Complexity trade-off in agglutinative languages

Wenchao Li (Zhejiang University) †

膠着型言語における複雑さのトレードオフ

李　文超 (浙江大学)

Abstract

This study utilised 25 treebanks of 16 agglutinative languages across nine language families, exploring the correlation between morphological richness and word order flexibility. Morphological richness was measured at moving-average morphological richness and moving-average mean size of paradigm. Cosine similarity and entropy were employed to capture word-order flexibility. Statistical analysis revealed that the richer the morphology, the more flexible the word order. A fairly strong correlation between morphology and syntax was confirmed, supporting the complexity trade-off hypothesis. Among nine language families, Austronesian, Afroasiatic, Dravidian languages were found morphologically and syntactically less diverse than Altaic, Indo-Aryan, and Uralic languages. Turkish, Uyghur, Basque, and North Sami showed the most balanced proportions of S, V, O combinations. Further, there are separations within the same language family, i.e. the Uralic Finnic branch was found more flexible than Finno-Ugric; the Altaic Mongolic branch was found more rigid than the Turkic branch.

Keywords: morphological richness, word order flexibility, natural language processing, mathematical linguistics

1. Introduction

Human language is a complex, dynamic, and hierarchically organised but regular system (Fenk-Oczlon and Pilz 2021). One such regularity is *complex trade-off*, i.e. if one component (e.g. phonology, morphology, syntax, semantics) in a language is sophistic, then another component is likely to be simplified, allowing languages to manifest roughly the same degree of complexity (Menzerath's law 1954; Shosted 2006; Sinnemäki 2014; Fenk-Oczlon and Fenk 2014). Complexity trade-off hypothesis has been addressed from a variety of linguistic components, i.e. phoneme-syllable-word (Coloma 2017), phonology-morphology (Shosted 2006), morphology-syntax (Sapir 1921; Jakobson 1936; McFadden 2003; Koplenig 2017; Yan and Liu 2021; Li, Liu and Xiong 2022). A crucial issue in the *trade-off* idea resides in: to what extent that trade-off holds, i.e. is it a case of piece-meal or whole-meal? Among the numbers of correlations, which linguistic component is the key factor? Regarding the first issue, most studies deemed that trade-off is some-participate, e.g. the more syllable per word, the fewer phonemes per syllable; the more phonemes per syllable, the fewer morphological cases, the more morphological cases, the more free of word order (Fenk-Oczlon and Fenk 1999; Shosted 2006; Sinnemäki 2008, 2014; Miestamo 2009; Koplenig et al. 2017). In terms of the second issue, after examining parallel data of translation, Fenk-Oczlon and Pilz (2021) contend that syllable complexity is a key factor in the correlations between phoneme inventory size, syllable size, length of words, length of clauses and population size.

This study targets 16 agglutinative languages, i.e. Basque, North Sami, Estonian, Finnish,

† widelia@zju.edu.cn

Hungarian, Japanese, Marathi, Tamil, Telugu, Turkish, Uyghur, Kazakh, Buryat, Wolof, Indonesian and Coptic. 23 annotated corpora from Universal Dependencies (of spoken, written and different genres) were analysed. By combining natural language processing technology that helps tokenisation, lemmatisation, POS and morphological features tagging, dependency parsing, this study aims to draw an insight of the associations between morphological diversity and word order flexibility of natural languages. Morphological richness is measured via two metrics: moving-average morphological richness (MAMR: Čech and Kubát 2018) and moving-average mean size of paradigm (MAMSP: Xanthos and Gillis 2010). Cosine similarity (COS) and entropy (ENTR: Shannon 1948, Chen et al. 2016, Bentz et al. 2017, Yan and Liu 2021, Li et al. 2022) are employed to capture word-order flexibility. Spearman's rank correlation coefficient is applied to search for the interactions of the linguistic components.

