Aufsätze

Siqi Liu / Jianwei Yan / Haitao Liu

The complexity trade-off between morphological richness and word order freedom in Romance languages: A quantitative perspective

https://doi.org/10.1515/zrp-2025-0032

Abstract: The complexity trade-off hypothesis suggests a balance between different linguistic features across human languages. This study investigates this hypothesis by quantitatively examining the evolution from Latin to Modern Romance languages. We focus on the trade-off between morphological richness and word order freedom, providing insights into their dynamic interrelations during linguistic evolution. Our analysis demonstrates that morphological richness and word order freedom are distributed along a continuum, with Latin exhibiting higher morphological complexity and freer word order and Modern Romance languages having lower morphological richness and more rigid word order. This evolution reflects the principles of efficiency in complex adaptive systems, showing a significant complexity trade-off where increases in one feature often result in decreases in the other. These findings indicate the adaptive nature of linguistic systems and offer valuable insights for diachronic typological research, enhancing our understanding of the complexity trade-off from a quantitative perspective.

Note: The authors have no competing interests to declare that are relevant to the content of this article.

Supported by: This work was supported by the National Social Science Foundation of China (Grant No. 24CYY064) the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities in China (Grant No. 22JJD740018), the Humanities and Social Sciences Youth Foundation of the Ministry of Education of China (Grant No. 22YJC740047), and the Hangzhou Normal University (Grant No. 4075C50223204130).

Correspondence address: Jianwei Yan, Department of Linguistics, Zhejiang University, Hangzhou, China, E-Mail: yanjianwei@aliyun.com
Siqi Liu, School of International Studies, Hangzhou Normal University, Hangzhou, China, E-Mail: liusiqi1225@foxmail.com
Haitao Liu, College of Foreign Languages and Literature, Fudan University, Shanghai, China, E-Mail: htliu@163.com

Keywords: complexity trade-off, morphological richness, word order freedom, Romance language

1 Introduction

In linguistic research the complexity trade-off hypothesis, also known as the negative correlation hypothesis, holds that the various elements of human languages exhibit negative correlations in terms of complexity (Hockett 1958; Crystal 1997; Coloma 2017). This hypothesis suggests that all languages strive to balance their overall complexity in order to efficiently convey similar content. Hockett (1958, 180–181) noted that the overall complexity of the grammar (including lexicon and syntax) of any language should be roughly the same, as all languages need to express content of roughly the same complexity. Thus, what cannot be expressed morphologically must be conveyed syntactically. In other words, if the morphological structure of a language is complex, then the syntax should be simple; conversely, if the syntax is complex, then the morphology should be simple. Crystal (1997, 5–6) also pointed out that all languages possess complex grammatical systems; they may be relatively simple in one aspect (such as having no inflections), but if so they usually become relatively complex in another aspect (such as word order), meaning that languages are self-regulating systems.

Morphological richness generally refers to the complexity and variety of word forms within a language, while word order freedom refers to the flexibility with which words can be arranged to form sentences. Previous studies have explored the relationship between morphological case and word order freedom, suggesting that there might exist a linguistic tendency towards a complexity trade-off by which languages with richer morphology tend to allow more flexible word order (Sapir 1921; McFadden 2003; Yan/Liu 2021; Li, et al. 2022; Li/Liu 2024). Research has also examined the role of processing complexity in word order changes, indicating that word order is a stable aspect of a particular language but it can vary significantly between languages in terms of the freedom allowed (Tily 2010). Furthermore, the balance between effort and information transmission during language acquisition has been studied, revealing correlations between case marking and constituent order freedom (Fedzechkina, et al. 2017). These studies offer a window into how humans encode and decode linguistic information, enriching our understanding of the dynamic relations between different components of human language (Yan/Liu 2021, 132). However, few studies to date have focused on this linguistic hypothesis from a diachronic perspective (Gulordava/Merlo 2015), which will provide us with another angle from which to approach language evolution and language processing.

In the field of historical linguistics the study of language evolution and structure is complex, intertwining various linguistic aspects such as morphology and syntax. Among the most persistent topics within this field is the diachronic changes in morphosyntactic features from Latin to the Romance languages (Schwegler 1990; Ledgeway 2012; Liu/Xu 2012). The relationship between Romance languages and their common ancestor, Latin, provides an exceptional opportunity to explore how languages evolve over time (Posner 1996; Maiden 2014). The evolution of Romance languages from their shared Latin or Proto-Romance ancestry presents a fascinating case for linguistic analysis (Ledgeway 2011; Buchi/Chauveau 2015; Vincent 2016). However, while some explorations of these typological changes have been conducted (e.g., Gulordava/Merlo 2015; Haspelmath/Michaelis 2017), investigations into the overall features of morphological richness and word order freedom remain insufficient.

Specifically, Latin is a synthetic or inflected language, meaning that the endings of words can change to indicate their grammatical function in a sentence. This allows for a high degree of word order or syntax flexibility. For instance, to express the idea of "*Caesar captured the city*", the Latin language may have six different variants:

SOV: *Caesar cepit urbem (Caesar the city captured)*
SVO: *Caesar urbem cepit (Caesar captured the city)*
VSO: *Cepit Caesar urbem (Captured Caesar the city)*
VOS: *Cepit urbem Caesar (Captured the city Caesar)*
OSV: *Urbem Caesar cepit (The city Caesar captured)*
OVS: *Urbem cepit Caesar (The city captured Caesar)*

In the above sentences "Caesar" is the nominative form, indicating the subject, while "cepit" is the third person singular perfect active indicative form of "capere" (to capture). "Urbem" is the accusative form of "urbs" (city), indicating the object. Although SOV and SVO are the most commonly used word orders in Latin, the meanings of all these six variants remain clear. The endings of the words indicate their grammatical roles (nominative for the subject, accusative for the object), and the relationship between the words is maintained regardless of word order. This means that the grammatical roles in Latin are clearly marked by the inflections, the words in the sentence can be rearranged in various orders, and the sentence will still be understood, showcasing the syntactic flexibility and rich morphology of Latin.

When we examine Modern Romance languages, however, we find that they tend to be less morphologically marked and more syntactically fixed. Despite the wealth of qualitative research on these languages, there remains a gap in understanding the historical trends and precise mechanisms that govern the relations between morphological richness and word order freedom in the evolution of the Romance language. This gap hinders a comprehensive understanding of language evolution and typology (Dahl 2004). In other words, there is still a noteworthy lack of quantitative and empirical investigation of the overall evolution of the features of morphological richness and word order freedom in Modern Romance languages.

Quantitative linguistics, a subdiscipline of general linguistics, explores language features through mathematical approaches, offering a lens for the quantitative investigation of morphosyntactic dynamics in this research. It aims to quantify, simulate, model, and explain linguistic phenomena and their ever-evolving nature, shedding light on the self-adaptive mechanisms and dynamic processes that drive the evolution of language over time (Köhler, et al. 2005; Köhler/Altmann 2008; Best/Rottmann 2017; Yan 2024). In this study, we utilize a quantitative linguistic approach to examine the complexity trade-off between morphological richness and word order freedom in Romance languages. To achieve this, suitable linguistic materials and quantitative methods must be identified. In the 1960s Greenberg proposed metrics to assess synthesis levels in languages by examining morpheme counts per word and implicational universals related to word order (Greenberg 1960; 1963). Despite these foundational efforts, debates persist regarding the best linguistic materials and evaluation methods for examining morphological richness and word order flexibility, reflecting the diverse morphological and syntactic properties of languages (Dahl 2004; Bickel/Nichols 2007; Haspelmath/Michaelis 2017).

