 annotated corpora from Universal Dependencies, this paper paid special attention to the correlations between morphology and syntax within Slavic languages by applying two metrics of morphological richness and two of word order freedom, respectively. Our findings are as follows. First, the quantitative metrics adopted can well capture the distributions of morphological richness and word order freedom of languages. Second, the metrics can corroborate the correlation between morphological richness and word order freedom. Within Slavic languages, this correlation is moderate and statistically significant. Precisely, the richer the morphology, the less strict the word order. Third, Slavic languages can be clustered into three subgroups based on classification models. Most importantly, ancient Slavic languages are characterized by richer morphology and more flexible word order than modern ones. Fourth, as two possible disturbing factors, corpus size does not greatly affect the results of the metrics, whereas corpus genre does play an important part in the measurements of word order freedom. Specifically, the word order of formal written genres tends to be more rigid than that of informal written and spoken ones. Overall, based on annotated corpora, the results verify the negative correlation between morphological richness and word order rigidity within Slavic languages, which might shed light on the dynamic relations between morphology and syntax of natural languages and provide quantitative instantiations of how languages encode lexical and syntactic information for the purpose of efficient communication.

Keywords: corpus linguistics, language complexity, linguistic typology, morphology, quantitative linguistics, Slavic, word order
For citation: Yan J., Liu H. Morphology and word order in Slavic languages: Insights from annotated corpora. *Voprosy Jazykoznanija*, 2021, 4: 131–159.
DOI: 10.31857/0373-658X.2021.4.131-159

Морфология и порядок слов в славянских языках: исследование на материале аннотированных корпусов

Янь Цзяньвэй[а]
Лю Хайтао[а, б, @]

[а]Чжэцзянский университет, Ханчжоу, Китай; [б]Гуандунский университет иностранных языков и внешней торговли, Гуанчжоу, Китай; htliu@163.com

Аннотация: Известно, что славянские языки обладают богатой морфологией, а также свободным порядком слов. Исследование взаимосвязи этих двух характеристик важно для понимания соотношения между морфологией и синтаксисом в естественных языках. Однако квантитативных

исследований этого вопроса на славянском материале существует немного. В данной статье на материале 34 аннотированных корпусов из Universal Dependencies исследуется корреляция между морфологией и синтаксисом в славянских языках с использованием двух метрик богатства морфологии и двух метрик свободы порядка слов. Результаты заключаются в следующем. Во-первых, принятые количественные метрики хорошо отражают связь между морфологическим богатством и свободой порядка слов в языках. Во-вторых, метрики подтверждают наличие корреляции между морфологическим богатством и свободой порядка слов (чем богаче морфология, тем менее строгий порядок слов). В славянских языках эта корреляция является умеренной и статистически значимой. В-третьих, славянские языки можно разделить на три подгруппы на основе классификационных моделей. В частности, древние славянские языки характеризуются более богатой морфологией и более гибким порядком слов, чем современные. В-четвертых, было установлено, что размер корпуса не сильно влияет на результаты анализа, но преобладающий в корпусе жанр имеет большое значение при измерении свободы порядка слов — а именно, порядок слов в формальных письменных текстах является более жестким, чем в неформальных письменных и в устных текстах. В целом анализ аннотированных корпусов подтверждает корреляцию между морфологическим богатством и свободой порядка слов в славянских языках, что может помочь нам в понимании динамических связей между морфологией и синтаксисом естественных языков и послужить квантитативной иллюстрацией того, как языки кодируют лексическую и синтаксическую информацию для эффективной коммуникации.

Ключевые слова: квантитативная лингвистика, корпусная лингвистика, лингвистическая типология, морфология, порядок слов, славянские языки, языковая сложность

Для цитирования: Yan J., Liu H. Morphology and word order in Slavic languages: Insights from annotated corpora. *Voprosy Jazykoznanija*, 2021, 4: 131–159.

DOI: 10.31857/0373-658X.2021.4.131-159

1. Introduction

Plenty of linguists hold the “negative correlation hypothesis” or the “complexity trade-off hypothesis”, which states that different components of human language (e.g., phonology, morphology, syntax, and semantics) are negatively correlated in terms of complexity (e.g., [Shosted 2006; Fenk-Oczlon, Fenk 2014; Coloma 2017]). In other words, if one component of a language is very sophisticated, another component of that language tends to be simplified. For example, Fenk-Oczlon and Fenk [2014] reported significant negative cross-linguistic correlations between syllable complexity and number of syllables per clause and per word, and an almost significant negative correlation between syllable complexity and number of morphological cases. Coloma [2017] examined the interrelationship among three components, phoneme per syllable, syllable per word, and word per clause, and found that they are negatively correlated between themselves, supporting the possible existence of complexity trade-offs. Shosted [2006] investigated the correlation between phonology and morphology, and found that the correlation is slightly positive but statistically insignificant. These studies on different levels provided a window into how human beings encode linguistic information, thus enriching our understanding of the dynamic relations of different components of human language.

When it comes to the correlation between the components of morphology and syntax, one commonly made cross-linguistic generalization is that languages with rich case-marking tend to have more freedom of word order than the languages without [Sapir 1921; Jakobson 1936; McFadden 2003]. Some studies have also proposed metrics or ways to quantify or explain this correlation. For instance, Sinnemäki [2014] tested whether complexity in case marking correlates with simplicity in word order cross-linguistically, and the results showed that languages with a lot of variety in case marking tend to have less variety in word order patterns and this correlation is strong in terms of the inventory of linguistic units and constructions. In addition, Koplenig et al. [2017] investigated the statistical trade-off between word order and word structure

of almost 1,200 different languages based on the metric of entropy, and found that when less information is carried by word structure, more information has to be conveyed by word order. Also, based on word order entropy, one of the findings of Levshina [2019: 562] is that some morphologically rich European VO and OV languages "tend to have high entropy of head-dependent orders". Moreover, McFadden [2003] explained the correlation between word order and morphology from the perspectives of language use, language acquisition, and language change. However, special attention still needs to be granted to the distinctive features of the Slavic group with more comprehensive and interpretable metrics.

Slavic languages are characterized as being morphologically conservative with rich fusional morphology and syntactically salient with so-called free word order [Comrie, Corbett (eds.) 1993: 6–7; Klein et al. (eds.) 2018]. As synthetic languages, they allow one word ending to express several categories at once, which makes them typically fusional or inflectional. In contrast, analytic languages, such as Chinese and Vietnamese, contain very little inflection and rely instead on features like word order to convey grammatical information. Meanwhile, English is often considered as one of the most analytic Indo-European languages though it is traditionally analyzed as a fusional one. Although a number of theoretical studies, computational research, and translational practices suggest that the free word order of Slavic languages may be correlated with their rich morphological features (e.g., [Gulordava, Merlo 2015; Maučec, Brest 2019]), the empirical investigations of this correlation and its ongoing changes within this language group are still rare.

Also, the emergence of large-scale annotated corpora (or treebanks) [Hajič 1998; Abeillé (ed.) 2003] and quantitative indicators paves the way for this study. On the one hand, annotated corpora are becoming a new linguistic resource for typological studies. There are already many quantitative studies based on treebanks, attempting to unveil the typological features of languages, e.g. [Liu 2010; Futrell et al. 2015; Alzetta et al. 2019]. Results based on annotated corpora are more profound and thorough, since they can better reflect the unique features of intra-linguistic variation [Levshina 2019], and conclusions drawn are based on "language sample used in practice instead of just on some simple sentences collected for the study" [Liu 2010: 1568]. On the other hand, the field of typology has a long history of adopting quantitative indicators to characterize linguistic features (e.g., [Greenberg 1960; 1963]). Also, quantitative metrics, combined with statistical measures, can better uphold the validity of the results. As put by Plungian [2018: 11], without statistics, the language system can generally not be fully understood, and statistics plays an essential role in linguistics of the 21st century. Hence, by adopting different metrics at each linguistic level, we aimed to quantify the correlation between morphological richness and word order freedom within the Slavic language group based on large-scale annotated corpora.

To this end, we employed 34 annotated corpora with morphological and syntactic annotations, which represent 13 Slavic and four non-Slavic languages (Vietnamese, Classical Chinese, Standard Chinese, and English). The reason why we adopted non-Slavic languages is that they range from typical analytic (Vietnamese, Classical Chinese), highly analytic (Standard Chinese) to a moderately analytic, but still the most analytic among Indo-European languages (English). We expect that the measured results of these languages also scatter consecutively in the typological continuum, and the comparison across Slavic and non-Slavic can better present the features of Slavic languages under discussion. Meanwhile, two metrics of morphological richness and two of word order freedom were adopted, and the correlations of the four metrics were also presented. The reason for choosing more than one indicator is to ensure the reliability of our measurements. Based on the prior introspective intuition and general quantitative findings mentioned above, we expect Slavic languages to conform to the typological generalization of "negative correlation". Moreover, we conducted a cluster analysis of Slavic languages to verify the robustness of the measures. In doing so, we hope that, based on four indicators, the analysis of the correlation between the morphological and the syntactic levels can better describe the linguistic characteristics of the Slavic languages, thus bringing new insights into the research of related languages and of how lexical and syntactic information is encoded.

Specifically, the research questions in this paper are as follows.

Question 1: Are the values of different indicators based on annotated corpora reliable to reflect the morphological richness and word order freedom of languages?

Question 2: Are the morphological richness and word order freedom of languages positively correlated? Can we see any variation of morphological richness and word order freedom within Slavic languages? If there is variation, does the well-known correlation hold?

Question 3: Where do languages (or subgroups) end up in the two-dimensional space? Are there any differences between modern and ancient Slavic languages?

Question 4: Are there any possible factors related to specific corpora that might have effects on the results of the metrics of morphological richness and word order freedom?

2. Materials and methods

2.1. Materials

In the current study, we adopted all Slavic treebanks in UD 2.5 [Zeman et al. 2019],[1] 24 treebanks in total, involving 13 Slavic languages, i.e., Belarusian (one treebank), Bulgarian (one treebank), Croatian (one treebank), Czech (five treebanks), Old Church Slavonic (one treebank), Old Russian (two treebanks), Polish (three treebanks), Russian (four treebanks), Serbian (one treebank), Slovak (one treebank), Slovenian (two treebanks), Ukrainian (one treebank) and Upper Sorbian (one treebank). Besides, we employed ten treebanks[2] of four non-Slavic languages for comparison, i.e., Standard Chinese (three treebanks), Classical Chinese (one treebank), English (five treebanks), and Vietnamese (one treebank). Altogether, 34 treebanks of 17 languages are used in this study. The details on the treebanks are shown in **Appendix A**.

All treebanks are annotated in CoNLL-U format according to dependency grammar [Tesnière 1959; 2015; Mel'čuk 1988] to describe linguistic relations of elements, i.e., the head and the dependent within sentences [Heringer 1993; Hudson 1995; Jiang, Liu (eds.) 2018]. To be specific, ten fields are annotated, i.e., ID (word index of the dependent), FORM, LEMMA, UPOS, XPOS (language-specific POS tag), FEATS (morphological features), HEAD (word index of the head), DEPREL (dependency relation), DEPS (enhanced dependency descriptions) and MISC (other miscellaneous annotation).[3] **Table 1** (p. 135) is a simplified CoNLL-U version of an example sentence.

2.2. Methods

Four metrics are used in this study, namely, moving-average morphological richness (MAMR), moving-average mean size of paradigm (MAMSP), cosine similarity (COSS), and word order entropy (ENTR). The first two are used to measure morphological richness, and the other two are used to measure word order freedom. On the one hand, these four measures can make full use of the annotated corpora; on the other hand, they are computable and interpretable.

[1] For more information, see https://universaldependencies.org/.

[2] We adopted all treebanks of Standard Chinese, Classical Chinese, English and Vietnamese in UD 2.5, except four of them, i.e., Chinese-HK (no annotation of LEMMA), Chinese-PUD (no annotation of LEMMA), English-ESL (no annotation of FORM and LEMMA), English-Pronouns (no declarative sentences with complete S/V/O combinations).

[3] For more information, see https://universaldependencies.org/format.html.