2. Methodology

2.1 Dataset

25 treebanks were adopted, which involves: a). four Uralic languages: Estonian (two treebanks), Finnish (two treebanks), North Sami (one treebank), Hungarian (one treebank); b). four Altaic languages: Buryat (one treebank), Kazakh (one treebank), Turkish (two treebanks), Uyghur (one treebank); c). two Dravidian languages: Tamil (two treebanks), Telugu (one treebank); d). one Indo-Aryan language Marathi (one treebank); e). one Afroasiatic Egyptian language Coptic (one treebank); f). one Niger-Congo Atlantic language Wolof (one treebank); g). one Basque language (one treebank); h). one Austronesian language: Indonesian (three tree banks); i). Japanese (three treebanks). Table 1 provides the details on the treebanks.

Table 1. Dataset

Treebanks	Text types	Words	Sentences	Treebanks	Text types	Words	Sentences
Basque-BDT	News	121,443	8993	**Marathi-UFAL**	Wiki, fiction	3,847	466
Buryat-BDT	Fiction, grammar-examples, news	10,185	927	**Indonesian-PUD**	News, wiki	19,446	1,000
Japanese-BCCWJ	Fiction, news, blog, conference, nonfiction	1,253,903	57,109	**Indonesian-GSD**	Blog, news	122,019	5,598
Japanese-GSDLUW	blog, news	150,243	8,100	**Wolof-WTB**	bible, wiki	44,258	2,107
Japanese-PUD	news, wiki	28,788	1,000	**Uyghur_UDT**	Fiction	40236	3,456
Tamil-TTB	News	9,581	600	**Indonesian-CSUI**	News, nonfiction	28,263	1,030
Tamil-MWTT	News	2,584	534	**Wolof-WTB**	Bible, wiki	44,258	2,107
Telugu-MTG	Grammar examples	6,465	1,328	**Coptic-Scriptorium**	Bible, fiction, nonfiction	55,858	2,163
Buryat-BDT	Grammar examples, news, fiction	10,185	927	**Estonian-EDT**	Fiction, academic, news, nonfiction	438,245	30,968
Kazakh-KTB	News, fiction, wiki	10,536	1,078	**Finnish-TDT**	Fiction, legal, news, blog, grammar-examples,	202,453	15,136
Turkish-Kenet	News, nonfiction	183,555	16,396	**Finnish-TDT**	Poetry, medical, social, web	19,382	2,122
Turkish-Boun	News, nonfiction	125,212	9,761	**North Sami-Giella**	News, nonfiction	26,845	3,122
Hungarian-Szeged	News	42,032	1,800				

2.2 Metrics

Morphological richness is measured via two metrics: moving-average morphological richness and moving-average mean size of paradigm. Cosine similarity and entropy were used to capture word-order flexibility.

2.2.1 Morphological richness

To minimise the influence of the corpus size, this study measured the morphological diversity of each language text by repeatedly calculating the type-token ratio (TTR) index for a subset of the text and the average. It further employs Covington and McFall's (2010) moving average TTR (MATTR) to calculate the word forms and lemma vocabulary richness. $\mathrm{MATTR\ (W)_{word\ form}}$ was obtained using the following formula:

$$\mathrm{MATTR\ (W)_{word\ form}} = \frac{\sum_{I=1}^{N-W+1} F_i}{W\ (N - W + 1)}$$

N is the size of the text, and W is the size of a randomly chosen window. F_i is a number in the extended form for a given window size. Regarding agglutinative languages such as Japanese, Turkish, there are two types of word forms. One is the agglutination of word root (*sabishii*.$_{\text{PRESENT TENSE}}$ 'lonely' → *sabishikatta*.$_{\text{PAST TENSE}}$ 'lonely) and the other is derivation (*sabishii* 'lonely → *sabishimi* 'loneliness'). In terms of inflectional languages, word forms can be inflectional (*kind* → *kinder*) and derivational (*kind*→*kindness*). The window size of this study is 500 words. $\mathrm{MATTR\ (W)_{word\ form}}$ is the TTR for each window obtained through conjugation. The TTR for each window was obtained by a lemma using a similar formula:

$$\mathrm{MATTR\ (W)_{word\ lemma}} = \frac{\sum_{I=1}^{N-W+1} L_i}{W\ (N - W + 1)}$$

N is the size of the text, and W is the size of a randomly chosen window. L_i is the number in the extended form for a constant window size. $\mathrm{MATTR\ (W)_{word\ lemma}}$ is the TTR for each window by conjugation. MAMR refers to the differences between word form diversity and lemma diversity. It is obtained via:

$$\mathrm{MAMR\ (W)} = \mathrm{MATTR\ (W)_{word\ form}} - \mathrm{MATTR\ (W)_{word\ lemma}}$$

The higher the $\mathrm{MAMR\ (W)}$, the richer the language's morphology. Another measure for capturing lexical diversity was the mean size of paradigm (MSP, Xanthos and Gillis 2010). It is obtained via two steps. First, divide the number of different inflections by the number of lemmas as follows.