The current study adopts the emerging cross-lingually consistent annotated corpora of Universal Dependencies (UD) as the linguistic materials, and the refined metrics of moving-average mean size of paradigm (MAMSP) and word order cosine similarity (COSS). The rationale for adopting these linguistic materials and quantitative metrics is twofold. First, UD is a database encompassing over 100 languages, designed to maximize cross-linguistic parallelism by consistently annotating different languages (Nivre 2015). A key advantage of this database for typological studies is its consistent and specific annotations, enabling easy comparison and interpretation across languages (Gerdes, et al. 2021; Yan/Liu 2023) (cf. Section 2.1 Materials for more details). Second, MAMSP and COSS are well-suited to the UD data, which provides information on word forms and word lemmas for each word, and syntactic relations for each sentence. MAMSP quantifies the average number of unique inflected forms per lemma, applying a moving-average operation to mitigate the impact of corpus size (Yan/Liu 2021; Li, et al. 2022; Li/Liu 2024), while COSS measures the similarity between actual word order variants and an ideal balanced word order distribution, namely the flexibility of word orders (Kuboň, et al. 2016). This approach emphasizes using quantitative, computable, and interpretable metrics to understand the evolution of morphosyntactic features across languages (cf. Section 2.2 Methods for more details). By utilizing large-scale corpora of UD with morphological and syntactic annotations and established metrics of MAMSP and COSS, this research explores how the morphology and syntax of Romance languages have changed compared to Latin, thereby capturing the patterns and laws beneath the diachronic changes in morphological richness and word order freedom in Romance languages. The primary research questions are as follows:

> ***Question 1:*** *How has the morphological richness of Romance languages evolved over time?*
> ***Question 2:*** *How has the word order freedom of Romance languages changed over time?*
> ***Question 3:*** *What are the correlations between morphological richness and word order freedom in Romance languages?*

This study aims to contribute to the broader understanding of language change and typology, offering insights into the cognitive and communicative factors that shape language structure. It will also provide valuable data for computational models of language processing.

The paper will be structured as follows: **Section 1** introduces the study and provides a literature review of relevant research. **Section 2** details the linguistic materials and the corpus-based approach used. **Section 3** presents the findings and interprets the implications of the results. Finally, **Section 4** summarizes the study and suggests directions for future research.

2 Materials and methods

2.1 Materials

In the current research we implemented a corpus-based approach, concentrating on Latin and seven Romance languages to comprise a total of eight languages across 23 corpora derived from UD version 2.5 (Zeman, et al. 2019).[1] Additionally, we incorporated four Chinese corpora covering both Modern (three corpora) and Classical Chinese (one corpus),[2] to serve as benchmarks for analysis. Consequently, our study encompasses 27 corpora across nine languages, as outlined in **Table 1**. More details on the corpora used can be found in **Appendix I**.

[1] Cf. <https://universaldependencies.org/>.

[2] In UD 2.5 there are six corpora for Chinese. Due to the lack of LEMMA annotations, we have excluded the Chinese-HK and Chinese-PUD corpora from our analysis.

Table 1: Basic information on the 27 corpora

#	Categories	Languages	Number of corpora	Dominant word order in WALS
1	**Research Focus**	Latin	3	No matching records found
2		Catalan	1	SVO
3		French	5	SVO
4		Galician	2	No matching records found
5		Italian	6	SVO
6		Portuguese	1	SVO
7		Romanian	3	SVO
8		Spanish	2	SVO
9	**Baseline**	Chinese	4	SVO[3]
		Grand Total	**27**	

As shown in **Table 1**, the ancestor of Romance languages, Latin, is listed in the first row, and the following seven Romance languages belong to the sub-branches of Ibero-Romance (Spanish, Portuguese, Galician, Catalan), Gallo-Romance (French), Italo-Romance (Italian), and Eastern Romance (Romanian). The Chinese language is considered typical analytic (Norman 1988; Li/Thompson 1989) contrary to the synthetic nature of Romance languages.

All UD corpora used are annotated in the CoNLL-U format, which follows the principles of dependency grammar (Tesnière 1959; Mel'čuk 1988). This format effectively captures the linguistic relationships between sentence elements, especially heads and dependents (Heringer 1993; Hudson 1995; Jiang/Liu 2018), supporting detailed linguistic analyses across diverse languages and language families. **Figure 1** is a tree representation of the sample sentence, *Andy drinks a few beers, and I drink some coffee*, based on the CoNLL-U format. It includes *Dependency Relations* between the heads and dependents, *XPOS Tags* (language-specific part-of-speech tags), *Word Forms*, *Word Lemmas*, and *UPOS Tags*.[4]

[3] In WALS (cf. <https://wals.info/>), Chinese is regarded as Mandarin.

[4] Cf. <https://universaldependencies.org/format.html>.

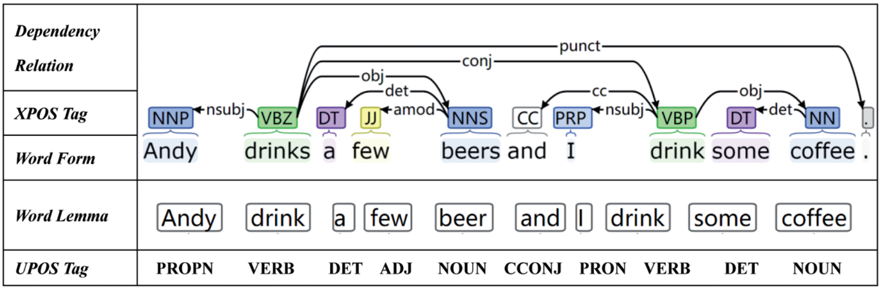

Dependency Relation
XPOS Tag
Word Form
Word Lemma
UPOS Tag
punct
conj
obj
det
nsubj
amod
cc
nsubj
obj
det
NNP
VBZ
DT
JJ
NNS
CC
PRP
VBP
DT
NN
.
Andy drinks a few beers and I drink some coffee .
Andy drink a few beer and I drink some coffee
PROPN VERB DET ADJ NOUN CCONJ PRON VERB DET NOUN

Figure 1: Tree representation of the sample sentence *Andy drinks a few beers, and I drink some coffee*[5]

Based on the *Word Forms* and *Word Lemmas* in **Figure 1**, we can see that the **VERB** *drinks* and the **NOUN** *beers* are morphologically marked, then we can quantify the morphological richness of the language under discussion and compare it with that of other languages. In addition, based on the *Dependency Relation* of ***nsubj*** between *Andy* and *drinks*, that of ***obj*** between *drinks* and *beers*, and the *UPOS Tag* of ***VERB*** (*drinks*), we know that the word order of the first main clause is SVO. Also, based on the *Dependency Relation* of ***nsubj*** between *I* and *drink*, that of ***obj*** between *drink* and *coffee*, and the *UPOS Tag* of ***VERB*** (*drink*), we know that the word order of the second main clause is SVO as well.

2.2 Methods

The metrics of moving-average mean size of paradigm (MAMSP) and cosine similarity of word order (COSS) were adopted as indicators of morphological richness and word order freedom, respectively.

2.2.1 Morphological richness

We employed the MAMSP metric to assess morphological richness. To understand how MAMSP is calculated, it is crucial to first grasp the concept of Mean Size of Paradigm (MSP), which was initially proposed by Xanthos and Gillis (2010). MSP quantifies morphological richness by calculating the average number of unique inflected forms per lemma. The formula for this calculation is as follows:

[5] Cf. <https://corenlp.run/>.

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

In this formula, F represents the number of distinct inflected word forms and L the number of distinct lemmas, either within a specific sentence or across an entire corpus. For instance, the total number of distinct word forms in the sample sentence *Andy drinks a few beers, and I drink some coffee.* is 10, including *Andy*, *drinks*, *a*, *few*, *beers*, *and*, *I*, *drink*, *some*, and *coffee* (F = 10). The number of distinct lemmas, however, is 9, counting *Andy*, *drink*, *a*, *few*, *beer*, *and*, *I*, *some*, and *coffee* (L = 9). Therefore, the MSP of the sentence is 10/9 ≈ 1.11.

Following Yan and Liu (2021), we refined the MSP algorithm by introducing the moving-average MSP (MAMSP). This adaptation utilizes a moving-average operation to reduce the impact of corpus size on the MSP results. Also, being different from the original methodology of Xanthos and Gillis (2010), which focused solely on verbs and nouns, our approach incorporates all *UPOS Tags* in calculating MAMSP. This extension is essential for providing a more accurate representation of inflectional languages with sophisticated case systems. The specific formula for MAMSP is as follows:

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

In this formula, N signifies the total number of tokens in a corpus while W denotes the window size, with the condition that $W < N$. Within each window, F_i represents the number of distinct word forms and L_i the number of distinct lemmas. For the practical application of this modified metric, we adopted a standard window size of 500 (W = 500), following Covington and McFall (2010) in their calculation of moving-average type-token ratio. Our calculation was based on the *Word Form* and *Word Lemma* information extracted from the UD corpora, and was performed using a self-written R script. A higher MAMSP value indicates a greater morphological complexity in the language under analysis. This approach allows us to analyze and compare morphological complexity across different languages.