Table 1

Simplified CoNLL-U version of the sentence *You like apples, he likes apples, she likes apples, and I like apples*

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	You	you	PRON	PRP	-	2	nsubj	NA	NA
2	like	like	VERB	VBP	-	0	root	NA	NA
3	apples	apple	NOUN	NNS	-	2	obj	NA	SpaceAfter=No
4	,	,	PUNCT	,	-	2	punct	NA	NA
5	he	he	PRON	PRP	-	6	nsubj	NA	NA
6	likes	like	VERB	VBZ	-	2	parataxis	NA	NA
7	apples	apple	NOUN	NNS	-	6	obj	NA	SpaceAfter=No
8	,	,	PUNCT	,	-	2	punct	NA	NA
9	she	she	PRON	PRP	-	10	nsubj	NA	NA
10	likes	like	VERB	VBZ	-	2	parataxis	NA	NA
11	apples	apple	NOUN	NNS	-	10	obj	NA	SpaceAfter=No
12	,	,	PUNCT	,	-	15	punct	NA	NA
13	and	and	CCONJ	CC	-	15	cc	NA	NA
14	I	I	PRON	PRP	-	15	nsubj	NA	NA
15	like	like	VERB	VBP	-	2	conj	NA	NA
16	apples	apple	NOUN	NNS	-	15	obj	NA	SpaceAfter=No
17	.	.	PUNCT	.	-	2	punct	NA	NA

2.2.1. Morphological richness

Moving-average morphological richness (MAMR) [Čech, Kubát 2018] is a measure that builds on the wordform vocabulary richness and the lemma vocabulary richness. Therefore, to explain how MAMR is calculated, we should first understand the calculation of vocabulary richness. Type-token ratio (TTR) is the most commonly used measure of vocabulary richness, and it is defined as vocabulary size divided by the text length. Besides, to minimize the influence of corpus size, Covington and McFall [2010] proposed to calculate TTR using a moving window to obtain moving average type-token ratio (MATTR). It is defined as:

$$MATTR(W)_{wordform} = \frac{\sum_{i=1}^{N-W+1} F_i}{W(N-W+1)}.$$

In this formula, N represents the number of tokens in a corpus, which can be divided into overlapped sub-corpora of the same token size or so-called "windows" with randomly chosen token size W ($W < N$). Typically, the windows are moved forward one step at a time. F_i is the number of word types (distinct word forms) in each window. Hence, the $MATTR(W)_{wordform}$ is the mean of a series of TTRs based on word forms for all windows.

Similarly, MATTR based on word lemmas can be defined as:

$$MATTR(W)_{lemma} = \frac{\sum_{i=1}^{N-W+1} L_i}{W(N-W+1)}.$$

Again, N represents the size of a corpus, W the window size W ($W < N$), and L_i the number of distinct lemmas in each window. Hence, the $MATTR(W)_{lemma}$ is the mean of a series of TTRs based on word lemma for all windows.

For example, the length (the number of tokens) of the example sentence in **Table 1** *You like apples, he likes apples, she likes apples, and I like apples* is 13 ($N = 13$) (We exclude punctuations during calculations). If a window size of 4 tokens ($W = 4$) is adopted, we obtain 10 windows — *you like apples he* | *like apples he likes* | *apples he likes apples* | *he likes apples she* | *likes apples she likes* | *apples she likes apples* | *she likes apples and* | *likes apples and I* | *apples and I like* | *and I like apples*. Then, according to the FORM field of **Table 1**, the numbers of distinct word forms in these 10 windows are 4 (*you, like, apples, he*), 4 (*like, apples, he, likes*), 3 (*apples, he, likes*), 4 (*he, likes, apples, she*), 3 (*likes, apples, she*), 3 (*apples, she, likes*), 4 (*she, likes, apples, and*), 4 (*likes, apples, and, I*), 4 (*apples, and, I, like*), and 4 (*and, I, like, apples*). Then, the $MATTR(4)_{wordform}$ of the sequence can be calculated as:

$$MATTR(W)_{wordform} = \frac{\sum_{i=1}^{N-W+1} F_i}{W(N-W+1)} = \frac{4+4+3+4+3+3+4+4+4+4}{4(13-4+1)} = 0.925.$$

Also, as shown in the LEMMA field of **Table 1,** we can obtain 10 windows of word lemmas, i.e., *you like apple he* | *like apple he like* | *apple he like apple* | *he like apple she* | *like apple she like* | *apple she like apple* | *she like apple and* | *like apple and I* | *apple and I like* | *and I like apple*. Then the numbers of distinct word lemmas in these 10 windows are 4 (*you, like, apple, he*), 3 (*like apple he*), 3 (*apple he like*), 4 (*he, like, apple, she*), 3 (*like apple she*), 3 (*apple she like*), 4 (*she, like, apple, and*), 4 (*like, apple, and, I*), 4 (*apple, and, I, like*) and 4 (*and, I, like, apple*). Hence, we can compute the $MATTR(4)_{lemma}$ of the sequence as follows:

$$MATTR(W)_{lemma} = \frac{\sum_{i=1}^{N-W+1} L_i}{W(N-W+1)} = \frac{4+3+3+4+3+3+4+4+4+4}{4(13-4+1)} = 0.9.$$

Accordingly, by definition, the MAMR of a sentence or a corpus can be calculated from the difference between the MATTR computed in word forms and the MATTR computed in lemmas [Čech, Kubát 2018]:

$$MAMR(W) = MATTR(W)_{wordform} - MATTR(W)_{lemma},$$

namely,

$$MAMR(W) = \frac{\sum_{i=1}^{N-W+1} F_i}{W(N-W+1)} - \frac{\sum_{i=1}^{N-W+1} L_i}{W(N-W+1)}.$$

Thus, the MAMR of the example sentence is $0.925 - 0.9 = 0.025$. The greater the difference, the higher the morphological richness of a sentence, text or corpus [Čech, Kubát 2018]. Previously, Čech and Kubát [2018] empirically showed the effectiveness of MAMR in genre classification. The present study aims to explore its applicability in the context of typological research.

In this study, we used a standard window size of 500 words ($W = 500$) following Covington and McFall [2010], and calculate the values of $MATTR(500)_{wordform}$ and $MATTR(500)_{lemma}$ using the MATTR software Version 2.0.[4] Specifically, the $MATTR(500)_{wordform}$ and $MATTR(500)_{lemma}$ for each language are calculated based on the FORM and LEMMA extracted from the UD treebanks, as shown in **Table 1**.

[4] For more information, see http://ai1.ai.uga.edu/caspr/.

The other metric used in this paper to calculate morphological richness was proposed by Xanthos and Gillis [2010]. It defines the morphological richness in terms of an average number of distinct inflected word forms per lemma, a simple version of the mean size of paradigm (MSP).[5] The algorithm is:

$$MSP = \frac{F}{L}.$$

In this formula, *F* and *L* represent the number of distinct inflected word forms and the number of distinct lemmas, respectively, in a sentence or a corpus. Thus, considering the example sentence *You like apples, he likes apples, she likes apples, and I like apples*, the number of distinct inflected word forms is 8 (*F* = 8) (*you, he, she, I, and, like, likes, apples*) and the number of the root lexicons or distinct lemmas is 7 (*L* = 7) (*you, he, she, I, and, like, apple*). Thus, *MSP* = 8 ÷ 7 ≈ 1.14.

In addition, the normalized MSP (NMSP) algorithm and robust MSP (RMSP) algorithms were also proposed [Xanthos et al. 2011; Xanthos, Gillis 2010; Xanthos, Guex 2015], aimed to reduce the influence of sample size and compensate for MSP's dependence on lexematic diversity.

In this paper, to be consistent with the first morphological measure (MAMR), we modified the MSP algorithm into the moving-average MSP (MAMSP). Also, instead of merely considering verbs and nouns in a corpus as Xanthos and Gillis [2010] did, we took all POS as observations when calculating MAMSP since inflectional languages (especially Slavic languages) also have extensive case systems for pronouns, adjectives, and determiners, etc. The formula of MAMSP is as follows:

$$MAMSP(W) = \frac{\sum_{i=1}^{N-w+1} \frac{F_i}{L_i}}{N - W + 1}.$$

N represents the number of tokens in a corpus, *W* the window size (*W* < *N*), F_i the number of distinct word forms in each window, and L_i the number of distinct word lemmas in each window. Take the example sentence for instance. When we adopt a moving window of 4 tokens (*W* = 4), 10 windows can be obtained. Based on **Table 1**, the numbers of distinct word forms in these 10 windows are 4, 4, 3, 4, 3, 3, 4, 4, 4, 4, and those of distinct lemmas are 4, 3, 3, 4, 3, 3, 4, 4, 4, 4. Then, the MSPs of all 10 windows are 4/4, 4/3, 3/3, 4/4, 3/3, 3/3, 4/4, 4/4, 4/4, and 4/4, respectively. Thus, the MAMSP of the example sentence is:

$$MAMSP(4) = \frac{\sum_{i=1}^{N-w+1} \frac{F_i}{L_i}}{N - W + 1} = \frac{\frac{4}{4}+\frac{4}{3}+\frac{3}{3}+\frac{4}{4}+\frac{3}{3}+\frac{3}{3}+\frac{4}{4}+\frac{4}{4}+\frac{4}{4}+\frac{4}{4}}{13-4+1} \approx 1.03.$$

The higher the value of MAMSP, the more complex the morphology of the language under investigation. Previously, empirical studies [Xanthos, Gillis 2010; Xanthos et al. 2011; Xanthos, Guex 2015] reported the effectiveness of MSP in reflecting morphological richness development in language acquisition. This study aims to investigate the applicability of modified MSP (MAMSP) in typological research.

Based on the FORM and LEMMA fields extracted from the UD treebanks, we likewise adopted a standard window size of 500 (*W* = 500) to compute MAMSP by a self-written R script.

[5] Size of paradigm refers to how many distinct word forms a lexeme has. For example, Sanskrit *a*-stem noun *dev-* 'god' has eight singular forms, viz., *dev-as*, *dev-a*, *dev-am*, *dev-ena*, *dev-āya*, *dev-āt*, *dev-asya*, and *dev-e* [Whitney 1889].

2.2.2. Word order freedom

Kuboň et al. [2016] compared four methods for calculating word order freedom based on the probabilities of the six variants of S/V/O word orders in corpora (SVO, OVS, VSO, VOS, SOV and OSV). It was found that "the phenomenon of word order freedom can be quantified practically by any reasonably selected measure" [Ibid.: 17]. Out of the four metrics in Kuboň et al. [2016], we adopted the third and the fourth ones, namely, the cosine similarity and the word order entropy. This is because that the first one, max–min distance (the maximal probability minus the minimal probability among the six variants), only considered the difference between the maximum and the minimum variants, ignoring the other four variants. The second metric, the standard Euclidean distance, is very similar to the third one.

The cosine similarity (COSS), which measures the similarity of two vectors, is widely used in information retrieval [Muflikhah, Baharudin 2009; Li, Han 2013]. In word order freedom calculation, it measures the distances between the actual probability of each S/V/O variant and the "ideal vector" with the equal frequency distribution of all six variants (i.e., 100 % / 6 ≈ 0.1667) [Kuboň et al. 2016]. The algorithm is as follows:

$$COSS = \frac{\sum_{i=1}^{n} P(x_i) \times P(y_i)}{\sqrt{\sum_{i=1}^{n} P(x_i)^2} \times \sqrt{\sum_{i=1}^{n} P(y_i)^2}}.$$

In this formula, the symbol $P(x_i)$ represents the probability or the relative frequency of each word-order variant in a given language and $P(y_i)$ is the "ideal vector" (0.1667) with equal distribution of frequencies. The values of COSS are increasing with the growth of word order freedom [Kuboň et al. 2016]. For example, if the word-order distribution for language *A* is SVO (30 %), OVS (20 %), VSO (10 %), VOS (15 %), SOV (15 %), and OSV (10 %), then its cosine similarity is:

$$COSS(A) = \frac{0.3 \times 0.1667 + 0.2 \times 0.1667 + 0.1 \times 0.1667 + 0.15 \times 0.1667 + 0.15 \times 0.1667 + 0.1 \times 0.1667}{\sqrt{0.3^2 + 0.2^2 + 0.1^2 + 0.15^2 + 0.15^2 + 0.1^2} \times \sqrt{0.1667^2 + 0.1667^2 + 0.1667^2 + 0.1667^2 + 0.1667^2 + 0.1667^2}} \approx 0.9245.$$

Similarly, the probabilities of the six S/V/O word-order variants in the example sentence from **Table 1** (we denote it as *B*) are SVO (100 %), OVS (0 %), VSO (0 %), VOS (0 %), SOV (0 %), and OSV (0 %), then:

$$COSS(B) = \frac{1 \times 0.1667}{\sqrt{1^2} \times \sqrt{0.1667^2 + 0.1667^2 + 0.1667^2 + 0.1667^2 + 0.1667^2 + 0.1667^2}} \approx 0.4083.$$

Since the value of *COSS*(*A*) is larger than that of *COSS*(*B*), the word order of language *A* is freer than that of language *B* (our example sentence).