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

Obtain MAMSP as:

$$\mathrm{MAMSP} = \frac{\sum_{I=1}^{N-W+1} \frac{F_i}{Li}}{W\ (N-W+1)}$$

The higher the MAMP and MAMSP values, the richer the morphology of the language.

2.2.2 Word order freedom

There are about 19 linguistic components involving word order (Tsunoda 2009 [1991]). This study narrowed down the components to the order of S, O, V. Six possible word-order patterning were

counted: SVO, OVS, VSO, VOS, SOV, OSV. Assuming *t* is the expected value and *s* is the observed value, the COS was obtained using the following formula:

$$\text{COS (s, t)} = \frac{\sum_{i=1}^{n} s_i t_i}{\sum_{i=1}^{n} s_i^2 \cdot \sum_{i=1}^{n} t_i^2}$$

The higher the COS (s, t), the greater the degree of flexibility in the word order of the sample. ENTR was obtained via:

$$\text{ENTR} = -\sum_{i=1}^{n} t_i \times \ln t_i$$

The higher the entropy of a particular language, the higher the word order freedom. This study examined both main sentences and sub sentences that consists of S, O, V.

3. Results

3.1 The reliability of the measures

To answer the research question 1, after obtaining the four metric values from 25 treebanks, Spearman's rank correlation coefficient analysis was performed. Figure 1 (left side) shows the scatter plot of MAMR vs. MAMSP values. The relationship between the metrics fits the regression line.

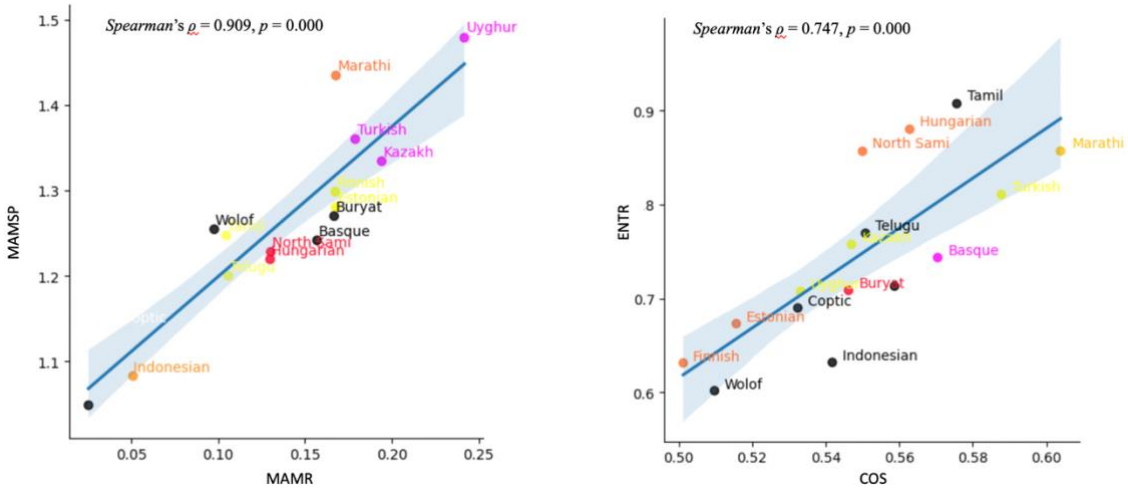

Figure 1. Scatter plot and regression line with MAMR, MAMSP; COS, ENTR

Further, Spearman's rank correlation coefficient between MAMR and MAMSP was $\rho = 0.909$ and $p = 0.000$. This indicates that the two metrics (MAMR and MAMSP) are capable of capturing morphological diversity. The Spearman's rank correlation coefficient between COS and ENTR was positive and strong, with $\rho = 0.747$ and $p = 0.000$. The regression line in Figure 3 (right side) fit the scatter plots of the correlation. This indicates COS and ENTR are capable to convey word order flexibility. Research question 1 is answered.