2.2.2 Word order freedom

Since the word order of the core arguments and the verb (S/V/O) of sentences has always been the most fundamental element of word order (Dryer 2005), we will focus on the flexibility of the word orders of the languages under investigation. Specifically, we extracted word order information from the UD corpora. One noteworthy point is that to minimize the impact of varying proportions of non-declarative sentences (i.e., imperative, interrogative, and exclamative) across the 27 corpora, we first excluded all non-declarative sentences using a self-written Perl script. Another important thing to note is that when it comes to the word order of S/V/O, some studies mixed main clauses and subordinated clauses altogether, while others focused on main clauses only (Greenberg 1963; Tomlin 1986; Dryer 1992; Courtin 2018; Yan/Liu 2021). Hence, in this study, we investigate the word order freedom in both situations to ensure the reliability of the measures used in the current research.

Technically, following the methodologies of Bonfante, et al. (2018) and Courtin (2018, 36–40), we calculated the relative frequencies or probabilities of the six S/V/O variants of the two situations in these corpora, utilizing specific extraction patterns based on the *UPOS Tags* and *Dependency Relations* as shown in **Figure 1**.

Specifically, the extraction pattern for the SVO of mixing main and subordinate clauses and the extraction pattern for the SVO of extracting main clauses are provided below.

SVO Pattern (Situation 1: both main and subordinate clauses):[6]

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

SVO Pattern (Situation 2: main clauses only):

```
pattern { V [upos=VERB]; V -[nsubj]-> S; V -[obj]-> O; S << V; V << O; S << O }
```

The other five extraction patterns (SOV, VSO, VOS, OVS, OSV) for each situation follow a similar pattern. Using the probabilities of all six variants in two different situations, we calculated the values of COSS, a metric that measures the similarity between two vectors and is widely utilized in information retrieval (Muflikhah/Baharudin 2009; Li/Han 2013; Yan/Liu 2021). In this research, COSS is employed to measure the distances between the actual probability of each S/V/O variant and the “ideal vector”, which represents an equal frequency distribution of all six variants (i.e., 100% / 6 ≈ 0.1667) (Kuboň, et al. 2016). The algorithm used is as follows:

$$COSS = \frac{\sum_{i=1}^{n} P(x_i) \, * \, P(y_1)}{\sqrt{\sum_{i=1}^{n} P(x_i)^2} \, * \, \sqrt{\sum_{i=1}^{n} P(y_i)^2}}$$

In this formula the symbol $P(x_i)$ represents the probability or the relative frequency of each word order variant in a given language, while $P(y_i)$ refers to the “ideal vector” (0.1667), which assumes an equal distribution of frequencies. The values of

[6] The extraction patterns are supported by the Grew software. For more information on Grew, cf. <-https://grew.fr/>.

COSS increase with the growth of word order freedom (Kuboň, et al. 2016). Therefore, the larger the COSS, the greater the freedom of word order.

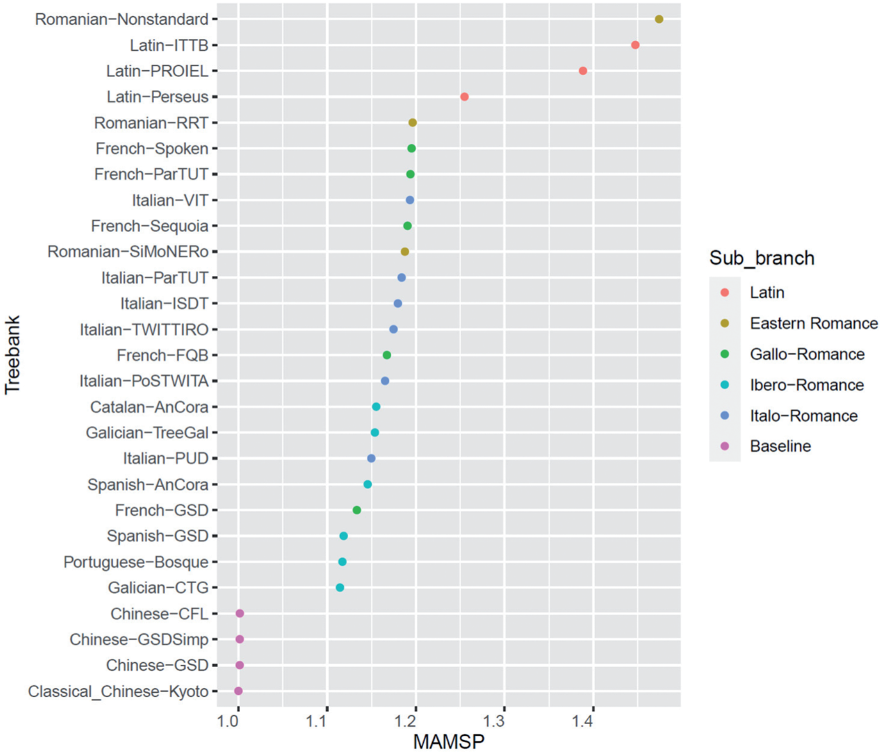
To summarize, we adopted a corpus-based approach in the current research due to its ability to facilitate large-scale, empirical analysis of linguistic data. The morphological richness and word order freedom of the languages under investigation, as well as their interrelations, were measured and quantified using computable and comprehensive measures of MAMSP and COSS derived from annotated corpora.

3 Results and discussion

Section 3.1 explores the diachronic changes in the morphological richness of Romance languages using the MAMSP metrics across all the corpora studied. **Section 3.2** focuses on the diachronic changes in word order freedom in Romance languages. In this part, we first examine the correlation between word order freedom based on the COSS metrics of two situations (for both main and subordinate clauses and for main clauses only). Finally, **Section 3.3** evaluates the trade-off between morphological richness and word order freedom, investigating whether they follow the complexity trade-off hypothesis and whether these metrics can assess the evolution and closeness of Romance languages.

3.1 The diachronic change of morphological richness in Romance languages

To evaluate the diachronic changes in morphological richness in Romance languages we calculated the MAMSP values, which serve as indices of morphological complexity, for the 27 corpora across the nine different languages, and ranked MAMSP in descending order as shown in **Figure 2**. The detailed MAMSP values are presented in **Appendix II**.



* The values are ranked in descending order according to the MAMSP values.

Figure 2: Morphological richness of all 27 corpora.

As previously mentioned, languages with higher MAMSP values demonstrate greater morphological complexity. **Figure 2** shows that Chinese languages have lower MAMSP values compared to Romance languages. This result indicates that typical analytical languages, in our case both Modern and Classical Chinese, are at the opposite end of the morphological spectrum from Romance languages, with the lowest MAMSP values in **Figure 2**. Conversely, the typical synthetic Indo-European Romance languages, especially the highly synthetic Latin language (all three Latin corpora with relatively high MAMSP), occupy the other end of the spectrum. This pattern supports our initial hypothesis that these languages are sequentially distributed along a continuum of morphological richness.