Finally, entropy (ENTR) is adopted as the second metric to measure the degree of word order freedom. Defined as a measure for the choice associated with symbols in strings [Shannon 1948], entropy is widely used in linguistic studies (e.g., [Chen et al. 2016; Bentz et al. 2017; Gutierrez-Vasques, Mijangos 2018]). The formula is as follows:

$$ENTR = -\sum_{i=1}^{n} P(x_i) \times \ln P(x_i).$$

Here, *ln* stands for natural logarithm. The values $P(x_i)$ in this formula are the probabilities of the six word-order variants. The entropy is maximal for the equal distribution of probabilities and

minimal for a language system that has only one acceptable word-order variant [Kuboň et al. 2016]. Thus, the higher the entropy for a particular language, the higher its degree of word order freedom.

Then the word order entropy of language *A* (SVO (30 %), OVS (20 %), VSO (10 %), VOS (15 %), SOV (15 %) and OSV (10 %)) is:

$$ENTR\ (A) = -\big(0.3 \times \ln(0.3) + 0.2 \times \ln(0.2) + 0.1 \times \ln(0.1) + 0.15 \times \ln(0.15) + 0.15 \times \ln(0.15) + 0.1 \times \ln(0.1)\big) \approx 1.7127.$$

Similarly, the word order entropy of the example sentence in **Table 1** (SVO (100 %), OVS (0 %), VSO (0 %), VOS (0 %), SOV (0 %) and OSV (0 %)) is:

$$ENTR(B) = -\big(1 \times \ln(1)\big) = 0.$$

The value of *ENTR*(*A*) is larger than that of *ENTR*(*B*); thus, the word order of language *A* is freer than that of language *B*.

As stated in [Kuboň et al. 2016: 13], "a typical mutual position of a subject, a predicate and an object constitute one of the basic typological characteristic[s] of a natural language" and "a combination of too many language phenomena in complicated sentences" might bias the final results; hence, we only focused on the order of the core arguments and verb (S, V, O) for the calculation of COSS and ENTR. Moreover, to avoid possible influence caused by different proportions of non-declarative sentences (i.e., imperative, interrogative and exclamative sentences) in 34 treebanks, we only focused on the declarative sentences in the current study. Therefore, technically, we first extracted all declarative sentences from the 34 treebanks by a self-written Perl script. Then, following the methods of Bonfante et al. [2018] and Courtin [2018: 36–40], we extracted the relative frequencies (or probabilities) of the six S/V/O word-order variants in the 34 treebanks based on the fourth column (UPOS) and the eighth column (DEPREL) in **Table 1** by specific extraction patterns of Grew. The examples of extraction patterns for SVO and VSO are given below.[6]

SVO Pattern:

pattern { V [upos=VERB]; V -[nsubj|csubj]-> S; V -[obj|iobj|xcomp|ccomp]-> O; S << V; V << O; S << O }

VSO Pattern:

pattern { V [upos=VERB]; V -[nsubj|csubj]-> S; V -[obj|iobj|xcomp|ccomp]-> O; V << S; S << O; V << O }

The other four extraction patterns are similar, and the resulting probabilities of all six variants of order of the subject, object and predicate verb, and numbers of declarative sentences are shown in **Appendix B**. Finally, based on the results, we can calculate the values of COSS and ENTR by a self-written R script.

To summarize, the correlations between morphological richness and word order freedom in Slavic languages are quantified by computable and comprehensive measures based on annotated corpora. At the operational level, MAMR is computed using MATTR software, and MAMSP, COSS, and ENTR are computed using self-written R scripts.

3. Results and discussion

Section 3.1 examines the reliability of the metrics adopted in reflecting morphological complexity and word order freedom, respectively. Then, **Section 3.2** first investigates the correlations between morphological richness and word order freedom to test whether the measures adopted

[6] For more information on the extraction patterns of Grew, see https://grew.fr/.

can be used to validate the trade-off hypothesis; then, it delves into the correlations within Slavic languages. **Section 3.3** tests whether these metrics can be applied to demonstrate the closeness of Slavic languages from a typological perspective. Finally, by subsetting treebanks into different sub-corpora and focusing on specific genres, the effects of corpus size and corpus genre on the results of the metrics are discussed in **Section 3.4**.

3.1. Reliability of the metrics of morphological richness and word order freedom

It is inherently difficult to quantify the degree of morphological richness and word order freedom of languages. As early as the 1960s, Greenberg proposed indicators to measure the degree of synthesis based on the number of morphemes per word, and the implicational Universals associated with word order properties [Greenberg 1960; 1963]. However, methods for measuring morphological richness and word order freedom remain controversial due to the variations of languages and their different ways of externalizing morphological and word order features. In attempting to test the reliability of the metrics adopted, we first calculated the values of MAMR and MAMSP, two measures of morphological complexity, of the 34 treebanks of 17 languages. The results are shown in **Appendix C**.

As introduced in **Section 2**, the higher values of MAMR and MAMSP correspond to the richer morphological complexity. **Appendix C** shows that Vietnamese, Chinese, and English have smaller MAMRs than Slavic languages. In terms of the MAMSP values, except for the English treebanks that are scattered among the treebanks of Slavic languages, all Vietnamese and Chinese treebanks own smaller values of MAMSP than Slavic ones do.

Moreover, the typical analytical languages (Vietnamese and Classical Chinese) rank at the far end of the morphological continuum with the lowest MAMR and MAMSP values in **Appendix C**, then come the highly analytical Standard Chinese language, the most analytic among Indo-European languages (English), and finally the typical fusional ones (Slavic languages). It confirms our assumption in **Section 1** (cf. **Introduction**) that these languages are consecutively distributed in the morphological continuum. Besides, to further examine the consistency and reliability of MAMR and MAMSP in reflecting morphological richness, we plotted a scatterplot with a regression line, as shown in **Figure 1**.

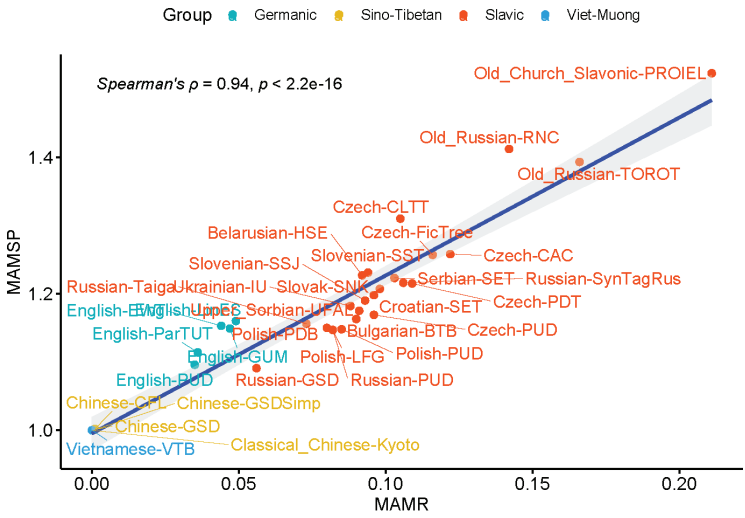

Figure 1. Scatterplot of MAMR and MAMSP with a regression line

Figure 1 shows that the relationship between MAMR and MAMSP fits the regression line well. Moreover, the Spearman's rank correlation coefficient between MAMR and MAMSP is positive, strong, and statistically significant: $\rho = 0.95$, $p < 0.001$.[7] It suggests that the two measures are homogeneous and consistent in reflecting the morphological richness of the language under discussion. Meanwhile, the Slavic group does exhibit more complex morphological features or a higher degree of morphological complexity [Comrie, Corbett (eds.) 1993] with relatively high values of MAMR and MAMSP. Compared with prior attempts to quantify the morphological complexity of Slavic languages, such as investigations based on plain texts [Popescu, Altmann 2008; Kelih 2010] or network parameters [Liu, Xu 2012; Liu, Cong 2013], the empirical results based on the FORM and LEMMA retrieved from annotated corpora can also provide new insights into approaching the concept of morphological richness properly.

We then calculated the values of COSS and ENTR, two measures of word order freedom. The results are shown in **Appendix D**.

As mentioned in **Section 2**, the higher values of COSS and ENTR correspond to the higher degree of word order freedom of the language in question. **Appendix D** shows that the COSS and ENTR values of Vietnamese, Chinese, and English are always smaller than those of Slavic languages. It suggests that Vietnamese, Chinese and English are less flexible than Slavic languages in terms of word order. Also, we drew a scatterplot with a regression line to visualize the relationship between COSS and ENTR, as shown in **Figure 2**.

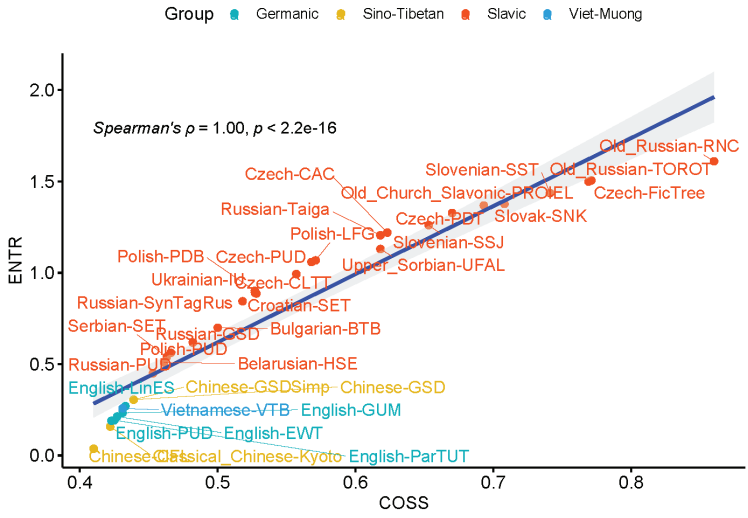

Figure 2. Scatterplot of COSS and ENTR with a regression line

Figure 2 shows that the regression line can well fit the relationship between COSS and ENTR, and the Spearman's rank correlation coefficient between COSS and ENTR is also positive, strong, and statistically significant, $\rho = 1.00$, $p < 0.001$. It indicates that the two metrics are

[7] Spearman's rank correlation, a nonparametric measure of the strength and direction of correlation that exists between two variables, was adopted here since the data violate the assumptions of Pearson product-moment correlation. We owe this point to Prof. Laura A. Janda. Conventionally, the closer the ρ (Spearman's *rho*) value to 1, the stronger the correlation. If the ρ value is equal to or greater than 0.7 or smaller than −0.7, the correlation is considered to be strong. If the ρ value is between 0.3 and 0.7 or between −0.3 and −0.7, it is moderate. If the ρ value is between 0 and 0.3 or 0 and −0.3, the correlation is weak.

homogeneous and reliable in capturing the degree of word order freedom of languages concerned. Meanwhile, Slavic languages are generally more flexible than Vietnamese, Chinese and English in terms of syntactic features [Comrie, Corbett (eds.) 1993]. Previously, based on entropy conditioned by unordered dependency graphs, Futrell et al. [2015] proposed a reasonable metric and applied it successfully to a collection of UD treebanks of 34 languages to demonstrate the negative correlation between word order freedom and the morphological richness; and later their findings were corroborated and further nuanced by Koplenig et al. [2017] and Levshina [2019]. In addition to one metric on entropy, the current study, by contrast, endeavored to capture the notion of word order freedom with two metrics, viz., word order entropy and cosine similarity. Both metrics are consistent with each other as discussed above, and more importantly, based on the validity of the measures, we would also pay special attention to one specific language group in the following discussions.

To summarize, based on information retrieved from large-scale annotated corpora, the morphological and syntactic metrics adopted in this study are computable, understandable, and can well capture the internal structures of human languages.

3.2. Correlation between morphological richness and word order freedom

In this section, we examine the correlations between morphological complexity and word order freedom across Slavic and non-Slavic languages first, then within Slavic ones.

We plotted the scatterplot matrix of the two morphological richness metrics and two word order freedom metrics, as shown in **Figure 3**.

