# Cross-linguistic harmony on word Order Features and Surface Word Order: Mandarin Chinese as a Mixed Featured Order with a Surface VO Order[1]

**Abstract:** This study is rooted in the grammatical typology of OV-VO word order and employs the Universal Dependency corpus to explore the relationship between word order feature and surface word order, and reexamining of word order classification models not only of Mandarin Chinese but also across diverse global languages. It evaluates the binary word orders of 78 languages and reveals the cross-linguistic harmony between word order features and surface word order features. Out of more than thirty dependency relations, six pairs display a highly significant bidirectional association with word order types. Moreover, by considering word order and straightforward textual frequency measurements, a substantial linear correlation emerges between word order features and surface word order across global languages, highlighting cross-linguistic harmony. Building upon this foundation, the paper introduces a two-dimensional model for word order classification. It contends that Mandarin typifies a surface VO structure, representing weakly harmonious language with feature mixing (include both “neutral feature” and “feature mixture”). This characteristic is likely linked to Mandarin's approach to encoding prototypical transitive events and its linguistic contact environment.

**Key words:** word order typology, Mandarin Chinese, harmony, annotated corpus

# 1. Introduction

The classification of Mandarin Chinese as a VO language, OV language, or a mixed OV-VO language has been a topic of debate among scholars. While there are differing views, an examination of simple textual frequency in corpora suggests that Mandarin Chinese predominantly exhibits a VO word order. However, some scholars argue for Mandarin Chinese is an OV (SOV) language. For instance, Li & Thompson (1974, 1975) propose that Mandarin Chinese underwent a transition from SVO to SOV. Nevertheless, Sun & Givon (1985: 329) refute this viewpoint based on textual and functional distribution, highlighting the absence of evidence for Mandarin Chinese being SOV from the perspective of child language acquisition as well. On the other hand, Hashimoto Bantaro (1996a) suggests the existence of a continuum from VO to OV in Mandarin Chinese based on geographical distribution. The syntax of sentence structure is frequently SOV in northern varieties and SVO in southern ones. Grammatical modifiers contrast between to modifier-modified word order in the north and modified-modifier in the south (Wadley 1996: 102).

An alternative perspective posits that the syntactic structure of the Chinese language

[1] This paper provides supplementary files at https://osf.io/eg83a/?view_only=352763e72a4c45bd905b9e5841bf5019.

does not conform to a singular, unadulterated syntactic type; rather, it exhibits characteristics of a "mixed order." Jin Lixin (2016) employed the branching direction theory (BDT) (Dryer 1992, 2009) in linguistic typology to investigate the word order type of Mandarin Chinese. By analyzing parameters related to OV-VO, Jin found that Mandarin Chinese exhibits a mixed word order with characteristics of both OV and VO. Out of the 12 parameters associated with Mandarin Chinese, 8 showed agreement with VO structures, while 9 exhibited agreements with OV structures. This results in a Mandarin Chinese word order type index of VO = 8/12 = 0.66 and OV = 9/12 = 0.75. The difference between 0.66 and 0.75 is less than 0.1, indicating a relatively small discrepancy, which is the evidence that Mandarin Chinese is a language with “mixed order”.

While this method considers Mandarin Chinese's word order characteristics based on the branching direction theory, which is relatively objective and has typological universality, it still has its limitations.

Firstly, the BDT lacks clarity in describing correlation. For instance, VO-OV languages and the postposition-preposition correlation in relative clauses are related. However, according to Dryer (1992) and the data from the World Atlas of Language Structures (WALS), the predictive accuracy for VO languages is 96.35%, while for OV languages, it is only 78.56%[2]. This discrepancy suggests an unequal strength in this "correlation." By assigning equal points for meeting this correlation, to some extent, the differences are obscured. Furthermore, if we consider text frequency, a language that allows both prepositions and postposition might have a very low frequency, even being an isolated occurrence, while another language may have frequent instances. In this case, “existence as evidence “would not be reasonable. Therefore, a new method is needed to accurately calculate the relevant factors for VO-OV word order and further measure and operationally define the "word order features."