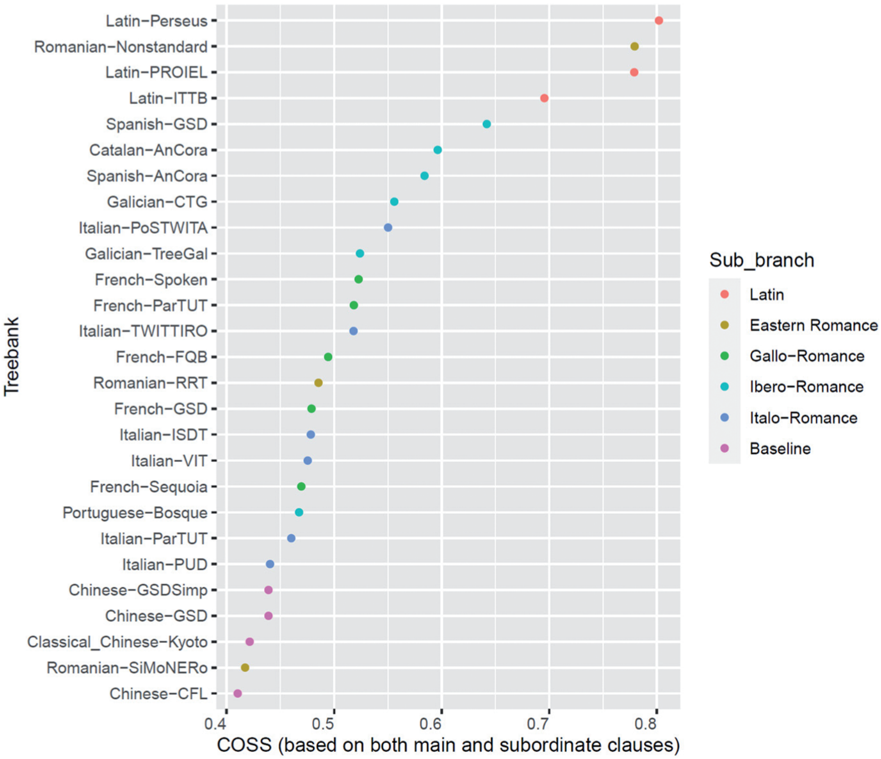
Another interesting feature of this morphological continuum is the exceptionally high MAMSP values observed in the Romanian-Nonstandard corpus. This peculiarity can be attributed to the inclusion in the corpus of various nonstandard language genres such as Old Romanian, Chat, and Folklore (Mărănduc/Malahov, et al. 2016; Mărănduc/Perez, et al. 2016; Mărănduc/Bobicev 2017). These nonstandard varieties of Romanian include various dialects and regional forms of the language and preserve historical linguistic features, retaining older forms and more complex morphological structures that might have been simplified or lost in the standard language. Also, the Romanian language has been influenced by multiple languages throughout its history, especially Slavic languages (Hitchins 1996; Schaller 2023). Due to geographical proximity and political and cultural exchanges, the geographical area of Romania had frequent contact with Slavic countries (a well-known language family for its rich morphological markings). This contact has led to the Romanian language borrowing many Slavic words, especially words related to religion, everyday life, agriculture, and social structure. The connection between Romanian and Slavic languages also explains why the other two Romanian corpora, i.e., Romanian-RRT and Romanian-SiMoNERo, also have high MAMSP.

However, the classifications of subbranches within other corpora of Modern Romance languages are less distinct, as they exhibit similar MAMSP values ranging from 1.11 to 1.19. In **Section 3.3**, we will further investigate the clustering of Romance languages to determine whether Modern Romance languages can be classified into corresponding subbranches based on both morphological richness and word order freedom metrics.

In summary, the quantitative analysis reveals significant diachronic changes in morphology from Latin to Modern Romance languages. Latin has a highly inflectional system with extensive case marking on nouns and verbs conjugated for person, number, tense, mood, etc., whereas the Romance languages have reduced or eliminated some of these inflectional categories. This reduction is evident through the decreased form counts for lemmas compared to Latin.

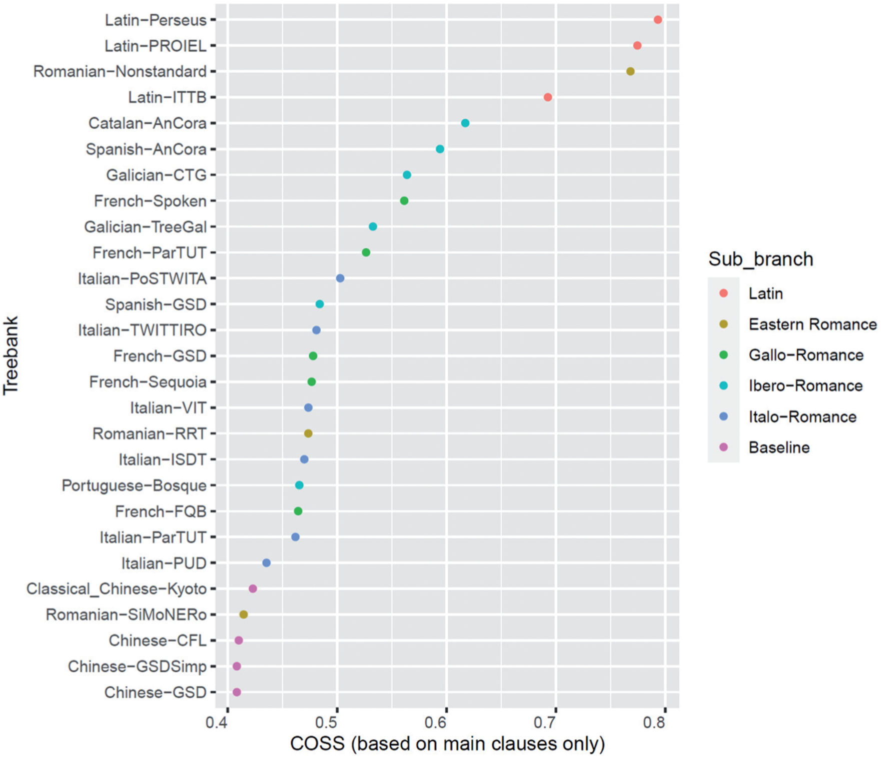
3.2 The diachronic change of word order freedom in Romance languages

We then examined the diachronic changes in word order freedom in Romance languages using the COSS metrics across all the corpora under investigation. Prior to this analysis we calculated the COSS values based on both main and subordinate clauses and on main clauses alone, to ensure the reliability of the metrics in measuring the word order freedom of the languages studied. The results are presented in **Figure 3** and **Figure 4**, respectively. The detailed COSS values for these two situations are presented in **Appendix III** and **Appendix IV**.



* The values are ranked in descending order according to the COSS values.

Figure 3: Word order freedom of all 27 corpora based on both main and subordinate clauses.



* The values are ranked in descending order according to the COSS values.

Figure 4: Word order freedom of all 27 corpora based on main clauses only.

As previously mentioned in **Section 2**, languages with higher COSS values exhibit greater word order flexibility. Both **Figure 3** and **Figure 4** demonstrate that the COSS values for Chinese are generally lower (ranging from 0.41 to 0.43 and from 0.41 to 0.42, respectively) than those for Romance languages, indicating that analytical languages have less word order flexibility than Romance languages. Additionally, Latin generally has the highest COSS values (ranging from 0.70 to 0.80 and from 0.69 to 0.79, respectively) among the languages studied, compared to other Modern Romance languages.

One noteworthy point is that the Romanian-Nonstandard and Romanian-SiMoNERo corpora exhibit distinct peculiarities. Romanian-Nonstandard demonstrates relatively high word order freedom, whereas Romanian-SiMoNERo shows comparatively low word order freedom. This may be attributed to two reasons. For one, the Romanian-Nonstandard corpus is composed of various nonstandard Romanian genres, as mentioned in **Section 3.1**, which means the corpus is likely to keep the flexible word order featured by dialects and historical forms. For another, the Romanian-SiMoNERo corpus is a contemporary medical corpus composed of three medical subdomains: cardiology, diabetes, and endocrinology (Mitrofan, et al. 2019; Barbu Mititelu/Mitrofan 2020). As suggested by previous research (Yan/Liu 2021), formal genres tend to exhibit more rigid word order. This tendency might explain why the COSS of the Romanian-SiMoNERo corpus is relatively low.

To further validate the reliability and precision of COSS as a measure of syntactic complexity, we created a scatterplot with a regression line to illustrate the relationship between COSS values based on both main and subordinate clauses and those based on main clauses only, as shown in **Figure 5**.

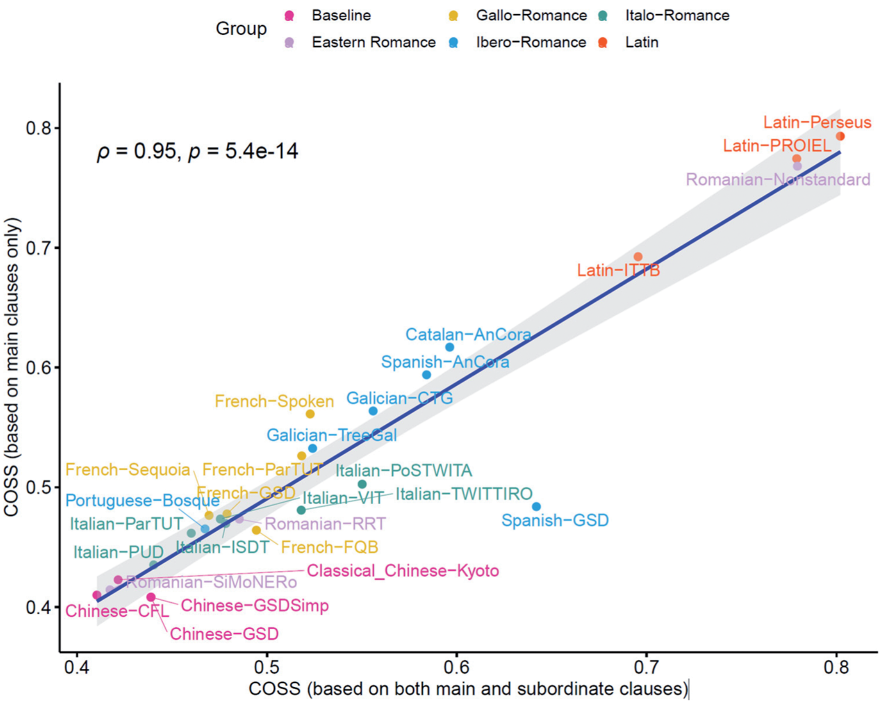


Figure 5: Word order freedom of all 27 corpora based on both main and subordinate clauses, and those based on main clauses only.