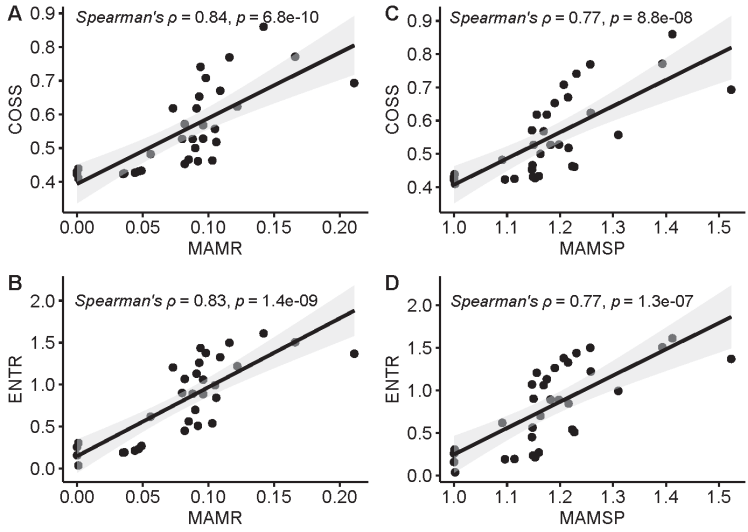

Figure 3. Scatterplot correlation matrix of MAMR vs. COSS, MAMR vs. ENTR, MAMSP vs. COSS, and MAMSP vs. ENTR with regression lines[8]

[8] As suggested by one of the three reviewers, an alternative way to present the multiple correlations of all four metrics is a correlogram with regression lines. That means we can combine **Figure 1**, **2** and **3** into one figure. However, we aimed to corroborate the reliability and consistency of the measures within domains first (i.e., morphology and syntax) (as shown in **Section 3.1**) and then the correlation across

Figure 3 shows that the correlations between morphological richness and word order freedom are all positive, strong, and statistically significant. Take plot A as an example. The Spearman's rank correlation coefficient between MAMR (one of the two measures of morphological complexity) and COSS (one of the two measures of word order freedom) is positive, $\rho = 0.84$, $p < 0.001$. It is also true for the other three panels. Specifically, the ρ values for the other three correlations are 0.83, 0.77, and 0.77, respectively, and their p-values are all statistically significant ($p < 0.001$). It suggests that as the morphological complexity of languages increases, their word order becomes freer or less rigid, which is consistent with the "complexity trade-off hypothesis". Hence, the statistical results provide empirical evidence for the qualitative hypothesis about morphology and syntax of languages [Sapir 1921; Jakobson 1936; McFadden 2003].

Moreover, taking **Figure 3 (A)** as an example, we plotted a scatterplot of MAMR and COSS with treebanks labels, as shown in **Figure 4,** to better observe the distribution of specific treebanks.

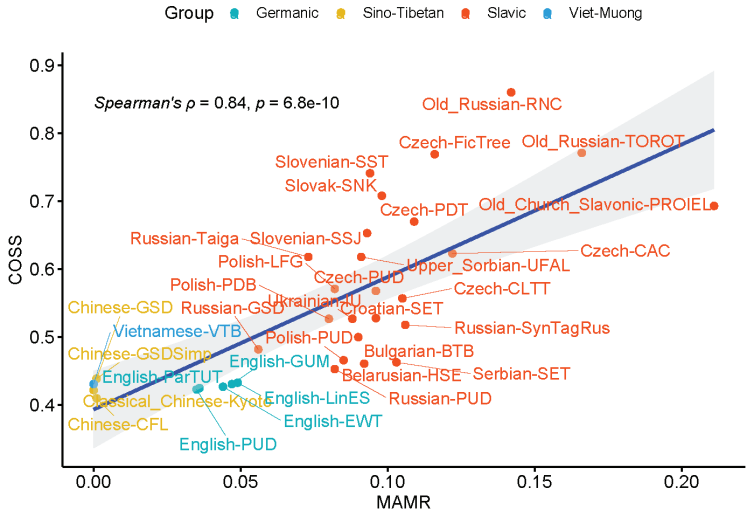

Figure 4. Scatterplot of MAMR and COSS with a regression line

Figure 4 further shows how typical analytic, highly analytic, the most analytic among Indo-European languages, and typical fusional Slavic treebanks scatter in the two-dimensional space, confirming that Vietnamese, Chinese and English treebanks do have a lower degree of morphological complexity and lower flexibility of word order than Slavic ones.

Combined with the findings in **Section 3.1**, the results here show that the metrics adopted are plausible to quantify the correlations between word order and morphology. It demonstrates that these metrics can be used to prove the "negative correlation hypothesis", which paves the way for the following examination of the correlations between morphological complexity and word order freedom within Slavic languages.

We plotted the scatterplot matrix of the four metrics within Slavic languages, as shown in **Figure 5**.

---

domains (i.e., whether word order correlates with morphological richness) (as shown in **Section 3.2**); Moreover, if we adopted a correlogram, the data points in the correlogram would not be labelled with treebank names, which, as suggested by our reviewer, may lead to difficulties in making detailed discussions. Hence, we presented the correlations within domains in **Figure 1** and **Figure 2**, and then across domains in **Figure 3** separately.

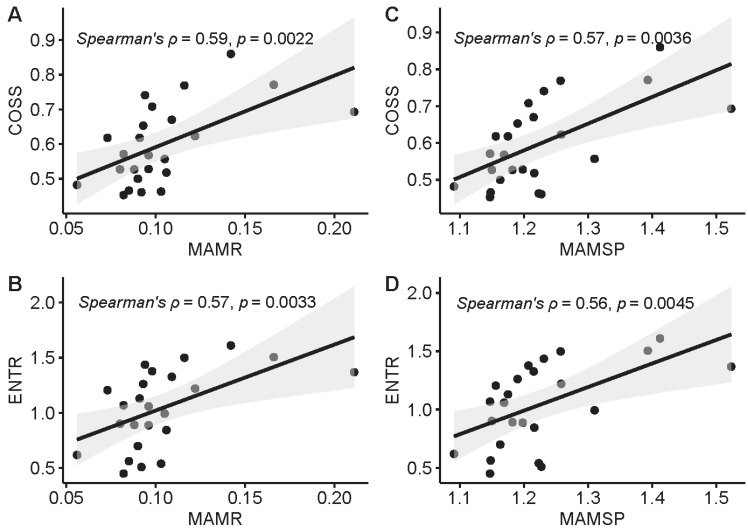

Figure 5. Scatterplot correlation matrix of MAMR vs. COSS, MAMR vs. ENTR, MAMSP vs. COSS, and MAMSP vs. ENTR with regression lines within Slavic languages

Figure 5 shows that the correlations between morphological richness and word order freedom within Slavic languages are also positive, moderate, and statistically significant. The Spearman's rank correlation coefficients between MAMR and COSS, MAMR and ENTR, MAMSP and COSS, and MAMSP and ENTR are 0.59, 0.57, 0.57, and 0.56, respectively.[9] It suggests that the "complexity trade-off hypothesis" still holds as far as only Slavic languages are considered, which further corroborates the negative correlation between morphology and syntax of Slavic languages [Comrie, Corbett (eds.) 1993: 6–7; Klein et al. (eds.) 2018].

Moreover, the Slavic treebanks adopted in the current study fall into 3 subgroups, the East branch (Russian, Belarusian, and Ukrainian), the West branch (Polish, Czech, Slovak, and Upper Sorbian), and the South branch (Bulgarian, Slovenian, Croatian, and Serbian), respectively.[10] The other two languages, Old Church Slavonic and Old Russian,[11] are considered here separately as ancient Slavic languages.

[9] When we focus on the modern Slavic languages only, the Spearman's rank correlations between MAMR and COSS, MAMR and ENTR, MAMSP and COSS, and MAMSP and ENTR are all positive and moderate, though they are not statistically significant (the ρ and p values for the these four correlations are $\rho = 0.43$, $p = 0.051$; $\rho = 0.40$, $p = 0.071$; $\rho = 0.39$, $p = 0.079$, and $\rho = 0.37$, $p = 0.095$). The insignificance of these correlations shows that modern Slavic languages are more closely connected with each other than ancient ones.

[10] According to Janda [2006: 415], and Sussex and Cubberley [2006: 2–6], modern Slavic languages can be divided into three subgroups: (1) the East branch, consisting of Russian, Belarusian, and Ukrainian; (2) the West branch, consisting of Polish, Czech, Slovak, and Sorbian; and (3) the South branch, consisting of Bulgarian, Macedonian, Slovenian, Bosnian, Croatian, and Serbian.

[11] Being the oldest attested Slavic language, Old Church Slavonic dates from the 10th or 11th century, and is now primarily used for religious purposes [Lunt 2001]. Together with Proto-Slavic, Old Church Slavonic is fundamentally important for the understanding of the modern Slavic languages [Sussex, Cubberley 2006: 2–6]. Old Russian, also known as Old East Slavic (or Common East Slavic), was used from the 10th to 15th centuries by the East Slavs. It finally developed into the Russian, Belarusian, Ukrainian and Rusyn languages, etc. [Krause, Slocum 2020].

For a better presentation of the correlations within Slavic languages, we took **Figure 5 (A)** as an example and plotted the correlation between MAMR and COSS with treebank labels, as shown in **Figure 6**.

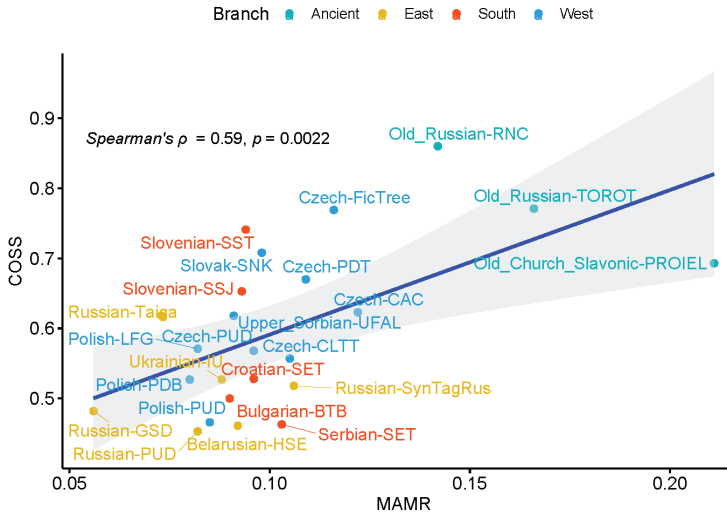

Figure 6. Scatterplot of MAMR and COSS with a regression line within Slavic languages

Figure 6 shows that, along with the regression line, the East branch (yellow dots) tends to appear on the bottom-left corner of the figure, then the South (red) and the West (blue), and finally the ancient languages (green) on the top-right. It is also true for the other three pairs of metrics, i.e., MAMR and ENTR, MAMSP and COSS, and MAMSP and ENTR. The results show that generally, the East, South and West branches of modern Slavic languages demonstrate their particular features with different levels of morphological richness and word order freedom, though the separation between these three branches is not sharp and clear-cut enough. In contrast, the distinctive features of the ancient Slavic languages emerge when compared with the modern ones.

To be specific, in terms of morphological features, the treebanks of Old_Church_Slavonic-PROIEL (a sample of Old Church Slavonic), Old_Russian-TOROT (a sample of Old Russian and Middle Russian) and Old_Russian-RNC (a sample of Middle Russian) have higher MAMR values than the three groups of modern Slavic languages do. It indicates a more complex morphological structure of old Slavic languages or a decreasing trend of the morphological features from ancient to modern ones. Smetonienė [2019] once compared the Slavic loan nouns from Petkevičius' Catechism with their equivalents in Slavic languages of the relevant period and found that patterns of morphological integration can function as an indication of Slavic language origin and development. Results here may provide another angle into the morphological changes of Slavic languages.

When it comes to word order, Old_Russian-RNC and Old_Russian-TOROT own the highest COSS values, and the COSS value of Old_Church_Slavonic-PROIEL ranks fifth among all treebanks. It means that the ancient Slavic languages are also likely to have a more flexible word order. The majority of prior studies on the Slavic word order system paid much attention to the discrete classification of word order types to cluster languages (e.g., [Zimmerling 2012]). In this study, we quantified the word order freedom continuously, and the results, as shown in **Figure 6**, can well reflect the stricter word order of the modern languages and freer word order of the ancient ones.