Table 2 provides a closer picture of values on morphological and word order diversity in each language. As indicated, the Turkic languages, i.e. Uyghur, Kazakh, Turkish present the highest value of MAMR and MAMSP, indicating that the three languages are morphologically most rich in all agglutinative languages in target. Then come the Indo-Aryan Marathi, an Indian language, belonging to the Indo-Aryan branch. Marathi is the only agglutinative language in the Indo-Aryan branch and it kept three grammatical genders: masculine, feminine and neuter. Other languages such as Urdu, Hindi are fusional and gender-less. A calculation of MAMR, MAMSP of Urdu and Hindi revealed Marathi presented the highest value of both morphology and syntax among the Indo-Aryan language. This inspired us to deduce that agglutinative language is morphologically richer and syntactically more

complex than fusional languages. The Uralic language family has seen variations, i.e. the Finnic language (Estonian and Finnish) presented higher MAMR, MAMSP value than Finno-Ugric language (North Sami and Hungarian). Dravidian languages Telugu and Tamil showed medium degree of morphological variety. The Atlantic language Wolof, Malayo-Polynesian language Indonesian, Egyptian language Coptic ranked at the end of MAMR and MAMSP values. Japanese was seen to bear the lowest degree of morphological richness.

Table 2. MAMR, MAMSP, COS and ENTR value of each language (total treebanks)

Language family	Branch	Languages	MAMR	MAMSP	COS	ENTR
Altaic	**Turkic**	**Uyghur**	**0.2418**	**1.4785**	**0.6039**	**0.8570**
Altaic	**Turkic**	**Kazakh**	**0.1942**	**1.3341**	**0.5877**	**0.8106**
Altaic	**Turkic**	**Turkish**	**0.1789**	**1.3600**	**0.5756**	**0.9074**
Indo-European	Indo-Aryan	Marathi	0.1678	1.4344	0.5704	0.7434
Uralic	**Finnic**	**Estonian**	**0.1677**	**1.2803**	**0.5512**	**0.6312**
Uralic	**Finnic**	**Finnish**	**0.1673**	**1.2985**	**0.5500**	**0.8567**
Altaic	Mongolic	Buryat	0.1668	1.2700	0.5508	0.7694
Basque	Basque	Basque	0.1569	1.2416	0.5470	0.7575
Uralic	**Finno-Ugric**	**North Sami**	**0.1303**	**1.2280**	**0.5462**	**0.7085**
Uralic	**Finno-Ugric**	**Hungarian**	**0.1299**	**1.2194**	**0.5157**	**0.6731**
Dravidian	**Dravidian**	**Telugu**	**0.1058**	**1.2000**	**0.5332**	**0.7074**
Dravidian	**Dravidian**	**Tamil**	**0.1046**	**1.2474**	**0.5312**	**0.6890**
Niger-Congo	Atlantic	Wolof	0.0977	1.2545	0.5097	0.6016
Austronesian	Malayo-Polynesian	Indonesian	0.0509	1.0829	0.5418	0.6318
Afroasiatic	Egyptian	Coptic	0.0446	1.1413	0.5324	0.6898
Japanese	Japanese	Japanese	0.0253	1.0488	0.5587	0.7130

Turning to word order flexibility, the Indo-Aryan language Marathi, Altaic Turkish were seen most flexible in word order, with ENTR ranging from 0.7434 to 0.9074. Afroasiatic Coptic, Austronesian Indonesian, Niger-Congo Wolof, Dravidian languages Tamil and Telugu were seen most rigid, cf. ENTR ranging from 0.6016 to 0.7074. Table 3 presents the distribution of word order patterns. The Turkic language Turkish and Uyghur, Basque, Finno-Ugric North Sami showed the most balanced proportions of S, V, O combinations.