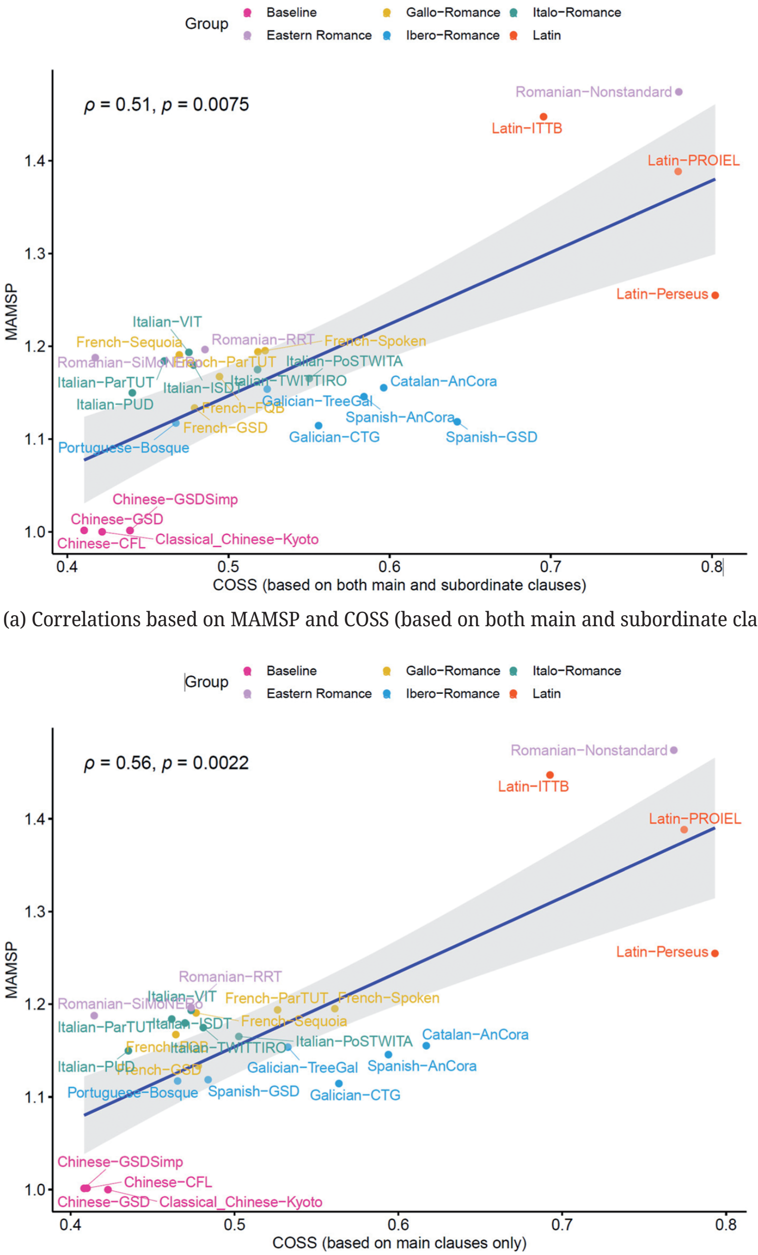
Figure 5 demonstrates that the regression line effectively captures the correlation between COSS values based on both main and subordinate clauses and COSS values based on main clauses only, with a positive, strong, and statistically significant Spearman's rank correlation coefficient (ρ = 0.95, p = 5.4e-14 < 0.001). This result suggests that the measures of word order freedom, whether for both main and subordinate clauses or based on main clauses alone, are consistent and reliable in reflecting the degree of word order flexibility in the languages under investigation. Furthermore, the visualization in **Figure 5** confirms that the Proto-Romance ancestry, Latin, exhibits greater syntactic flexibility compared to modern ones and significantly more than our baseline, Chinese.

In a difference from previous studies (e.g., Yan/Liu 2021) that relied on the COSS of both main clauses and subordinate ones, the present research sought to measure word order flexibility in main clauses only, as well to examine the word order freedom in two dimensions. Our result of a significant correlation further demonstrates the effectiveness of cosine similarity measures in quantifying the word order flexibility of the languages under investigation.

To summarize, this section has validated the reliability of the metrics in measuring the word order freedom of the languages, and presented our findings related to diachronic changes in the word order freedom of Romance languages. Specifically, our quantitative methods provide evidence for the significant syntactic changes that have occurred during the evolution from Latin to Modern Romance languages. While Latin is characterized by relatively free word order due to its rich case system allowing for flexible constituent placement within sentences, most Modern Romance languages display more rigid SVO order in both main and subordinate clauses.

3.3 The trade-off relationship between morphological complexity and syntactic flexibility in Romance languages

The previous sections have demonstrated that the morphological and syntactic metrics we employed are computationally viable, easy to interpret, and effective in capturing the complex structures of human languages. Building on our analysis and validation of these measures of morphological richness and word order freedom, we subsequently examined the trade-off relationships between these two components in Romance languages. **Figure 6** illustrates the evolutionary dynamics of Romance languages as they developed from Latin.



(a) Correlations based on MAMSP and COSS (based on both main and subordinate clauses).

(b) Correlations based on MAMSP and COSS (based on main clauses only).

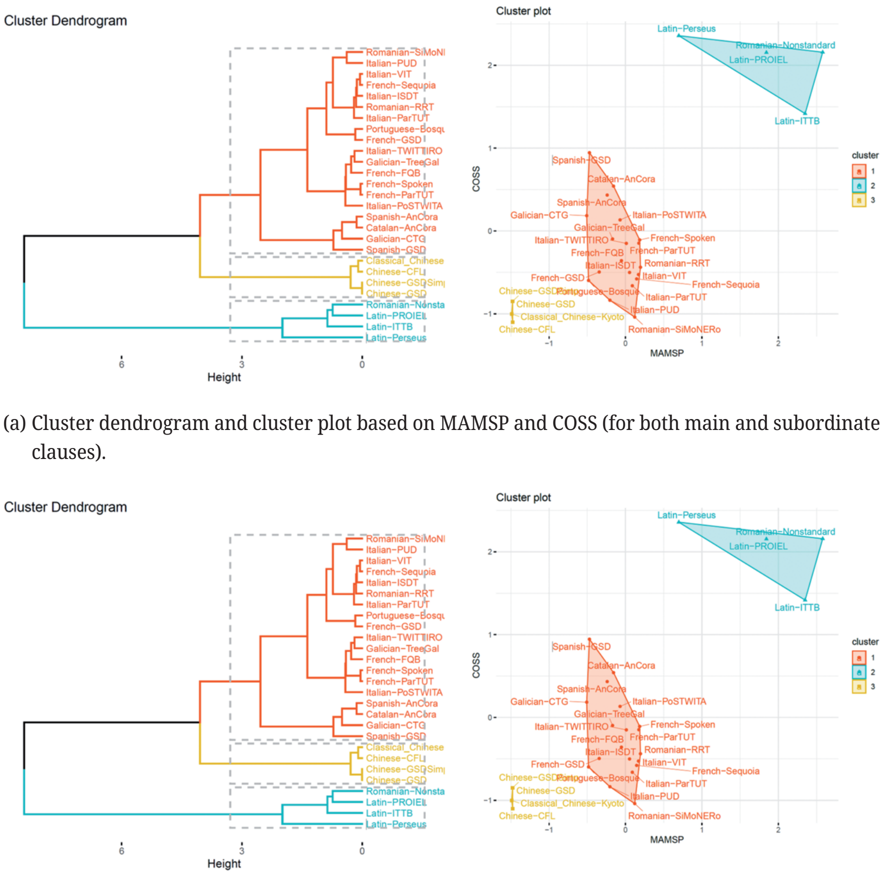
Figure 6: Correlations based on morphological richness and word order freedom (based on both main and subordinate clauses and main clauses only, respectively) of all 27 corpora.