Secondly, when we focus solely on "word order features," we to some extent overlook the OV/VO word order itself and the interaction between OV/VO and the correlated parameters. For instance, Liu (2010) identified, through dependency-based classification of word order, that Chinese occupies an extreme position within the continuum of Head-Dependent word order. This finding contradicts the notion of a "mixed word order" proposed by Jin Lixin (2016). Therefore, it is imperative to provide a coherent explanation for the underlying mechanisms that give rise to such a conflict. The influence of word order features on word order is highly complex, as the causes of word order variation can stem from language history, context, pragmatic conventions, phonological morphology, and other factors. Therefore, it is necessary to measure both the OV/VO word order and word order parameters, discover their interactive relationship, identify typological correlations, and further infer possible typological motivations.

Thirdly, given that this approach aims to achieve a refined measurement of word order, utilizing annotated corpora to investigate word order of Chinese and world languages is a promising means. The use of annotated corpora or treebanks (Abeillé 2003) as a novel resource for studying word order typology has gained momentum (Gerdes et al. 2021), which can contribute to a research based on Corpus. Cross-linguistic corpora have

[2] This data is circulated based on Feature 85A & Feature 83A on WLAS (Dryer 2013)

become new source for validating theoretical linguistic hypotheses or discovering the further linguistic universal. For instance, Liu (2010) discovered a head-dependent continuum for world languages through annotated treebanks, while Yan & Liu (2023) statistically validated the first five implicational universals proposed by Greenberg (1963), and Levshina (2019) verified principles such as case strategy and word order strategy (cf. Siewierska & Bakker 2008), and also the Heaviness Principle proposed by Hawkins (1983). The emerging trend from these studies suggests that corpus-based typological research might offer a better description and capture the unique features of intra-language variation, serving as evidence or refutation for existing qualitative research, which is also referred to as token-based typology (Levshina 2019). The typological findings derived from annotated corpora based on substantial real language data (rather than relying solely on reference grammars or introspective data) have propelled the advancement of typology, especially in the goal of inducing universal correlations and implicational relationships directly from usage events (Bickel 2010, Bickel 2015).

Based on these discussions, this paper considers the approach proposed by Jin Lixin (2016) as deserving of refinement and extension. In light of this, the present study aims to reevaluate the syntactic word order types in Standard Mandarin Chinese. It endeavors to devise a novel methodology for assessing syntactic word order, encompassing both the frequency of Verb-Object (VO) and Object-Verb (OV) textual occurrences, thereby offering a two-dimensional dataset for syntactic word order characteristics. Furthermore, the study aspires to unveil latent cross-linguistic syntactic commonalities and potentially establish novel paradigms for syntactic classification.

This paper is arranged as following: In Section 2, this paper introduces the corpus used in the study. In Section 3, the paper analyzes the significance and validity of the correlation of word order parameters in syntax through an analogy with the "confusion matrix" in machine learning. It utilizes the "typological matrix" and two derived parameters: "PRP (Forward Predicting Rates)"&"FPR (Retrorse Predicting Rates)" to detect six core word order parameters, referred to as "word order features." In Section 4, based on measurements along two dimensions, the paper Measures the two-dimensional word order of languages and discovers a linear correlation between the surface word order and featured word order of world languages, indicating cross-linguistic harmony, which provided a futher evidence of word order distribution and cognitive principle word order rules. It also proposes a typological classification system for word order based on these two dimensions. In Section 5, the paper conducts a case analysis of several samples, revealing two major configurations of mixed features: "neutral features " and " features balanced" Section 6 briefly discusses the motivations behind such mixed word orders from both intra-linguistic and extra-linguistic perspectives.

## 2. Corpus

The quantitative research in this paper is based on the Universal Dependencies (UD) corpus, annotated corpora or treebanks (Abeillé 2003) have been using in investigating word order universal or variety (e.g., Liu 2010, Yan & Liu 2023, Levshina 2019, Gerdes et

al. 2021). UD is a cross-linguistic grammatical annotation framework that is primarily grounded in dependency grammar. Its goal is to provide a consistent annotation scheme for various natural languages, enabling both intra-linguistic and cross-linguistic comparisons and research[3]. Under the annotation guidelines of UD, each word is annotated with its part-of-speech (POS) tag, and syntactic relationships are primarily represented by the dependency relations and directions between words. These POS tags and syntactic relationships align closely with the terminology used in traditional grammar analysis. Here is an example of an English annotation in the UD framework:

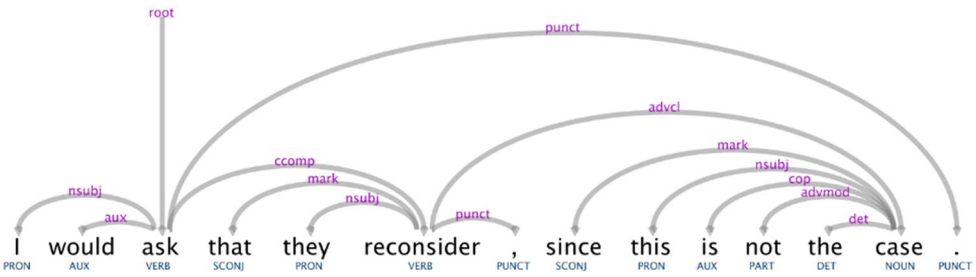


Figure 1: English example annotated in the UD (Universal Dependencies)

UD, as a cross-linguistic corpus based on dependency annotation, has aimed to facilitate typological research since its inception. This research has unique advantages in investigating language-specific word order patterns. Due to the fact that dependency relations only involve two grammatical elements and their relative positions, they can be readily transformed into word order patterns. For instance, in the sentences presented in Figure 1, the dependency relation between "ask" and "I" is nsubj (nominal subject), with a right-to-left dependency direction. From this, we can infer the SV (subject-verb) word order. This type of typological research conducted through the analysis of exemplar data is referred to as token-based typology (Levshina 2019), and it provides quantitative evidence for linguistic universals.

However, it should be noted that the part-of-speech and dependency annotations in UD are largely aligned with traditional grammar, which means that they primarily represent descriptive categories rather than comparative concepts in linguistic typology. Comparative concepts, as discussed by Haspelmath (2010), are more suitable for cross-linguistic comparisons. Therefore, the annotations in UD cannot be easily substituted for the concepts used in general typological research.

This study selected corpora from 78 languages, ensuring that each language had a minimum of 1000 sentences in the sample. The selected languages include major Indo-European languages, major Asian languages, as well as some African languages, and a small number of languages from other regions. It also includes several diachronic languages such as Classical Chinese, Classical Greek, and Classical Hebrew, as well as regional dialects such as Cantonese. Although the selected sample is not perfectly balanced in terms of affiliation and geographic distribution. But considering that this represents the maximum number of languages that can be obtained while ensuring an

[3] For detail information: http://universaldependencies.org/introduction.html.

adequate corpus size, overall, the linguistic data for the study is sufficiently comprehensive and generally meets the requirements for typological sampling. The following provides sample information on the regions, featured classifications, corpora, and their sizes for selected languages, the full list of corpora is included in supplementary files:

Table 1: Corpora and their sizes for selected languages in the study

| Language | Region | Affiliation | Corpus | Sentences |
|---|---|---|---|---|
| **Chinese** | East Asia | Sino-Tibetan | UD_ChineseGSDSimp@2.11 | 4,997 |
| **English** | Europe/America | Germanic | UD_English-EWT@2.11 | 16,623 |
| **Akkadian** | Africa | Semitic | UD_Akkadian-RIAO@2.11 | 1,874 |
| **Turkish** | Asia/Europe | Turkic | UD_Turkish-Tourism@2.11 | 19,833 |
| **Hungarian** | Europe | Uralic | UD_Hungarian-Szeged@2.11 | 1,800 |

# 3. Six Word Order Parameters Identified from the Word Order Matrix

## 3.1 Word Order Matrix and Forward/Retrorse Predicting Rates

This study introduces the concept of the Word Order Matrix as an analogy to the Confusion Matrix commonly used in machine learning. Additionally, two related metrics, the Forward Predicting Rate (FPR) and the Retrorse Predicting Rate (RPR), are proposed to address the efficiency issue in describing language correlation.

Firstly, we select two binary features from world languages. The first feature is VO-OV word order, and the second feature is another word order related feature, such as adjctive clause modifier (namely, relative clause).