Table 3. The distribution of word order patterns in agglutinative languages

Language	SVO	SOV	VSO	VOS	OVS	OSV
Basque	58.84%	30.39%	0.34%	1.70%	3.97%	4.76%
Buryat	68.95%	28.77%	0.23%	0.00%	0.00%	2.05%
Coptic	72.58%	24.60%	0.40%	0.00%	0.00%	2.42%
Estonian	73.29%	20.43%	0.00%	0.08%	3.30%	2.90%
Finnish	79.71%	16.48%	0.16%	0.00%	1.11%	2.54%
Hungarian	68.69%	22.43%	0.47%	0.00%	2.80%	5.61%
Indonesian	68.53%	31.33%	0.14%	0.00%	0.00%	0.00%
Japanese	64.60%	34.13%	0.48%	0.00%	0.00%	0.79%
Kazakh	69.97%	25.74%	0.99%	0.00%	0.00%	3.30%

Korean	87.17%	12.57%	0.00%	0.00%	0.00%	0.26%
Marathi	52.63%	42.11%	0.00%	0.00%	0.00%	5.26%
North Sami	80.75%	16.07%	0.79%	0.20%	0.99%	1.19%
Tamil	67.24%	18.97%	0.00%	0.00%	1.72%	12.07%
Telugu	67.61%	26.15%	0.15%	0.00%	0.00%	6.09%
Turkish	53.89%	43.80%	0.86%	0.86%	0.29%	0.29%
Uyghur	72.32%	25.11%	0.86%	0.21%	0.43%	1.07%
Wolof	77.34%	20.70%	0.00%	0.00%	0.00%	1.95%

3.2 Interactions between word form diversity, word order flexibility

The previous section has confirmed the validity of the four metrics for capturing morphological and syntactic complexity, i.e. MAMR, MAMSP, COS and ENTR. This section proceeds to pursue the second research question, i.e. whether morphologically richer languages are likely to be syntactically more free? The Spearman's rank correlation coefficient analysis was carried out between the morphological metrics vs. word order metrics. A positive and fairly strong correlation was confirmed, i.e. MAMR vs. COS: $\rho = 0.735$ and $p = 0.001$; MAMR vs. ENTR: $\rho = 0.656$ and $p = 0.006$; MAMSP vs. COS: $\rho = 0.624$ and $p = 0.010$; MAMSP vs. ENTR: $\rho = 0.565$ and $p = 0.023$. we plotted the scatterplot matrix of the four metrics in Figure 2.