Figure 6 reveals a trade-off correlation between morphological complexity and syntactic flexibility. The Spearman's rank correlation coefficient between MAMSP and COSS (considering both main and subordinate clauses, **Figure 6 (a)**) is moderate and statistically significant ($\rho = 0.51$, $p = 0.0075 < 0.01$). Similarly, the Spearman's rank correlation coefficient between MAMSP and COSS (considering main clauses only, **Figure 6 (b)**) is also moderate and statistically significant ($\rho = 0.56$, $p = 0.0022 < 0.01$).

These findings indicate that as Latin evolved into the Romance languages its rich system of morphological markings was progressively simplified, while word order became increasingly rigid. In other words, the diachronic analysis provides evidence for a compensatory relationship between morphological and syntactic change during the transition from synthetic Latin to the analytic Romance languages. This suggests that a decrease in morphological complexity correlates with less word order flexibility, supporting the complexity trade-off hypothesis. Consequently, these statistical findings provide empirical support for qualitative assumptions regarding the interplay between morphology and syntax in linguistics (Sapir 1921; McFadden 2003).

An additional consideration is the effectiveness of the morphological and syntactic metrics used in this study to categorize Romance languages into distinct subgroups. By adopting a morphological richness metric and two word-order freedom metrics, we conducted an Agglomerative Clustering Analysis using Euclidean distance as the similarity measure. The resulting cluster dendrograms and cluster plots, based on MAMSP and COSS (considering both main and subordinate clauses) and MAMSP and COSS (considering main clauses only), are presented in **Figure** 7. The correlation coefficients of 0.935 and 0.932 respectively indicate that the cluster trees effectively represent the relationships between the languages.[7]

[7] While the high correlations indicate that the clustering process accurately represents the original distances of the morphosyntactic features among the language corpora, they do not automatically mean that languages from the same subbranches are grouped together. That's why we further investigated the clustering output, including the dendrogram and specific clusters, to confirm how well the clustering aligns with known subbranches. For instance, in **Figure 7(a)**, Romanian-SiMoNERo and Italian-PUD are grouped into a small subset within the larger red subset representing Modern Romance languages. This does not imply that certain varieties of Romanian are more closely related to Italian than to other Romanian varieties; rather, it indicates that these two corpora exhibit more similarities to each other based on the morphosyntactic features and clustering algorithm used in this study. More importantly, both still fall under the broader subset of Modern Romance languages, meaning they exhibit greater similarity to other corpora within this red subset compared to those in the yellow and green subsets.



(a) Cluster dendrogram and cluster plot based on MAMSP and COSS (for both main and subordinate clauses).

(b) Cluster dendrogram and cluster plot based on MAMSP and COSS (for main clauses only).

Figure 7: Cluster dendrograms and cluster plots based on MAMSP and COSS (for both main and subordinate clauses) and MAMSP and COSS (for main clauses only).

The cluster dendrograms and cluster plots in **Figure** 7 demonstrate the high accuracy of the classification models in differentiating between the baseline (represented by the yellow subset) and Romance languages (represented by the red and green subsets). Furthermore, the dendrogram highlights the distinctiveness of ancient Romance languages (denoted by the green subset, which includes Latin-Perseus, Latin-PROIEL, and Latin-ITTB), as they consistently group together in a separate category, along with Romanian-Nonstandard. The underlying reasons for this grouping of Romanian-Nonstandard has been discussed in **Sections 3.1** and **3.2**. This is also the case for the cluster plots shown in the right panels of **Figure** 7.

In other words, in these two-dimensional spaces the sub-branches represent the typical and highly analytic languages of Chinese, the more synthetic Romance languages, and the most synthetic Latin. This distribution demonstrates that Chinese corpora exhibit a lower degree of morphological complexity and lower word order flexibility compared to Romance languages, which are themselves less complex and flexible than Latin.

An interesting finding here is the dynamic adaptation between morphological richness and word order freedom over time. As Modern Romance languages have reduced their morphological features, they have correspondingly developed more rigid syntactic word orders. This aligns with the theoretical assumption that trade-offs between morphology and syntax reflect the principle of least effort and synergetic linguistics (Zipf 1965; Köhler 1987, 2005; Koplenig, et al. 2017; Yan/Liu 2021; Li/Liu 2024), suggesting that humans strive to encode and decode linguistic information efficiently.

To be specific, flexible word order demands greater cognitive resources to encode and decode more detailed morphological information. Conversely, fixed word order reduces the cognitive load required to encode and decode complex morphological rules, allowing grammatical relationships to be conveyed more straightforwardly (Yan/Liu 2021, 148; Li, et al. 2022). Our empirical results support this hypothesis, showing that the trade-off between morphology and syntax in Romance languages has evolved to optimize communication efficiency. This quantitative perspective provides valuable evidence for the principle of efficient communication in linguistic evolution.

4 Conclusion

This study employs 23 corpora made up of Latin and seven Modern Romance languages from the UD 2.5 database as primary research objects, complemented by four corpora of typical analytic languages (Classical and Modern Chinese) as baselines. The research investigates the correlation between morphological richness and word order freedom within the Romance languages, and explores the potential of clustering these languages based on their morphosyntactic features.

Using the quantitative metrics of MAMSP and COSS, morphological and syntactic features were extracted and quantified from annotated corpora. The statistical results confirm the consistency and reliability of the adopted indicators, especially the COSS metrics considering both main and subordinate clauses and those considering main clauses only. Notably, the well-known complexity trade-off hypothesis holds within Romance languages, indicating a negative relationship between morphological richness and word order rigidity. Additionally, clustering analysis based on these indicators effectively differentiates Chinese from Romance languages: the ancestor of Romance languages, Latin, exhibits distinct morphological and syntactic characteristics compared to Modern Romance languages.

Unlike previous studies that primarily relied on qualitative analyses, this research adopts a quantitative approach to explore the diachronic evolution of Romance languages. By systematically investigating historical changes in morphological and syntactic structures, it uncovers broader trends in the typological development of languages over time. The study not only enriches our understanding of language change within the Romance family but also opens up new perspectives for typological inquiry through corpus-based methods.

The findings highlight the importance of examining linguistic evolution from a diachronic perspective, and demonstrate how quantitative analysis can reveal patterns and trends over time. Future research could explore additional aspects of language change within this language family, or could conduct comparative studies with other language families to gain deeper insights into the broader dynamics of linguistic evolution.

5 Bibliography

Barbu Mititelu, Verginica/Mitrofan, Maria, *The Romanian Medical Treebank — SiMoNERo*, in: Barbu Mititelu, Verginica/Irimia, Elena/Tufiş, Dan/Cristea, Dan (edd.), *Proceedings of the 15th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing-ConsILR*, 2020, 7–16.

Best, Karl-Heinz/Rottmann, Otto, *Quantitative Linguistics, An Invitation*, Lüdenschied, Ram-Verlag, 2017.

Bickel, Balthasar/Nichols, Johanna, *Inflectional synthesis of the verb*, in: Haspelmath, Martin/Dryer, Matthew S./Gil, David/Comrie, Bernard (edd.), *The World Atlas of Language Structures*, Oxford, Oxford University Press, 2005, 94–97.

Bonfante, Guillaume/Guillaume, Bruno/Perrier, Guy, *Application of Graph Rewriting to Natural Language Processing*, Hoboken, John Wiley & Sons, 2018.

Buchi, Éva/Chauveau, Jean-Paul, *From Latin to Romance*, in: Müller, Peter/Ohnheiser, Ingeborg/Olsen, Susan/Rainer, Franz (edd.), *Word-Formation: An International Handbook of the Languages of Europe*, vol. 3, Berlin, De Gruyter, 2015, 1931–1957.

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

Courtin, Marine, *Mesures de Distances Syntaxiques Entre Langues Àpartir de Treebanks*, Paris, Université Paris III — Sorbonne Nouvelle, 2018.

Covington, Michael/McFall, Joe, *Cutting the Gordian knot: The moving-average type-token ratio (MATTR)*, Journal of Quantitative Linguistics 17 (2010), 94–100.

Crystal, David, *The Cambridge Encyclopedia of Language (2nd edition)*, Cambridge, Cambridge University Press, 1997.

Dahl, Östen, *The Growth and Maintenance of Linguistic Complexity*, Amsterdam, John Benjamins, 2004.