For example, assuming that acl-N (adjective clause-Noun) and OV word order co-occur, we label aclN as the "ov" feature and similarly label N-acl as the "vo" feature. Based on this, any distribution of these two features can be summarized in the following matrix, where the numbers in parentheses represent the frequency of the corresponding dominant[4] word order patterns observed in the 78 languages:

Table2: Word Order Matrix of VO-OV and acl/N

| | OV | VO |
|---|---|---|
| Feature $^{ov}$（acl-N） | $T^{ov}$（11） | $F^{vo}$（6） |
| Feature $^{vo}$（N-acl） | $F^{ov}$（9） | $T^{vo}$（42） |

[4] In this paper, “dominant” is simply defined as a frequency greater than 0.5, because there is rare evidence to provide a more reasonable threshold, despite a higher threshold is applied in other studies (e.g., Dryer 2013, Yan & Liu 2023).

According to the word order matrix, we can calculate two metrics:

(1) Forward Predicting Rate (FPR):

FPR represents the extent to which the OV-VO word order can predict a word order feature. A higher FPR indicates that the surface word order can better predict a particular word order feature. For example:

**For VO languages, $FPR^{vo} = T^{vo} / (T^{vo} + F^{vo}) = 55.0\%$.**
**For OV languages, $FPR^{ov} = T^{OV} / (T^{OV} + F^{OV}) = 87.5\%$.**

The data observed in this study shows that among languages with a dominant word order of VO dominant, 55.0% of them use dominant acl-N order, while among languages with a dominant order of OV, only 70.5% of them use dominant N-acl order.

(2) Retrorse Predicting Rate (RPR):

RPR represents the extent to which a word order feature can predict the OV-VO word order. A higher RPR indicates that a particular word order feature can better predict the surface word order. For example:

**For VO languages, $RPR^{vo} = T^{vo} / (T^{vo} + F^{OV}) = 82.4\%$.**
**For OV languages, $RPR^{ov} = T^{OV} / ( T^{OV} + F^{vo}) = 64.7\%$.**

The data observed in this study shows that among all languages with dominant N-acl order, 82.4% of them are VO languages, while among all languages with acl-N order, only 64.7% of them are OV languages.

As a result, a description of the correlation between O/V order and adposition is calculated as table 3:

Table 3: the FPR & RPR of acl/N

| Patterns | $FPR^{OV}$ | $FPR^{VO}$ | $RPR^{VO}$ | $RPR^{OV}$ |
|---|---|---|---|---|
| **acl/N** | 0.550 | 0.875 | 0.824 | 0.647 |

Based on these findings, we can observe that although it is commonly said that VO languages tend to use N-acl order and OV languages tend to use acl-N order, the specific extent of this tendency and whether it holds true can be examined or validated using the FPR and RPR values from the word order matrix. For a clearer illustration of the calculation of FPR&PRP, figure 2 is exhibited.

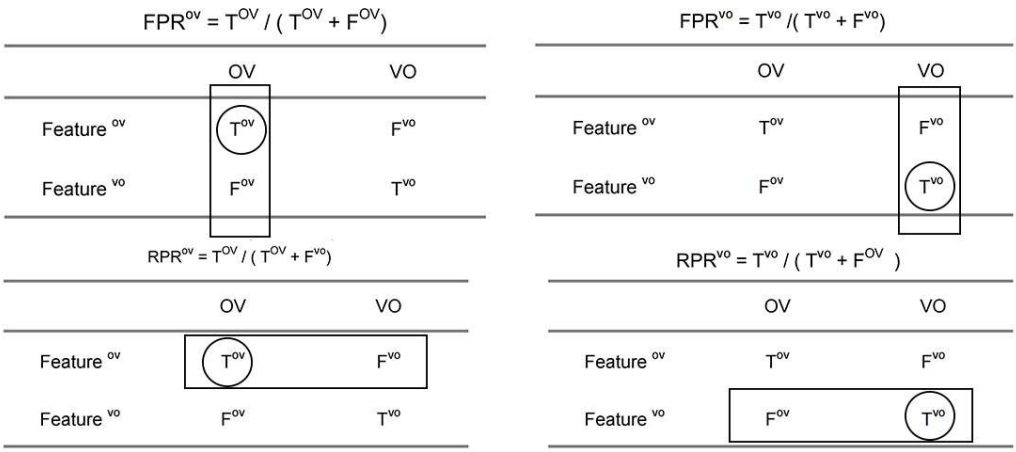


figure 2: the calculation of FPR&RPR

Next more word order patterns are to be examined. In the Word Order Patterns, for every pattern X/Y, the left side represents the dependent item X, while the right side represents the head of the dependency Y. The head is constrained to four part-of-speech types: N (noun), V (verb), A (adjective), and X (unrestricted head). To exclude potential error in annotation or extreme cases, when determining the word order category, the frequency less than 10 was excluded.