Results show that although the correlation between morphology and syntax within Slavic languages is not as high as that of all sample treebanks, it is still moderate and significant, confirming

the typological generalization of "negative correlation". Put differently, the morphological richness of Slavic languages decreases as the syntactic structure becomes more rigid. Moreover, the ancient Slavic languages are characterized as being morphologically richer and syntactically more flexible than modern ones.

3.3. Language clustering within Slavic languages

A further concern would be whether the morphological and syntactic metrics in the current study can well cluster Slavic languages into subgroups. To adopt at least one metric of morphological complexity and one of word order freedom, respectively, nine different combinations of these four indicators can be obtained, i.e., (1) MAMR and COSS; (2) MAMR and ENTR; (3) MAMR, COSS and ENTR; (4) MAMSP and COSS; (5) MAMSP and ENTR; (6) MAMSP, COSS and ENTR; (7) MAMR, MAMSP and COSS; (8) MAMR, MAMSP and ENTR; (9) MAMR, MAMSP, COSS and ENTR.

We took these nine combinations as inputs for Agglomerative Clustering Analysis with Euclidean distance. The values of the correlation coefficient produced are (1) 0.65, (2) 0.74, (3) 0.73, (4) 0.57, (5) 0.66, (6) 0.65, (7) 0.78, (8) 0.69 and (9) 0.65, respectively, showing that the cluster trees generated can well reflect the relationship under discussion. It is noteworthy that four among nine combinations, viz., (2) MAMR and ENTR; (3) MAMR, COSS and ENTR; (8) MAMR, MAMSP and ENTR; and (9) MAMR, MAMSP, COSS, and ENTR, can cluster Slavic languages ideally. Moreover, the ancient Slavic languages are always clustered together for all nine combinations. We took the combination (9) as an example in **Figure 7**.

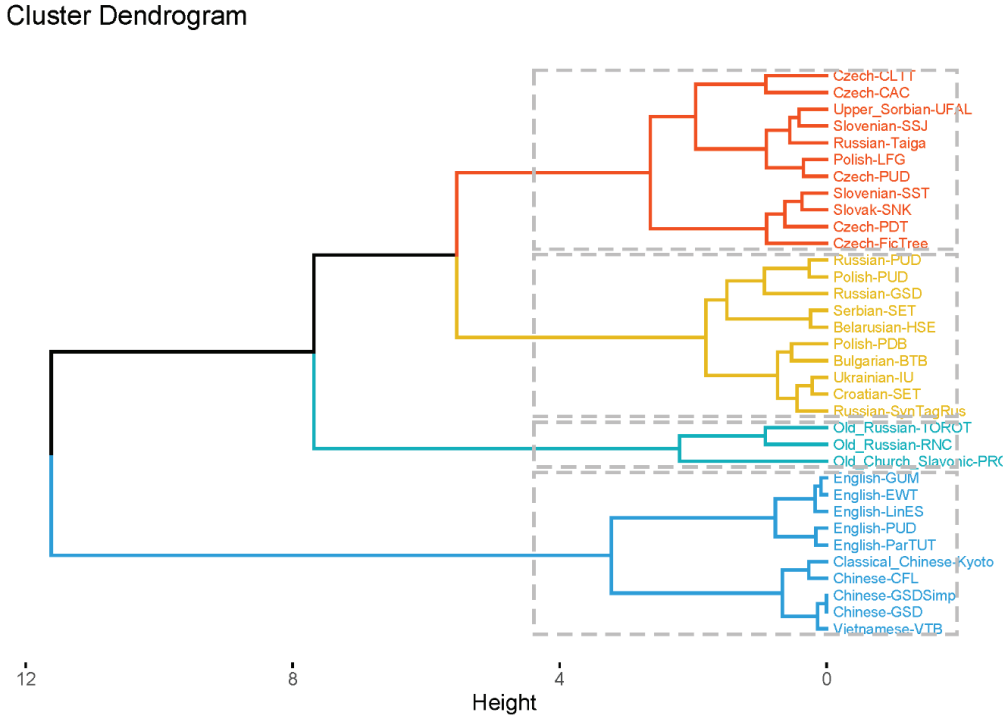

Figure 7. Cluster dendrogram based on MAMR, MAMSP, COSS, and ENTR

Figure 7 shows that the classification model can accurately distinguish Slavic languages from non-Slavic ones (the blue subset). Moreover, the ancient Slavic languages (the green subset: Old_Russian-TOROT, Old_Russian-RNC, and Old_Church_Slavonic-PROIEL) show their uniqueness in the dendrogram, clustering into one category.

In this regard, Liu and Cong [2013] once reported favorable clustering results of 12 Slavic languages based on 15 network parameters of 12 parallel novels. Similarly, though with four parameters and unparalleled materials, results here are generally satisfactory. To be specific, the 11 modern Slavic languages in **Figure 7** are subdivided into the red subset and the yellow subset. The red subset includes all four West Slavic languages, namely, all five treebanks of Czech, the only treebank of Upper Sorbian, one treebank of Polish (the other two are grouped into the East), and the only treebank of Slovak. Meanwhile, the yellow subset includes all three East Slavic languages, i.e., the only treebank of Belarusian, the only treebank of Ukrainian, and three treebanks of Russian (the other one is grouped into the West). Besides, the South branch is mixed into the East and West ones (specifically, one treebank of Croatian, one treebank of Serbian, and one treebank of Bulgarian are grouped into the East branch; two treebanks of Slovenian are grouped into the West).

One interesting question then may arise: Why are ancient Slavic languages, i.e., Old Church Slavonic (a South Slavic liturgical and literary language [Sussex, Cubberley 2006: 2]) and Old Russian (a common parent to the East Slavic languages [Krause, Slocum 2020]), so close to each other? As suggested by Trubetzkoy [1927] and Durnovo [1932], the underlying reason might be that they are variants of one single language, "Late Common Slavic", and their similarities far outweigh their differences. As Lunt [1987: 134] stated, "the differences between standard early Russian and Old Church Slavonic, though striking, simply do not justify the sort of linguistic distance scholars have posited". Precisely, the rich inflectional system of Old Church Slavonic generally "coincides with" that of Old Russian, and the exceptions are "surprisingly minor" [Lunt 1987: 148]. In fact, Old Church Slavonic and Old Russian shared virtually all morphological elements [Lunt 1987: 136]. This might explain why the MAMR and MAMSP values of Old Church Slavonic and Old Russian in **Appendix C** are so close to each other and why they are clustered together in **Figure 7** from the morphological perspective.

As for their similarities in terms of the "free word order", the word "free" means that the position of an element within a sentence is not directly determined by its syntactic function [Mathesius 1942; Firbas 1992: 118]. Although logically, there are six different permutations of subject, object and verb, the Slavic languages are generally classified as free SVO languages [Siewierska, Uhlířová 1998: 107; Dryer 2013]. As put by Plungian [2018: 10–11], all diachronic processes revolve around the frequency of linguistic forms. The relative frequency in **Appendix B** shows that within this dominant SVO order, the ancient Slavic languages exhibit lower proportions of SVO structures (Old_Church_Slavonic-PROIEL — 54.50 %, Old_Russian-RNC — 29.00 %, Old_Russian-TOROT — 45.80 %) and more balanced proportions of other five S/V/O combinations than modern ones (the average SVO proportion of modern Slavic is 70.88 % and the other five S/V/O combinations are rare). The average proportion of SVO structures of the ten non-Slavic treebanks (Vietnamese, Chinese and English) is 95.35 %. Therefore, it not only confirms that ancient Slavic languages share great similarities regarding the flexibility of word order (this is also consistent with the COSS and ENTR values in **Appendix D**), but also demonstrates a tendency of Slavic languages towards strict SVO languages diachronically. Thus, this explains why the Old Church Slavonic and Old Russian are grouped into one subset in **Figure 7** in terms of word order.

Another noteworthy aspect is that the diachronic changes of morphological richness and word order seem to be dynamically adapted with each other. Specifically, with losses of morphological features, modern Slavic languages are becoming more rigid in terms of syntactic word order. Koplenig et al. [2017: 4] once suggested that the trade-offs between morphology and syntax can theoretically be credited as a reflection of the least effort principle [Zipf 1965] or be understood under the framework of synergetic linguistics [Köhler 1987; 2005], indicating that human

beings tend to encode linguistic information efficiently. In this case, if a modern Slavic language uses more fixed word order to convey grammatical relationships, then human cognitive capacity would be overloaded to encode too many morphological rules in word structure; Conversely, when the constraint on word order is more flexible, more cognitive efforts can be spared to encode more information in morphological features. Hence, our empirical results might provide another evidence to this assumption of efficient communication from a diachronic perspective.

To summarize, based on the metrics of morphological richness and word order freedom, subgroups of Slavic languages can be well clustered. More importantly, the dynamic relations between morphological and syntactic features might shed new light on how language evolved diachronically and how human beings encode linguistic information for efficient communication. In all, the results confirm the applicability of morphological and syntactic metrics in language clustering.

3.4. Possible factors that might influence the quantitative results

Finally, it is still of great interest to investigate whether the corpus size (or text size) and corpus genre (or text type) would affect our quantitative results. For example, why are the treebanks Polish-LFG and Russian-Taiga located in different clusters with respect to other Polish and Russian treebanks in **Figure 7**? Due to the limitations of linguistic materials,[12] the effects of corpus size was investigated by subsetting all treebanks into individual sub-corpora, and the effects of corpus genre by delving into two specific treebanks.

First, we will focus on the possible disturbing effects of corpus size. Since the token count of the smallest corpus among our 34 corpora is 6339, we subsetted all 34 corpora cumulatively into sub-corpora with 5000 tokens as steps (5000, 10 000, 15 000, 20 000, … 95 000, 100 000). We then computed the MAMR and MAMSP values for each sub-corpus (W = 500) to compare with the MAMR and MAMSP values of the whole treebanks. Taking the MAMR value of each sub-corpus as an example, we plotted a line chart of all 34 treebanks, as shown in **Figure 8** (p. 149).

Figure 8 shows that the MAMR values of all sub-corpora are generally similar to the MAMR values of the entire treebanks, especially when the token size reaches 50 000. Moreover, to test whether there exists significant difference between MAMRs of different token sizes, we followed Wang and Liu [2017] by generating two linear regression models. The first one was fitted with MAMR and token size as the dependent and independent variables, respectively. The model is not significant (F = 0.07094, df_1 = 1, df_2 = 456, p = 0.7901 > 0.05, adjusted R^2 = −0.002037).[13] The second model predicts MAMRs from token sizes with interaction with different treebanks. It is highly significant (F = 2842, df_1 = 67, df_2 = 390, p = 2.2e-16 < 0.0001, adjusted R^2 = 0.9976). A likelihood ratio test between these two models shows that the regression model with the interaction of treebanks is significantly different from the model without it (p < 0.0001). It shows that differences of MAMRs between treebanks are statistically significant, and the effect of the correlation between MAMRs and the interaction between token sizes and treebanks is strong

[12] On the one hand, the tokens of the corpora under investigation range from 6339 (Chinese-CFL) to 1 285 509 (Czech-PDT) (see **Appendix A**), and the numbers of declarative sentences of the corpora range from 412 (Chinese-CFL) to 84 884 (Czech-PDT) (see **Appendix B**); on the other hand, the genres of the corpora cover wiki, blog, nonfiction, social, news, spoken, reviews, legal, medical, web, grammar-examples, the Bible, and so on. Hence, it is impossible to adopt corpora of the same size with the same genre for the research topic of the current study.

[13] Adjusted R^2 is an important indicator in a significance test of linear regression models. In accordance with Gries [2013: 265], "R^2 is adjusted such that you incur a slight penalty for every predictor included in your model".