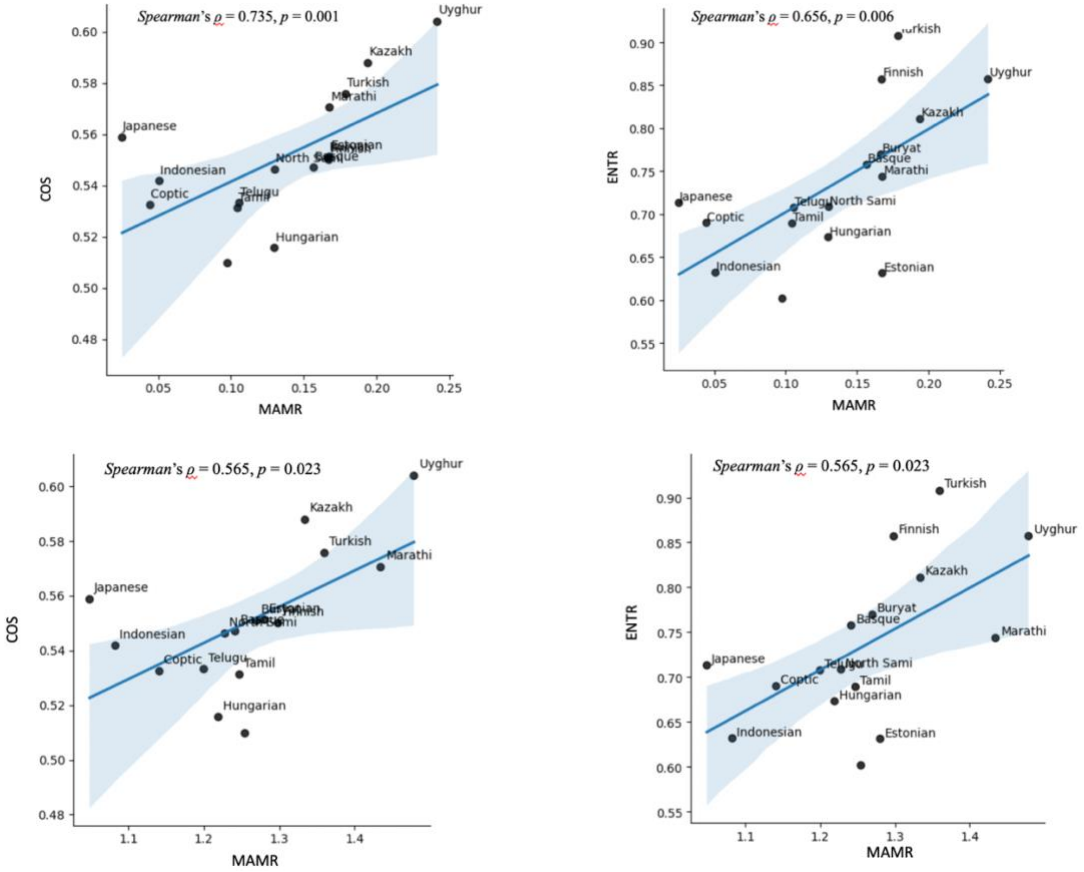

Figure 2. Spearman's rank correlation coefficient between MAMR, MAMSP, COS, ENTR

Figure 2 indicates a positive correlation between the four values, i.e. when the morphological richness of languages increases, the word order flexibility ascends. This finding is consistent with the

'complexity trade-off' hypothesis, as in Slavic languages (Yan and Liu 2021), written Japanese (Li et al. 2022).

Figure 3 and 4 showed the clustering of the agglutinative languages of nine language families based on morphological and syntactic diversity. A separation of Finnic, and Finno-Ugric branches of Uralic language family; Mongolic and Turkic branches of Altaic language family is seen. Specially, the Turkic branch is morphologically richer and syntactically more flexible than the Mongolic branch. Japanese is morphologically less rich compared with the rest 15 agglutinative languages but its word order freedom are in medium degree.

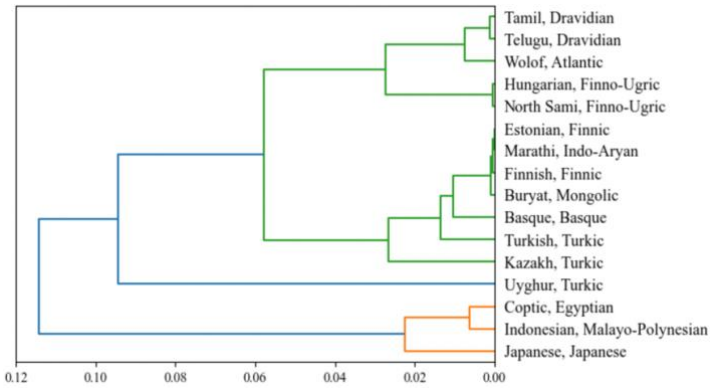

Figure 3. A clustering of the agglutinative languages based on morphological richness

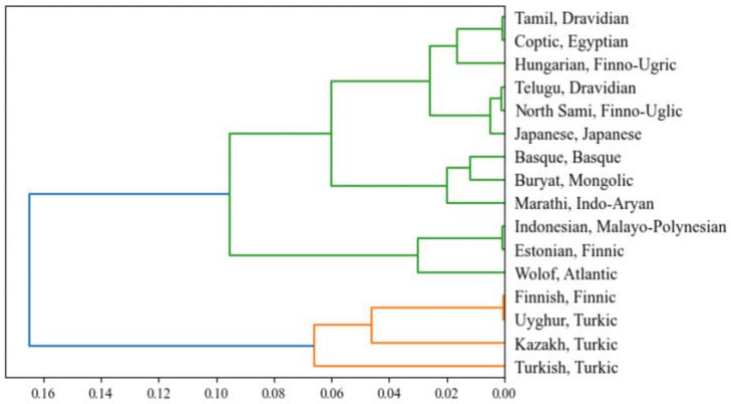

Figure 4. A clustering of the agglutinative languages based on word order flexibility