For the co-occurrence hypothesis of word order features, when X is dependent and Y is head, a classic theory argues that word order patterns satisfy the XY corresponds to OV feature and YX corresponds to VO feature (cf. Lehmann 1973, 1978, Vennemann 1973, 1974a, 1974b, 1976), which is designed because of the belief that O is the dependent while V is the Head. Consequently, Dryer (1992) argued this predictive approach is the Head-Dependent Theory (HDT), which assumes that all head-dependent items have the same word order. However, for patterns such as aux/V, Case/N, cop/N, cop/A, cop/X, and numd/N, the FPR and the RPR values predicted by HDT are very low. Therefore, for this set of word order patterns, we hypothesize that YX is the OV feature and XY is the VO feature, with the latter being considered a modification of the Branching Direction Theory (BDT) proposed by Dryer (1992) in relation to the HDT theory. Namely, for those patterns, when X is dependent and Y is head, most word order patterns satisfy the YX corresponds to OV feature and XY corresponds to VO feature, which will be called “non-HDT model” in this paper. Additionally, this paper provides detailed data for these two patterns in an additional dataset, which offers a clear comparation between two different calculations (namely, HDT/non-HDT).

Basing on above adjustment, the FPR and RPR values calculated based on the word order matrix are provided (a file in supplementary files provides a further data). The word order patterns not satisfying HDT will be marked with an asterisk (*) in the table, the unmarked patterns are of HDT patterns. A significance levels of the chi-square analysis is also exhibited (Fisher's exact test), which determinate whether there is a statistically significant association between the category of each pattern and the O/V category:

Table 4: Correlation & FPR and RPR for 32 word order patterns

| Number | PPatterns | p-value | $FPR^{OV}$ | $FPR^{VO}$ | $RPR^{VO}$ | $RPR^{OV}$ |
|---|---|---|---|---|---|---|
| **1** | obj/V（O/V） | *NA* | 1.00 | 1.00 | 1.00 | 1.00 |
| **2** | acl/N | ***<0.01*** | 0.55 | 0.88 | 0.82 | 0.65 |
| **3** | advcl/V | ***<0.01*** | 0.70 | 0.79 | 0.85 | 0.59 |
| **4** | advcl/A | ***<0.05*** | 0.47 | 0.89 | 0.82 | 0.62 |
| **5** | advcl/X | ***<0.01*** | 0.65 | 0.81 | 0.84 | 0.60 |
| **6** | advmod/V | *0.092* | 0.96 | 0.20 | 0.92 | 0.35 |
| **7** | advmod/A | *0.211* | 0.95 | 0.08 | 0.80 | 0.28 |
| **8** | advmod/X | *0.162* | 0.96 | 0.17 | 0.90 | 0.34 |
| **9** | amod/N | *0.242* | 0.74 | 0.43 | 0.79 | 0.36 |
| **10*** | aux/V | ***<0.01*** | 0.55 | 0.91 | 0.82 | 0.55 |
| **11*** | case/N | ***<0.01*** | 0.71 | 0.87 | 0.87 | 0.71 |
| **12** | cc/N | *0.069* | 0.87 | 0.04 | 0.40 | 0.28 |
| **13** | ccomp/V | ***<0.01*** | 0.39 | 1.00 | 0.79 | 1.00 |
| **14** | ccomp/A | ***<0.01*** | 0.43 | 0.97 | 0.89 | 0.75 |
| **15** | ccomp/X | ***<0.01*** | 0.39 | 0.98 | 0.79 | 0.90 |
| **16*** | cop/N | ***<0.01*** | 0.74 | 0.94 | 0.90 | 0.74 |
| **17*** | cop/A | ***<0.01*** | 0.75 | 0.92 | 0.92 | 0.75 |
| **18*** | cop/X | ***<0.01*** | 0.74 | 0.92 | 0.90 | 0.74 |
| **19** | csubj/V | *0.127* | 0.53 | 0.76 | 0.78 | 0.50 |
| **20** | csubj/N | *0.302* | 0.39 | 0.84 | 0.77 | 0.50 |
| **21** | csubj/A | *0.149* | 0.36 | 0.90 | 0.75 | 0.63 |
| **22** | csubj/X | *0.163* | 0.41 | 0.80 | 0.78 | 0.44 |
| **23** | det/N | *0.445* | 0.83 | 0.09 | 0.56 | 0.29 |
| **24** | mark/X | ***<0.01*** | 0.41 | 1.00 | 0.81 | 0.41 |
| **25** | nmod/N | ***<0.01*** | 0.65 | 0.83 | 0.84 | 0.63 |
| **26** | nsubj/V | *0.121* | 1.00 | 0.13 | 1.00 | 0.34 |
| **27** | nsubj/N | *0.075* | 1.00 | 0.06 | 1.00 | 0.30 |
| **28** | nsubj/A | *0.482* | 0.95 | 0.08 | 0.80 | 0.29 |
| **29** | nsubj/X | *0.121* | 1.00 | 0.13 | 1.00 | 0.34 |
| **30*** | numd/N | *0.092* | 0.92 | 0.00 | 0.00 | 0.29 |
| **31** | obl/N | *0.036* | 0.74 | 0.60 | 0.85 | 0.42 |
| **32** | xcomp/V | ***<0.01*** | 0.48 | 0.98 | 0.83 | 0.91 |
| **33** | xcomp/A | *0.087* | 0.22 | 0.94 | 0.81 | 0.50 |
| **34** | xcomp/X | ***<0.01*** | 0.48 | 0.98 | 0.83 | 0.91 |