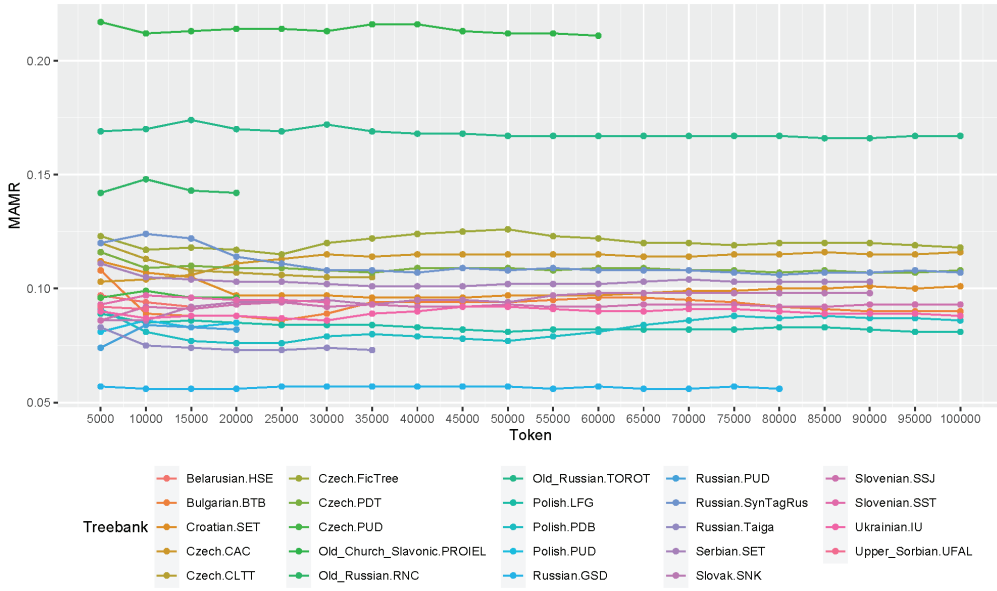

Figure 8. MAMR values of 34 treebanks with different token sizes (5000 ≤ Token ≤ 100 000)

(adjusted $R^2 = 0.9976$), whereas the effect of the correlation between MAMRs and token size is trivial (adjusted $R^2 = 0.002037$). It is also true for the measurements of MAMSP. Specifically, the differences of MAMSPs between treebanks are statistically significant, and the effect of the correlation between MAMSPs and the interaction between token sizes and treebanks is strong ($F = 1254$, $df_1 = 67$, $df_2 = 390$, $p = 2.2e\text{-}16 < 0.05$, adjusted $R^2 = 0.9946$), whereas the effect of the correlation between MAMSPs and token size is small ($F = 0.02776$, $df_1 = 1$, $df_2 = 456$, $p = 0.8678 > 0.05$, adjusted $R^2 = -0.002132$). The results suggest that different treebanks affect the MAMR and MAMSP values to a large degree, and the token sizes have little effects.

Similarly, we subsetted all 34 corpora cumulatively into sub-corpora with 400 sentences as steps (400, 800, 1200, 1600, … 7600, 8000) since the sentence number of the smallest corpus among all 34 corpora is 412. We then computed the COSS and ENTR values of each sub-corpus. We found that the COSS value of sub-corpus is generally similar to the COSS of the whole treebank when the sentence number reaches 4000. Moreover, based on regression models, it can also be found that the differences of COSSs between treebanks are statistically significant, and the effect of the correlation between COSSs and the interaction between sentence numbers and treebanks is strong ($F = 953.5$, $df_1 = 67$, $df_2 = 352$, $p = 2.2e\text{-}16 < 0.05$, adjusted $R^2 = 0.9935$), but the effect of the correlation between COSSs and sentence number is extremely small ($F = 7.709$, $df_1 = 1$, $df_2 = 418$, $p = 0.005741 > 0.05$, adjusted $R^2 = 0.01576$). Also, the differences of ENTRs between treebanks are statistically significant, and the effect of the correlation between ENTRs and the interaction between sentence numbers and treebanks is strong ($F = 1018$, $df_1 = 67$, $df_2 = 352$, $p = 2.2e\text{-}16 < 0.05$, adjusted $R^2 = 0.9939$), whereas the effect of the correlation between ENTRs and sentence number is trivial ($F = 10.75$, $df_1 = 1$, $df_2 = 418$, $p = 0.001131 > 0.05$, adjusted $R^2 = 0.02274$). The results suggest that the values of COSS and ENTR are affected by the treebanks to a large extent. In contrast, sentence numbers have little effect on the COSS and ENTR values.

Therefore, we can see that corpus size in tokens or sentences does not greatly affect the indicators adopted. It might be related to the fact that we have adopted the moving-window operation and the relative-frequency operation (cf. **Methods**) to minimize the effects of corpus size.

Now we will focus on the possible disturbing effects of corpus genre (text type). We will consider Polish-LFG and Russian-Taiga treebanks to discuss the effects of genre on the results of the quantitative metrics.[14]

In **Appendices C** and **D**, it can be found that the morphological richness metrics (MAMR and MAMSP) for Polish-LFG (ranked second among three Polish treebanks in terms of MAMR and ranked third in terms of MAMSP) and Russian-Taiga (ranked third among four Russian treebanks in terms of MAMR and ranked second in terms of MAMSP) are similar to those of other Polish and Russian treebanks. However, the COSS and ENTR values of Polish-LFG and Russian-Taiga are the highest among all Polish and Russian treebanks, respectively. Do the effects of genre contribute to this phenomenon? We computed the COSS and ENTR values of each genre of Polish-LFG, as shown in **Table 2**.

Table 2

COSS and ENTR values of different genres in the Polish-LFG treebank

Genre	COSS	Sentences	Genre	ENTR	Sentences
Legal	0.505	11	Legal	0.530	11
News	0.558	6 090	Academic	0.994	48
Academic	0.560	48	News	1.004	6 090
Fiction	0.562	6 215	Fiction	1.020	6 215
Nonfiction	0.572	1 105	Nonfiction	1.086	1 105
Blog	0.593	121	Blog	1.087	121
Spoken Media	0.660	105	Spoken Media	1.149	105
Social	0.695	374	Social	1.400	374
Spoken Prepared	0.741	263	Spoken Prepared	1.419	263
Spoken Conversational	0.779	581	Spoken Conversational	1.455	581
Whole treebank	**0.571**	**14 913**	**Whole treebank**	**1.068**	**14 913**

Table 2 shows that the COSS values of formal written genres (legal, news, academic, fiction, nonfiction) in Polish-LFG are generally approximant or below the COSS value of the whole treebank. In contrast, those of informal written genre (blog) or spoken genres (spoken media, social, spoken prepared, spoken conversational) are generally above the COSS of the whole treebank. It is also true for ENTR values. The results indicate that formal written genres are prone to demonstrate fixed syntactic structures, whereas informal written and spoken genres are likely to allow more flexibility in terms of word order.

As we can see in **Appendix A**, the other two Polish treebanks (Polish-PDB and Polish-PUD) are generally composed of formal written genres (one is composed of fiction, news, nonfiction; and the other of news and wiki). Hence, it is the factor of genre that renders Polish-LFG different from the other Polish treebanks, as shown in **Figure 7**.

Similarly, we also examined the syntactic features of each genre of Russian-Taiga in **Table 3**.

[14] Among all 34 treebanks, three treebanks (i.e., Belarusian-HSE, Polish-LFG and Russian-Taiga) are annotated with comments like "# genre = news", which makes it possible to extract sub-corpora of specific genres from these treebanks. However, the Belarusian-HSE treebank is too small (11 250 tokens, 617 sentences), not to mention the fact that it needs to be divided into four different genres. Hence, we focused on Polish-LFG and Russian-Taiga to investigate the possible effects of corpus genre on quantitative metrics.

Table 3

COSS and ENTR values of different genres in the Russian-Taiga treebank

Genre	COSS	Sentences	Genre	ENTR	Sentences
News	0.438	15	News	0.257	15
Social	0.569	1 846	Social	1.046	1 846
Poetry	0.707	785	Poetry	1.384	785
Whole treebank	**0.618**	**2 646**	**Whole treebank**	**1.205**	**2 646**

Table 3 shows that, as a formal written genre, the news genre in Russian-Taiga has fixed word order with extremely low COSS and ENTR values (although the sentence number of this genre is small). The genre "social", as a spoken genre, does show its flexible syntax. Finally, being a written genre, poetry is featured by greater word order flexibility, though. It is consistent with the prior finding that "their [poetries'] forms of language are similar to the spoken language and their syntax is flexible as they permit themselves considerable freedom in word order for different purposes" [Nofal 2014: 283]. Therefore, the overwhelming proportion of social and poetry genres in the Russian-Taiga treebank results in high indicators of word order freedom. This makes it different from the other three Russian treebanks (Russian-GSD, Russian-PUD, Russian-SynTagRus), classified into a different cluster, as shown in **Figure 7**.

Hence, based on the treebanks Polish-LFG and Russian-Taiga, we investigated the effects of corpus genre on our quantitative measures. It was found that genres do not significantly affect morphological metrics, but they tend to do so with regard to syntactic indicators. Generally, formal written genres tend to have more fixed word order, while informal and spoken ones are likely to be more syntactically flexible. However, due to the limited genre annotation of the treebanks, we only examined two treebanks in this section. The findings above are interesting and worthy of further investigations.

4. Conclusion

This study adopts 24 treebanks of 13 Slavic languages in UD 2.5 database as the research object, and ten treebanks of typical analytic languages (Vietnamese, Classical Chinese), a highly analytic language (Standard Chinese) and the most analytic among Indo-European languages (English) as comparative materials. It investigates the correlation between morphological richness and word order freedom within Slavic languages and its possible implications.

Based on four quantitative metrics, morphological and syntactic features are extracted and quantified from annotated corpora. The statistical results show that the indicators adopted are consistent and reliable. Moreover, it can be found that the well-known "trade-off hypothesis" holds within Slavic languages, viz., the morphological richness and the rigidity of word order in Slavic languages are negatively related. Besides, we also examined the clustering of Slavic languages based on the above indicators from a practical perspective. It was found that the combination of morphological and syntactic metrics can differentiate the Slavic languages from the non-Slavic ones. More importantly, the ancient Slavic languages demonstrate morphological and syntactic characteristics which clearly distinguish them from the modern ones. The diachronic changes of morphological richness and word order might reflect how human beings encode linguistic information morphologically and syntactically. Finally, we also investigated the effects of two factors on the values of quantitative metrics. It was found that the corpus size does not have a great effect on the indicators used in this study. In contrast, the corpus genre might greatly affect the syntactic indicators, viz., the more formal the genre is, the more fixed the word order.

This study provides evidence for the "negative correlation hypothesis" or the "complexity trade-off hypothesis". With special attention paid to Slavic languages, it examines the correlation between morphological typology and word order typology, applies quantitative metrics to an annotated database, provides insights into adopting treebanks as resources for typological research, shedding new light on how natural languages encode lexical and syntactic information for efficient communication.

For future studies, paralleled corpora with more balanced genre distribution, covering more levels of the language are highly desirable to uncover the dynamic relations of different components of human languages.

Appendix A: Information on language groups, text types (or genres), and token counts of the treebanks

#	Treebank	Group	Text Type	Tokens
1	Belarusian-HSE	Slavic	fiction, legal, news, nonfiction	11 250
2	Bulgarian-BTB	Slavic	fiction, legal, news	134 091
3	Chinese-CFL	Sino-Tibetan	learner-essay	6 339
4	Chinese-GSD	Sino-Tibetan	wiki	106 226
5	Chinese-GSDSimp	Sino-Tibetan	wiki	106 203
6	Classical_Chinese-Kyoto	Sino-Tibetan	nonfiction	74 770
7	Croatian-SET	Slavic	news, web, wiki	175 244
8	Czech-CAC	Slavic	legal, medical, news, nonfiction, reviews	434 256
9	Czech-CLTT	Slavic	legal	31 106
10	Czech-FicTree	Slavic	fiction	135 261
11	Czech-PDT	Slavic	news, nonfiction, reviews	1 285 509
12	Czech-PUD	Slavic	news, wiki	15 986
13	English-EWT	Germanic	blog, email, reviews, social	224 964
14	English-GUM	Germanic	academic, fiction, news, nonfiction, spoken, web, wiki	88 128
15	English-LinES	Germanic	fiction, nonfiction, spoken	82 712
16	English-ParTUT	Germanic	legal, news, wiki	43 837
17	English-PUD	Germanic	news, wiki	18 725
18	Old_Church_Slavonic-PROIEL	Slavic	bible	57 563
19	Old_Russian-RNC	Slavic	legal, nonfiction	15 762
20	Old_Russian-TOROT	Slavic	legal, nonfiction	149 780
21	Polish-LFG	Slavic	fiction, news, nonfiction, social, spoken	105 147
22	Polish-PDB	Slavic	fiction, news, nonfiction	292 133
23	Polish-PUD	Slavic	news, wiki	15 731
24	Russian-GSD	Slavic	wiki	79 875