4. Summary

This study explored the correlation between morphological richness and word order flexibility by utilizing 25 treebanks of 16 agglutinative languages across nine language families. The metrics, MAMR and MAMSP; COS and ENTR were found to be valid to capture the morphological and syntactic diversity. Statistical analysis revealed that the richer the morphology, the more flexible the word order. A fairly strong correlation between morphology and syntax was confirmed in all agglutinative languages, which supports the complexity trade-off hypothesis. Among the nine language families, Austronesian, Afroasiatic, Dravidian languages were found morphologically and syntactically the least diverse. The languages of high morphological and syntactic diverse go to the Altaic Turkic branch, the In-European Indo-Aryan branch, Uralic Finnic branch. Turkish, Uyghur, Basque, and North Sami showed the most balanced proportions of S, V, O combinations. A separation of Finnic, and Finno-Ugric branch of Uralic language family; Mongolic and Turkic branch of Altaic language family was confirmed. The Finnic was found more flexible than Finno-Ugric; the Mongolic was found more rigid than the Turkic.

Acknowledgement

This paper is based on work that was supported by the National Foundation of Social Sciences of China (22BYY186).

References

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

Čech, R., Kubát, M. 2018. Morphological Richness of Text. In: Fidler, M., Cvrček, V. (eds.) Taming the Corpus. *From Inflection and Lexis to Interpretation. Quantitative Methods in the Humanities and Social Sciences*. Cham: Springer, 63-77.

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

Coloma, G. 2017. The Existence of Negative Correlation between Linguistic Measures across Languages. *Corpus Linguistics and Linguistic Theory* 13: 1-26.

Fenk-Oczlon, G., and Fenk, A. 1999. Cognition, quantitative linguistics, and systemic typology. *Linguist. Typol.* 3, 151–177. doi: 10.1515/lity.1999.3.2.151

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

Fenk-Oczlon G., Pilz J. 2021. Linguistic Complexity: Relationships Between Phoneme Inventory Size, Syllable Complexity, Word and Clause Length, and Population Size. *Front. Commun*. 6:626032. doi: 10.3389/fcomm.2021.626032

Jakobson, R. 1936. Beitrag zur allgemeinen Kasuslehre, Gesamtbedeutungen der russischen Kasus. In: *PLingCP* 6: 240-288.

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

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

Li, W., Liu, H., Xiong, Z. 2022. A quantitative analysis of word order freedom and the abundance of case markers in Japanese [Nihongo ni okeru gojun no jiyuu-do to kakuhyooshiki no hoofu-sa ni kansuru keiryoo-teki kenkyuu]. *Mathematical linguistics* 33 (5): 325-340.

McFadden, T. 2003. On morphological case and word-order freedom. *Berkeley Linguistics Society* 29: 295–306.

Menzerath, P. 1954. *Die Architektonik des Deutschen Wortschatzes*. Hannover; Stuttgart: Dümmler.

Miestamo, M. 2009. Implicational hierarchies and grammatical complexity. In G. Sampson, D. Gil and P. Trudgill (Eds.), *Language Complexity as an Evolving Variable* (pp. 81-97). Oxford: Oxford University Press

Sapir, E. 1921. *Language: An Introduction to the Study of Speech*. New York: Harcourt, Brace & World Inc., 33-35.

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

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

Sinnemäki, K. 2008. Complexity trade-offs in core argument marking. In M. Miestamo, K. Sinnemäki and F. Karlsson (Eds.), Language complexity: Typology,contact, change (pp. 67-88). Amsterdam, Philadelphia: Benjamins

Sinnemäki, K. 2014. Complexity trade-offs: A case study. In: F.J. Newmeyer and L.B. Preston (eds.),

Measuring Grammatical Complexity, 179–201. Oxford: Oxford University Press.
Xanthos, A., & Gillis, S. 2010. Quantifying the development of inflectional diversity. *First Language* 30, 175–198.
Yan, J., Liu, H. 2021. Morphology and word order in Slavic languages: Insights from annotated corpora. *Voprosy Jazykoznanija* 4: 131–159.