Dryer, Matthew S., *The Greenbergian word order correlations*, Language 68 (1992), 81–138.

Dryer, Matthew S., *Order of subject, object and verb*, in: Haspelmath, Martin/Dryer, Matthew S./Gil, David/Comrie, Bernard (edd.), *The World Atlas of Language Structures*, Oxford, Oxford University Press, 2005, 330–333.

Fedzechkina, Maryia/Newport, Elissa/Jaeger, Florian, *Balancing effort and information transmission during language acquisition: Evidence from word order and case marking*, Cognitive Science 41 (2017), 416–446.

Gerdes, Kim/Kahane, Sylvain/Chen, Xinying, *Typometrics: From implicational to quantitative universals in word order typology*, Glossa: A Journal of General Linguistics 6 (2021), 17.

Greenberg, Joseph Harold, *A quantitative approach to the morphological typology of language*, International Journal of American Linguistics 26 (1960), 178–194.

Greenberg, Joseph Harold, *Some universals of grammar with particular reference to the order of meaningful elements*, in: Greenberg, Joseph Harold (ed.), *Universals of Language*, London, MIT Press, 1963, 73–113.

Gulordava, Kristina/Merlo, Paola, *Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of Latin and ancient Greek*, in: Nivre, Joakim/Hajičová, Eva (edd.), *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, Uppsala, Uppsala University, 2015, 121–130.

Haspelmath, Martin/Michaelis, Susanne Maria, *Analytic and synthetic*, in: Buchstaller, Isabelle/Siebenhaar, Beat (edd.), *Language Variation-European Perspectives VI: Selected Papers from the Eighth International Conference on Language Variation in Europe*, Amsterdam, John Benjamins, 2017, 3–21.

Heringer, Hans Jürgen, *Dependency syntax: Basic ideas and the classical model*, in: Jacobs, Joachim/von Stechow, Arnim/Sternefeld, Wolfgang/Vennemann, Theo (edd.), *Syntax: An International Handbook of Contemporary Research*, vol. 1, Berlin, De Gruyter, 1993, 298–316.

Hitchins, Keith, *The Romanians, 1774–1866*, Oxford, Oxford University Press, 1996.

Hockett, Charles, *A Course in Modern Linguistics*, New York, Macmillan Publishers, 1958.

Hudson, Richard, Measuring syntactic difficulty, 1995, Available at <https://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf>.

Jiang, Jingyang/Liu, Haitao (edd.), *Quantitative Analysis of Dependency Structures*, Berlin, De Gruyter, 2018.

Köhler, Reinhard/Altmann, Gabriel, *Aims and scope of quantitative linguistics*, in: Mohanty, Panchanan /Köhler, Reinhard (edd.), *New Readings in Quantitative Linguistics*, New Delhi, Indian Institute of Language Studies, 2008, 1–32.

Köhler, Reinhard/Altmann, Gabriel/Piotrowski, Rajmund G. (edd.), *Quantitative Linguistics – An International Handbook*, Berlin, De Gruyter, 2005.

Kuboň, Vladislav/Lopatková, Markéta/Hercig, Tomás, *Searching for a measure of word order freedom*, in: Brejová, Brona (ed.), *Proceedings of the 16th ITAT Conference Information Technologies — Applications and Theory*, CEUR Workshop Proceedings, 2016, 11–17.

Ledgeway, Adam, *Syntactic and morphosyntactic typology and change*, in: Maiden, Martin/Smith, John Charles/Ledgeway, Adam (edd.), *The Cambridge History of the Romance Languages*, Cambridge, Cambridge University Press, 2011, 382–471.

Ledgeway, Adam, *From Latin to Romance: Morphosyntactic Typology and Change*, Oxford, Oxford University Press, 2012.

Li, Baoli/Han, Liping, *Distance weighted cosine similarity measure for text classification*, in: Yin, Hujun/Tang, Ke/Gao, Yang/Klawonn, Frank/Lee, Minho/Weise, Thomas/Li, Bin/Yao, Xin (edd.), *Intelligent Data Engineering and Automated Learning — IDEAL 2013*, New York, Springer, 2013, 611–618.

Li, Charles/Thompson, Sandra (edd.), *Mandarin Chinese: A Functional Reference Grammar*, California, University of California Press, 1989.

Liu, Haitao/Xu, Chunshan, *Quantitative typological analysis of Romance languages*, Poznań Studies in Contemporary Linguistics 48 (2012), 597–625.

Li, Wenchao/Liu, Haitao, *Complexity trade-off in morphosyntactic module: Suggestions from Japanese dialects*, Poznan Studies in Contemporary Linguistics 60 (2024), 159–187.

Li, Wenping/Liu, Haitao/Xiong, Zihan, *A quantitative study of word order freedom and case marking richness in Japanese*, Mathematical Linguistics 33 (2022), 325–340.

Maiden, Martin, *Linguistic History of Italian*, New York, Routledge, 2014.

Mărănduc, Cătălina/Malahov, Ludmila/Perez, Cenel-Augusto/Colesnicov, Alexander, *RoDia project of a regional and historical corpus for Romanian*, in: Cojocaru, Svetlana/Nikitchenko, Mykola/Drugus, Ioachim/Iftene, Adrian (edd.), Proceedings of *Conference on Mathematical Foundations of Informatics*, Chisinau, Academy of Sciences of Moldova, 2016, 268–284.

Mărănduc, Cătălina/Perez, Cenel-Augusto/Simionescu, Radu, *Social media-processing Romanian chat and discourse analysis*, Computación y Sistemas 20 (2016), 405–414.

Mărănduc, Cătălina/Bobicev, Victoria, *Non standard treebank Romania – Republic of Moldova in the Universal Dependencies*, in: Cojocaru, Svetlana/Gaindric, Constantin/Drugus, Ioachim (edd.), Proceedings of *Conference on Mathematical Foundations of Informatics*, Chisinau, Academy of Sciences of Moldova, 2017, 111–116.

McFadden, Thomas, *On morphological case and word-order freedom*, in: Nowak, Pawel/Yoquelet, Corey/Mortensen, David (edd.), *Proceedings of the 29th Annual Meeting of the Berkeley Linguistics Society, General Session and Parasession on Phonetic Sources of Phonological Patterns: Synchronic and Diachronic Explanations*, Berkeley, Berkeley Linguistics Society, 2003, 295–306.

Mel'čuk, Igor Aleksandrovic, *Dependency Syntax: Theory and Practice*, New York, State University of New York Press, 1988.

Mitrofan, Maria/Barbu Mititelu, Verginica/Mitrofan, Grigorina, *MoNERo: A biomedical gold standard corpus for the Romanian language*, in: Demner-Fushman, Dina/Cohen, Kevin Bretonnel/Ananiadou, Sophia/Tsujii, Junichi (edd.), *Proceedings of the 18th BioNLP Workshop and Shared Task*, Florence, Association for Computational Linguistics, 2019, 71–79.

Muflikhah, Lailil/Baharudin, Baharum, *Document clustering using concept space and cosine similarity measurement*, in: Jusoff, Hj. Kamaruzaman/Othman, Mohamed/Xie, Yi (edd.), *2009 International Conference on Computer Technology and Development (ICCTD 2009)*, Washington, IEEE Computer Society, 2009, 58–62.

Nivre, Joakim, *Towards a universal grammar for natural language processing*, in: Gelbukh, Alexander (ed.), *Lecture Notes in Computer Science*, vol. 9041, Cham, Springer, 2015, 3–16.

Norman, Jerry, *Chinese*, Cambridge, Cambridge University Press, 1988.

Posner, Rebecca, *The Romance Languages*, Cambridge, Cambridge University Press, 1996.

Sapir, Edward, *Language: An Introduction to the Study of Speech*, New York, Harcourt, Brace & World, 1921.

Schaller, Helmut, *Slavic Elements in Modern Romanian and Their History*, Vienna, Austrian Academy of Sciences Press, 2023.

Schwegler, Armin, *Analyticity and Syntheticity: A Diachronic Perspective with Special Reference to Romance Languages*, Berlin, De Gruyter, 1990.

Tesnière, Lucien, *Eléments de la Syntaxe Structurale*, Paris, Klincksieck, 1959.

Tily, Harry, *The Role of Processing Complexity in Word Order Variation and Change*, Stanford, Stanford University, 2010.

Tomlin, Russell S, *Basic Word Order: Functional Principles*, London, Croom Helm, 1986.