#	Treebank	Group	Text Type	Tokens
25	Russian-PUD	Slavic	news, wiki	16 378
26	Russian-SynTagRus	Slavic	fiction, news, nonfiction	904 227
27	Russian-Taiga	Slavic	news, poetry, social	32 182
28	Serbian-SET	Slavic	news	85 333
29	Slovak-SNK	Slavic	fiction, news, nonfiction	86 913
30	Slovenian-SSJ	Slavic	fiction, news, nonfiction	122 072
31	Slovenian-SST	Slavic	spoken	28 026
32	Ukrainian-IU	Slavic	blog, email, fiction, grammar-examples, legal, news, reviews, social, web, wiki	98 974
33	Upper_Sorbian-UFAL	Slavic	nonfiction, wiki	9 175
34	Vietnamese-VTB	Viet-Muong	news	37 431

Appendix B: Probability (or relative frequency) of S/V/O word order combinations and numbers of declarative sentences in the treebanks

#	Treebank	SVO	SOV	VSO	OSV	OVS	VOS	Sentences
1	Belarusian-HSE	0.884	0.028	NA	0.023	0.051	0.014	617
2	Bulgarian-BTB	0.807	0.061	0.002	0.008	0.106	0.016	10 382
3	Chinese-CFL	0.995	0.003	NA	0.003	NA	NA	412
4	Chinese-GSD	0.929	0.028	NA	0.043	NA	NA	4 981
5	Chinese-GSDSimp	0.929	0.028	NA	0.043	NA	NA	4 981
6	Classical_Chinese-Kyoto	0.968	0.026	NA	0.006	NA	NA	15 115
7	Croatian-SET	0.764	0.072	0.014	0.058	0.073	0.019	8 866
8	Czech-CAC	0.630	0.058	0.077	0.032	0.132	0.071	24 440
9	Czech-CLTT	0.717	0.024	0.099	0.037	0.102	0.022	1 125
10	Czech-FicTree	0.457	0.191	0.045	0.085	0.147	0.075	10 990
11	Czech-PDT	0.570	0.083	0.068	0.042	0.169	0.069	84 884
12	Czech-PUD	0.705	0.085	0.057	0.033	0.085	0.035	985
13	English-EWT	0.954	NA	0.001	0.038	0.004	0.003	14 404
14	English-GUM	0.947	0.001	0.001	0.047	0.003	0.001	5 106
15	English-LinES	0.942	0.002	0.001	0.044	0.008	0.003	4 867
16	English-ParTUT	0.959	0.002	0.001	0.035	NA	0.002	2 046
17	English-PUD	0.964	NA	0.001	0.023	0.004	0.008	986
18	Old_Church_Slavonic-PROIEL	0.545	0.109	0.107	0.041	0.043	0.155	6 335

#	Treebank	SVO	SOV	VSO	OSV	OVS	VOS	Sentences
19	Old_Russian-RNC	0.290	0.301	0.110	0.175	0.049	0.075	583
20	Old_Russian-TOROT	0.458	0.131	0.181	0.045	0.066	0.118	16 942
21	Polish-LFG	0.700	0.078	0.031	0.031	0.086	0.073	14 913
22	Polish-PDB	0.767	0.046	0.031	0.024	0.067	0.065	20 574
23	Polish-PUD	0.875	0.011	0.031	0.009	0.039	0.035	987
24	Russian-GSD	0.841	0.010	0.006	0.026	0.099	0.018	5 004
25	Russian-PUD	0.899	0.008	0.009	0.018	0.059	0.008	987
26	Russian-SynTagRus	0.780	0.042	0.010	0.041	0.089	0.038	57 496
27	Russian-Taiga	0.635	0.137	0.060	0.067	0.069	0.032	2 646
28	Serbian-SET	0.878	0.018	0.014	0.030	0.053	0.006	4 308
29	Slovak-SNK	0.518	0.152	0.035	0.053	0.184	0.057	9 721
30	Slovenian-SSJ	0.585	0.113	0.035	0.076	0.167	0.024	7 661
31	Slovenian-SST	0.485	0.178	0.055	0.137	0.121	0.024	2 886
32	Ukrainian-IU	0.766	0.069	0.013	0.055	0.068	0.030	6 424
33	Upper_Sorbian-UFAL	0.622	0.162	0.141	0.012	0.046	0.017	641
34	Vietnamese-VTB	0.948	0.018	0.000	0.027	NA	0.007	2 783

Appendix C: Values of two metrics of morphological richness for each treebank

Treebank	MAMR	Group	Treebank	MAMSP	Group
Old_Church_Slavonic-PROIEL	0.211	Slavic	Old_Church_Slavonic-PROIEL	1.523	Slavic
Old_Russian-TOROT	0.166	Slavic	Old_Russian-RNC	1.412	Slavic
Old_Russian-RNC	0.142	Slavic	Old_Russian-TOROT	1.393	Slavic
Czech-CAC	0.122	Slavic	Czech-CLTT	1.310	Slavic
Czech-FicTree	0.116	Slavic	Czech-CAC	1.258	Slavic
Czech-PDT	0.109	Slavic	Czech-FicTree	1.257	Slavic
Russian-SynTagRus	0.106	Slavic	Slovenian-SST	1.231	Slavic
Czech-CLTT	0.105	Slavic	Belarusian-HSE	1.227	Slavic
Serbian-SET	0.103	Slavic	Serbian-SET	1.223	Slavic
Slovak-SNK	0.098	Slavic	Russian-SynTagRus	1.216	Slavic
Croatian-SET	0.096	Slavic	Czech-PDT	1.215	Slavic
Czech-PUD	0.096	Slavic	Slovak-SNK	1.207	Slavic
Slovenian-SST	0.094	Slavic	Croatian-SET	1.198	Slavic
Slovenian-SSJ	0.093	Slavic	Slovenian-SSJ	1.190	Slavic

Treebank	MAMR	Group	Treebank	MAMSP	Group
Belarusian-HSE	0.092	Slavic	Ukrainian-IU	1.182	Slavic
Upper_Sorbian-UFAL	0.091	Slavic	Upper_Sorbian-UFAL	1.175	Slavic
Bulgarian-BTB	0.090	Slavic	Czech-PUD	1.169	Slavic
Ukrainian-IU	0.088	Slavic	Bulgarian-BTB	1.163	Slavic
Polish-PUD	0.085	Slavic	English-LinES	1.160	Germanic
Polish-LFG	0.082	Slavic	Russian-Taiga	1.156	Slavic
Russian-PUD	0.082	Slavic	English-EWT	1.153	Germanic
Polish-PDB	0.080	Slavic	Polish-PDB	1.150	Slavic
Russian-Taiga	0.073	Slavic	English-GUM	1.149	Germanic
Russian-GSD	0.056	Slavic	Polish-PUD	1.148	Slavic
English-LinES	0.049	Germanic	Russian-PUD	1.147	Slavic
English-GUM	0.047	Germanic	Polish-LFG	1.147	Slavic
English-EWT	0.044	Germanic	English-ParTUT	1.114	Germanic
English-ParTUT	0.036	Germanic	English-PUD	1.096	Germanic
English-PUD	0.035	Germanic	Russian-GSD	1.091	Slavic
Chinese-CFL	0.001	Sino-Tibetan	Chinese-CFL	1.002	Sino-Tibetan
Chinese-GSD	0.001	Sino-Tibetan	Chinese-GSDSimp	1.001	Sino-Tibetan
Chinese-GSDSimp	0.001	Sino-Tibetan	Chinese-GSD	1.001	Sino-Tibetan
Classical_Chinese-Kyoto	0.000	Sino-Tibetan	Vietnamese-VTB	1.000	Viet-Muong
Vietnamese-VTB	0.000	Viet-Muong	Classical_Chinese-Kyoto	1.000	Sino-Tibetan

(Note: Ranked in descending order according to the MAMR and MAMSP values.)

Appendix D: Values of two metrics of word order freedom for each treebank

Treebank	COSS	Group	Treebank	ENTR	Group
Old_Russian-RNC	0.860	Slavic	Old_Russian-RNC	1.610	Slavic
Old_Russian-TOROT	0.771	Slavic	Old_Russian-TOROT	1.505	Slavic
Czech-FicTree	0.769	Slavic	Czech-FicTree	1.499	Slavic
Slovenian-SST	0.741	Slavic	Slovenian-SST	1.436	Slavic
Slovak-SNK	0.708	Slavic	Slovak-SNK	1.377	Slavic
Old_Church_Slavonic-PROIEL	0.693	Slavic	Old_Church_Slavonic-PROIEL	1.368	Slavic
Czech-PDT	0.670	Slavic	Czech-PDT	1.327	Slavic
Slovenian-SSJ	0.653	Slavic	Slovenian-SSJ	1.261	Slavic

Treebank	COSS	Group	Treebank	ENTR	Group
Czech-CAC	0.623	Slavic	Czech-CAC	1.220	Slavic
Upper_Sorbian-UFAL	0.618	Slavic	Russian-Taiga	1.205	Slavic
Russian-Taiga	0.618	Slavic	Upper_Sorbian-UFAL	1.130	Slavic
Polish-LFG	0.571	Slavic	Polish-LFG	1.068	Slavic
Czech-PUD	0.568	Slavic	Czech-PUD	1.059	Slavic
Czech-CLTT	0.557	Slavic	Czech-CLTT	0.993	Slavic
Croatian-SET	0.528	Slavic	Polish-PDB	0.901	Slavic
Ukrainian-IU	0.527	Slavic	Ukrainian-IU	0.891	Slavic
Polish-PDB	0.527	Slavic	Croatian-SET	0.887	Slavic
Russian-SynTagRus	0.518	Slavic	Russian-SynTagRus	0.844	Slavic
Bulgarian-BTB	0.500	Slavic	Bulgarian-BTB	0.699	Slavic
Russian-GSD	0.482	Slavic	Russian-GSD	0.618	Slavic
Polish-PUD	0.466	Slavic	Polish-PUD	0.562	Slavic
Serbian-SET	0.463	Slavic	Serbian-SET	0.539	Slavic
Belarusian-HSE	0.461	Slavic	Belarusian-HSE	0.509	Slavic
Russian-PUD	0.453	Slavic	Russian-PUD	0.450	Slavic
Chinese-GSDSimp	0.439	Sino-Tibetan	Chinese-GSDSimp	0.305	Sino-Tibetan
Chinese-GSD	0.439	Sino-Tibetan	Chinese-GSD	0.305	Sino-Tibetan
English-LinES	0.433	Germanic	English-LinES	0.270	Germanic
Vietnamese-VTB	0.431	Viet-Muong	Vietnamese-VTB	0.258	Viet-Muong
English-GUM	0.431	Germanic	English-GUM	0.234	Germanic
English-EWT	0.427	Germanic	English-EWT	0.213	Germanic
English-ParTUT	0.425	Germanic	English-ParTUT	0.193	Germanic
English-PUD	0.423	Germanic	English-PUD	0.190	Germanic
Classical_Chinese-Kyoto	0.422	Sino-Tibetan	Classical_Chinese-Kyoto	0.158	Sino-Tibetan
Chinese-CFL	0.410	Sino-Tibetan	Chinese-CFL	0.037	Sino-Tibetan

(Note: Ranked in descending order according to the COSS and ENTR values.)

REFERENCES

Abeillé (ed.) 2003 — Abeillé A. (ed.). *Treebanks: Building and using parsed corpora*. Dordrecht: Kluwer Academic Publ., 2003.

Alzetta et al. 2019 — Alzetta C., Dell'Orletta F., Montemagni S., Venturi G. Inferring quantitative typological trends from multilingual treebanks. A case study. *Lingue e linguaggio*, 2019, XVIII(2): 209–242.

Bentz et al. 2017 — Bentz C., Alikaniotis D., Cysouw M., Ferrer-i-Cancho R. The entropy of words. Learnability and expressivity across more than 1000 languages. *Entropy*, 2017, 19(6): 1–32.

Bonfante et al. 2018 — Bonfante G., Guillaume B., Perrier G. *Application of graph rewriting to natural language processing*. Hoboken (NJ): John Wiley & Sons Inc., 2018.

Čech, Kubát 2018 — Čech R., Kubát M. Morphological richness of text. *Taming the corpus: From inflection and lexis to interpretation*. Fidler M., Cvrček V. (eds.). Cham: Springer, 2018, 63–77. DOI: 10.1007/978-3-319-98017-1_4.

Chen et al. 2016 — Chen R., Liu H., Altmann G. Entropy in different text types. *Digital scholarship in the humanities*, 2016, 32(3): 528–542.

Coloma 2017 — Coloma G. The existence of negative correlation between linguistic measures across languages. *Corpus Linguistics and Linguistic Theory*, 2017, 13(1): 1–26.

Comrie, Corbett (eds.) 1993 — Comrie B., Corbett G. G. (eds.). *The Slavonic languages*. London: Routledge, 1993.

Courtin 2018 — Courtin M. *Mesures de distances syntaxiques entre langues àpartir de treebanks*. Paris: Université Paris III — Sorbonne Nouvelle, 2018.

Covington, McFall 2010 — Covington M. A., McFall J. D. Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 2010, 17(2): 94–100.

Dryer 2013 — Dryer M. S. Order of subject, object and verb. *The world atlas of language structures online*. Dryer M. S., Haspelmath M. (eds.). Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. http://wals.info/chapter/81 (accessed 12 April 2020).

Durnovo 1932 — Дурново Н. Н. К вопросу о времени распада общеславянского языка. [Durnovo N. N. On the time of the split of Common Slavic.] *Sborník prací I. sjezdu slovanských filologů v Praze*, 1932, 514–526.

Fenk-Oczlon, Fenk 2014 — Fenk-Oczlon G., Fenk A. Complexity trade-offs do not prove the equal complexity hypothesis. *Poznań Studies in Contemporary Linguistics*, 2014, 50(2): 145–155.

Firbas 1992 — Firbas J. *Functional sentence perspective in written and spoken communication*. Cambridge: Cambridge Univ. Press, 1992.

Futrell et al. 2015 — Futrell R., Mahowald K., Gibson E. Quantifying word order freedom in dependency corpora. *Proc. of the 3rd International Conf. on Dependency Linguistics (Depling 2015)*. Nivre J., Hajičová E. (eds.). Uppsala: Uppsala Univ., 2015, 91–100.

Greenberg 1960 — Greenberg J. H. A quantitative approach to the morphological typology of language. *International Journal of American Linguistics*, 1960, 26(3): 178–194.

Greenberg 1963 — Greenberg J. H. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*. Greenberg J. H. (ed.). Cambridge (MA): MIT Press, 1963, 73–113.

Gries 2013 — Gries S. T. *Statistics for linguistics with R*. Berlin: De Gruyter, 2013.

Gulordava, Merlo 2015 — Gulordava K., Merlo P. diachronic trends in word order freedom and dependency length in dependency-annotated corpora of Latin and Ancient Greek. *Proc. of the 3rd International Conf. on Dependency Linguistics (Depling 2015)*. Nivre J., Hajičová E. (eds.). Uppsala: Uppsala Univ., 2015, 121–130.

Gutierrez-Vasques, Mijangos 2018 — Gutierrez-Vasques X., Mijangos V. Comparing morphological complexity of Spanish, Otomi and Nahuatl. *Proc. of the Workshop on Linguistic Complexity and Natural Language Processing*. Becerra-Bonache L., Jiménez-López M. D., Martín-Vide C., Torrens-Urrutia A. (eds.). Santa Fe (NM): Association for Computational Linguistics, 2018, 30–37.

Hajič 1998 — Hajič J. Building a syntactically annotated corpus: The Prague dependency treebank. *Issues of valency and meaning*. Hajičová E. (ed.). Prague: Charles Univ. Press, 1998, 106–132.

Heringer 1993 — Heringer H. J. Dependency syntax: Basic ideas and the classical model. *Syntax: An international handbook of contemporary research*. Vol. 1. Jacobs J., von Stechow A., Sternefeld W., Vennemann T. (eds.). Berlin: Walter de Gruyter, 1993, 298–316.

Hudson 1995 — Hudson R. Measuring syntactic difficulty. Ms., 1995. https://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf.

Jakobson 1936 — Jakobson R. Beitrag zur allgemeinen Kasuslehre: Gesamtbedeutungen der russischen Kasus. *Travaux du Cercle Linguistique de Prague*, 1936, 4: 240–288.

Janda 2006 — Janda L. A. Slavic languages. *Encyclopedia of language and linguistics*. Brown K. (ed.). Amsterdam: Elsevier, 2006, 415–418.

Jiang, Liu (eds.) 2018 — Jiang J., Liu H. (eds.). *Quantitative analysis of dependency structures*. Berlin: De Gruyter, 2018.

Kelih 2010 — Kelih E. The type-token relationship in Slavic parallel texts. *Glottometrics*, 2010, 20: 1–11.

Klein et al. (eds.) 2018 — Klein J., Joseph B., Fritz M. (eds.). *Handbook of comparative and historical Indo-European linguistics*. Berlin: De Gruyter Mouton, 2018.

Köhler 1987 — Köhler R. System theoretical linguistics. *Theoretical Linguistics,* 1987, 14(2–3): 241–258.

Köhler 2005 — Köhler R. Synergetic linguistics. *Quantitative linguistics: An international handbook*. Köhler R., Altmann G., Piotrowski R. G. (eds.). Berlin: De Gruyter, 2005, 760–774.

Koplenig et al. 2017 — Koplenig A., Meyer P., Wolfer S., Müller-Spitzer C. The statistical trade-off between word order and word structure: Large-scale evidence for the principle of least effort. *PLOS ONE*, 2017, 12(3): e0173614.

Krause, Slocum 2020 — Krause T. B., Slocum J. Online lessons at the Linguistics Research Center at the University of Texas at Austin: Old Russian. 2020.

Kuboň et al. 2016 — Kuboň V., Lopatková M., Hercig T. Searching for a measure of word order freedom. *Proc. of the 16th ITAT Conf. Information Technologies — Applications and Theory (Tatranské Matliare, 2016)*. Brejová B. (ed.). CreateSpace Independent Publishing Platform, 2016, 11–17. http://ceur-ws.org/Vol-1649/11.pdf.

Levshina 2019 — Levshina N. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 2019, 23(3): 533–572.

Li, Han 2013 — Li B., Han L. Distance weighted cosine similarity measure for text classification. *Intelligent Data Engineering and Automated Learning — IDEAL 2013.* Yin H., Tang K., Gao Y., Klawonn F., Lee M., Li B., Weise Th., Yao X. (eds.). Berlin: Springer, 2013, 611–618.

Liu 2010 — Liu H. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 2010, 120(6): 1567–1578.

Liu, Cong 2013 — Liu H., Cong J. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 2013, 58(10): 1139–1144.

Liu, Xu 2012 — Liu H., Xu C. Quantitative typological analysis of Romance languages. *Poznań Studies in Contemporary Linguistics*, 2012, 48(4): 597–625.

Lunt 1987 — Lunt H. G. On the relationship of Old Church Slavonic to the written language of early Rus'. *Russian Linguistics*, 1987, 11(2/3): 133–162.

Lunt 2001 — Lunt H. G. *Old Church Slavonic grammar*. New York: Mouton de Gruyter, 2001.

Mathesius 1942 — Mathesius V. Ze srovnávacích studií slovosledných. *Časopis pro moderní filologii*, 1942, 28: 181–190, 302–307.

Maučec, Brest 2019 — Maučec M. S., Brest J. Slavic languages in phrase-based statistical machine translation: A survey. *Artificial Intelligence Review*, 2019, 51(1): 77–117.

McFadden 2003 — McFadden T. On morphological case and word-order freedom. *Proc. of the 29th Annual Meeting of the Berkeley Linguistics Society. General session and parasession on phonetic sources of phonological patterns: Synchronic and diachronic explanations*. Nowak P. M., Yoquelet C., Mortensen D. (eds.). Berkeley (CA): Sheridan Books, 2003, 295–306.

Mel'čuk 1988 — Mel'čuk I. *Dependency syntax: Theory and practice*. Albany (NY): State Univ. of New York Press, 1988.

Muflikhah, Baharudin 2009 — Muflikhah L., Baharudin B. Document clustering using concept space and cosine similarity measurement. *ICCTD 2009 — 2009 International Conf. on Computer Technology and Development*. Jusoff H. K., Othman M., Xie Y. (eds.). Institute of Electrical and Electronic Engineers, 2009, 58–62.

Nofal 2014 — Nofal K. H. Syntactic deviations / stylistic variants in poetry: Chaucer and T. S. Eliot as models. *International Journal of English Language and Literature Studies*, 2014, 3(4): 283–310.

Plungian 2018 — Плунгян В. А. Лингвистика в XXI веке: проблемы, перспективы, точки роста. *Слово.ру: Балтийский акцент*, 2018, 9(1): 7–12. [Plungian V. A. Linguistics in the 21st century: Problems, Prospects, and Growth Points. *Slovo.ru: Baltic Accent*, 2018, 9(1): 7–12.]

Popescu, Altmann 2008 — Popescu I-I., Altmann G. Hapax legomena and language typology. *Journal of Quantitative Linguistics*, 2008, 15(4): 370–378.

Sapir 1921 — Sapir E. *Language: An introduction to the study of speech*. New York: Harcourt, Brace, 1921.

Shannon 1948 — Shannon C. E. A mathematical theory of communication. *Bell System Technical Journal*, 1948, 27(4): 623–656.

Shosted 2006 — Shosted R. K. Correlating complexity: A typological approach. *Linguistic Typology*, 2006, 10(1): 1–40.

Siewierska, Uhlířová 1998 — Siewierska A., Uhlířová L. An overview of word order in Slavic languages. *Constituent order in the languages of Europe*. Siewierska A. (ed.). Berlin: Mouton de Gruyter, 1998, 105–150.

Sinnemäki 2014 — Sinnemäki K. Complexity trade-offs: A case study. *Measuring grammatical complexity*. Newmeyer F. J., Preston L. B. (eds.). New York: Oxford Univ. Press, 2014, 179–201.

Smetonienė 2019 — Smetonienė A. Patterns of morphological integration of Slavic loan nouns in Petkevičius' Catechism (1598) as an indication of their origin and chronology. *Studia z Filologii Polskiej i Słowiańskiej*, 2019, 54. DOI: 10.11649/sfps.1766.

Sussex, Cubberley 2006 — Sussex R., Cubberley P. *The Slavic languages*. Cambridge: Cambridge Univ. Press, 2006.

Tesnière 1959 — Tesnière L. *Eléments de la syntaxe structurale*. Paris: Klincksieck, 1959.

Tesnière 2015 — Tesnière L. *Elements of structural syntax.* Transl. from French by Osborne T., Kahane S. Amsterdam: John Benjamins, 2015.

Trubetzkoy 1927 — Трубецкой Н. С. *К проблеме русского самопознания*. Париж: Евразийское книгоизд-во, 1927. [Trubetzkoy N. S. *K probleme russkogo samopoznaniya* [On the problem of Russian self-awareness]. Paris: Eurasian Publishing House, 1927.]

Wang, Liu 2017— Wang Y., Liu H. The effects of genre on dependency distance and dependency direction. *Language Sciences*, 2017, 59(866): 135–147.

Whitney 1889 — Whitney W. D. *Sanskrit Grammar*. Cambridge (MA): Harvard Univ. Press, 1889.

Xanthos, Gillis 2010— Xanthos A., Gillis S. Quantifying the development of inflectional diversity. *First Language*, 2010, 30(2): 175–198.

Xanthos, Guex 2015 — Xanthos A., Guex G. On the robust measurement of inflectional diversity. *Recent contributions to quantitative linguistics.* Tuzzi A., Benešová M., Macutek J. (eds.). Berlin: De Gruyter Mouton, 2015, 241–254.

Xanthos et al. 2011 — Xanthos A., Laaha S., Gillis S., Stephany U., Aksu-Koç A., Christofidou A., Gagarina N., Hrzica G., Ketrez F. N., Kilani-Schoch M. et al. On the role of morphological richness in the early development of noun and verb inflection. *First Language*, 2011, 31(4): 461–479.

Zeman et al. 2019 — Zeman D., Nivre J., Abrams M., Aepli N., Agić Ž., Ahrenberg L., Aleksandravičiūtė G., Antonsen L., Aplonova K., Aranzabe M. J. et al. Universal Dependencies 2.5. Universal Dependecies Consortium, 2019. https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3105.

Zimmerling 2012 — Циммерлинг А. В. Системы порядка слов в славянских языках. *Вопросы языкознания*, 2012, 5: 3–37. [Zimmerling A. V. Word-order systems in Slavic languages. *Voprosy Jazykoznanija*, 2012, 5: 3–37.]

Zipf 1965 — Zipf G. K. *Human behavior and the principle of least effort: An introduction to human ecology*. New York: Hafner Publishing Company, 1965.

Получено / received 31.07.2020 *Принято / accepted 16.03.2021*