Vincent, Nigel, *Continuity and change from Latin to Romance*, in: Adams, James Noel/Vincent, Nigel (edd.), *Early and Late Latin: Continuity or Change?*, Cambridge, Cambridge University Press, 2016, 1–13.

Xanthos, Aris/Gillis, Steven, *Quantifying the development of inflectional diversity*, First Language 30 (2010), 175–198.

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

Yan, Jianwei/Liu, Haitao, *Basic word order typology revisited: A cross-linguistic quantitative study based on UD and WALS*, Linguistics Vanguard 9 (2024), 73–85.

Yan, Jianwei, *The current state and prominent features of quantitative linguistics through the lens of QUALICO 2023: A conference report*, Journal of Quantitative Linguistics 31 (2024), 54–67.

Zeman, Daniel/Nivre, Joakim/Abrams, Mitchell/Aepli, Noëmi/Agić, Željko/Ahrenberg, Lars, et al., *Universal Dependencies 2.5. Universal Dependecies Consortium*, 2019, Available at <https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3105>.

Appendices

Appendix I Details on the 27 corpora

Corpus	Language	Sub-branch
Catalan-AnCora	Catalan	Ibero-Romance
French-FQB	French	Gallo-Romance
French-GSD	French	Gallo-Romance
French-ParTUT	French	Gallo-Romance
French-Sequoia	French	Gallo-Romance
French-Spoken	French	Gallo-Romance
Galician-CTG	Galician	Ibero-Romance
Galician-TreeGal	Galician	Ibero-Romance
Italian-ISDT	Italian	Italo-Romance
Italian-ParTUT	Italian	Italo-Romance
Italian-PoSTWITA	Italian	Italo-Romance
Italian-PUD	Italian	Italo-Romance
Italian-TWITTIRO	Italian	Italo-Romance
Italian-VIT	Italian	Italo-Romance
Latin-ITTB	Latin	Latin
Latin-Perseus	Latin	Latin
Latin-PROIEL	Latin	Latin
Portuguese-Bosque	Portuguese	Ibero-Romance
Romanian-Nonstandard	Romanian	Eastern Romance
Romanian-RRT	Romanian	Eastern Romance
Romanian-SiMoNERo	Romanian	Eastern Romance
Spanish-AnCora	Spanish	Ibero-Romance
Spanish-GSD	Spanish	Ibero-Romance

Corpus	Language	Sub-branch
Chinese-CFL	Chinese	Sino-Tibetan
Chinese-GSD	Chinese	Sino-Tibetan
Chinese-GSDSimp	Chinese	Sino-Tibetan
Classical_Chinese-Kyoto	Classical Chinese	Sino-Tibetan

Appendix II: Morphological richness of the 27 corpora

Corpus	MAMSP	Language	Sub-branch
Romanian-Nonstandard	1.474214	Romanian	Eastern Romance
Latin-ITTB	1.447406	Latin	Latin
Latin-PROIEL	1.388395	Latin	Latin
Latin-Perseus	1.254863	Latin	Latin
Romanian-RRT	1.196385	Romanian	Eastern Romance
French-Spoken	1.195184	French	Gallo-Romance
French-ParTUT	1.193890	French	Gallo-Romance
Italian-VIT	1.193363	Italian	Italo-Romance
French-Sequoia	1.190611	French	Gallo-Romance
Romanian-SiMoNERo	1.187633	Romanian	Eastern Romance
Italian-ParTUT	1.183984	Italian	Italo-Romance
Italian-ISDT	1.179742	Italian	Italo-Romance
Italian-TWITTIRO	1.174830	Italian	Italo-Romance
French-FQB	1.167354	French	Gallo-Romance
Italian-PoSTWITA	1.165279	Italian	Italo-Romance
Catalan-AnCora	1.155326	Catalan	Ibero-Romance
Galician-TreeGal	1.153818	Galician	Ibero-Romance
Italian-PUD	1.149822	Italian	Italo-Romance
Spanish-AnCora	1.145682	Spanish	Ibero-Romance
French-GSD	1.133480	French	Gallo-Romance
Spanish-GSD	1.118590	Spanish	Ibero-Romance
Portuguese-Bosque	1.117207	Portuguese	Ibero-Romance
Galician-CTG	1.114442	Galician	Ibero-Romance
Chinese-CFL	1.001563	Chinese	Baseline
Chinese-GSDSimp	1.001454	Chinese	Baseline
Chinese-GSD	1.001391	Chinese	Baseline
Classical_Chinese-Kyoto	0.999993	Classical Chinese	Baseline

* The values are ranked in descending order according to the MAMSP values.

Appendix III: Word order freedom of the 27 corpora based on both main clauses and subordinate clauses

Corpus	COSS (main and subordinate clauses)	Language	Sub-branch
Latin-Perseus	0.802026	Latin	Latin
Romanian-Nonstandard	0.779436	Romanian	Eastern Romance
Latin-PROIEL	0.779048	Latin	Latin
Latin-ITTB	0.695514	Latin	Latin
Spanish-GSD	0.641961	Spanish	Ibero-Romance
Catalan-AnCora	0.596312	Catalan	Ibero-Romance
Spanish-AnCora	0.584069	Spanish	Ibero-Romance
Galician-CTG	0.555904	Galician	Ibero-Romance
Italian-PoSTWITA	0.550137	Italian	Italo-Romance
Galician-TreeGal	0.524038	Galician	Ibero-Romance
French-Spoken	0.522754	French	Gallo-Romance
French-ParTUT	0.518288	French	Gallo-Romance
Italian-TWITTIRO	0.518036	Italian	Italo-Romance
French-FQB	0.494413	French	Gallo-Romance
Romanian-RRT	0.485427	Romanian	Eastern Romance
French-GSD	0.478951	French	Gallo-Romance
Italian-ISDT	0.478287	Italian	Italo-Romance
Italian-VIT	0.475431	Italian	Italo-Romance
French-Sequoia	0.469490	French	Gallo-Romance
Portuguese-Bosque	0.467377	Portuguese	Ibero-Romance
Italian-ParTUT	0.460128	Italian	Italo-Romance
Italian-PUD	0.440346	Italian	Italo-Romance
Chinese-GSDSimp	0.438902	Chinese	Baseline
Chinese-GSD	0.438897	Chinese	Baseline
Classical_Chinese-Kyoto	0.421589	Classical Chinese	Baseline
Romanian-SiMoNERo	0.417338	Romanian	Eastern Romance
Chinese-CFL	0.410452	Chinese	Baseline

* The values are ranked in descending order according to the COSS values.

Appendix IV: Word order freedom of the 27 corpora based on the main clauses only

Corpus	COSS (main clauses)	Language	Sub-branch
Latin-Perseus	0.793214	Latin	Latin
Latin-PROIEL	0.774430	Latin	Latin
Romanian-Nonstandard	0.768125	Romanian	Eastern Romance
Latin-ITTB	0.692569	Latin	Latin
Catalan-AnCora	0.617043	Catalan	Ibero-Romance
Spanish-AnCora	0.593932	Spanish	Ibero-Romance
Galician-CTG	0.563713	Galician	Ibero-Romance
French-Spoken	0.561221	French	Gallo-Romance
Galician-TreeGal	0.532583	Galician	Ibero-Romance
French-ParTUT	0.526311	French	Gallo-Romance
Italian-PoSTWITA	0.502648	Italian	Italo-Romance
Spanish-GSD	0.483907	Spanish	Ibero-Romance
Italian-TWITTIRO	0.480938	Italian	Italo-Romance
French-GSD	0.477861	French	Gallo-Romance
French-Sequoia	0.476555	French	Gallo-Romance
Italian-VIT	0.473558	Italian	Italo-Romance
Romanian-RRT	0.473501	Romanian	Eastern Romance
Italian-ISDT	0.469827	Italian	Italo-Romance
Portuguese-Bosque	0.465281	Portuguese	Ibero-Romance
French-FQB	0.464207	French	Gallo-Romance
Italian-ParTUT	0.461698	Italian	Italo-Romance
Italian-PUD	0.435240	Italian	Italo-Romance
Classical_Chinese-Kyoto	0.422787	Classical Chinese	Baseline
Romanian-SiMoNERo	0.414387	Romanian	Eastern Romance
Chinese-CFL	0.410004	Chinese	Baseline
Chinese-GSD	0.408248	Chinese	Baseline
Chinese-GSDSimp	0.408248	Chinese	Baseline

* The values are ranked in descending order according to the COSS values.