## 3.2 The exploring of Word Order features

Before exploring word order features, we need to provide an operational definition for them:

***Word Order features***: Word order features refer to a group of word order patterns that highly co-occur with VO-OV word orders or exhibit a strong correlation with them. These patterns can reliably predict (or imply) surface word orders across languages, and they can

also be strongly predicted (or implied) by surface word orders. If there is a bidirectional predictive or implicational relationship, it can be considered "***word order features***."

Therefore, a word order feature must satisfy two criteria. Firstly, it should demonstrate significant cross-linguistic correlations. Secondly, its predictions or implications should hold true for both OV and VO languages and both on FPR and PRP.

When we focus only on the significant correlations between word order parameters and OV-VO word order, we can identify the following significantly correlated parameters:

Table 5: Significantly correlated word order parameters

| Number | Patterns | p-value | $FPR^{OV}$ | $FPR^{VO}$ | $RPR^{VO}$ | $RPR^{OV}$ |
|---|---|---|---|---|---|---|
| **1** | acl/N | ***<0.01*** | 0.55 | 0.88 | 0.82 | 0.65 |
| **2** | advcl/V | ***<0.01*** | 0.70 | 0.79 | 0.85 | 0.59 |
| **3** | advcl/A | ***<0.05*** | 0.47 | 0.89 | 0.82 | 0.62 |
| **4** | advcl/X | ***<0.01*** | 0.65 | 0.81 | 0.84 | 0.60 |
| **5** | aux/V | ***<0.01*** | 0.55 | 0.91 | 0.82 | 0.55 |
| **6** | case/N | ***<0.01*** | 0.71 | 0.87 | 0.87 | 0.71 |
| **7** | ccomp/V | ***<0.01*** | 0.39 | 1.00 | 0.79 | 1.00 |
| **8** | ccomp/A | ***<0.01*** | 0.43 | 0.97 | 0.89 | 0.75 |
| **9** | ccomp/X | ***<0.01*** | 0.39 | 0.98 | 0.79 | 0.90 |
| **10** | cop/N | ***<0.01*** | 0.74 | 0.94 | 0.90 | 0.74 |
| **11** | cop/A | ***<0.01*** | 0.75 | 0.92 | 0.92 | 0.75 |
| **12** | cop/X | ***<0.01*** | 0.74 | 0.92 | 0.90 | 0.74 |
| **13** | mark/X | ***<0.01*** | 0.41 | 1.00 | 0.81 | 0.41 |
| **14** | nmod/N | ***<0.01*** | 0.65 | 0.83 | 0.84 | 0.63 |
| **15** | xcomp/V | ***<0.01*** | 0.48 | 0.98 | 0.83 | 0.91 |
| **16** | xcomp/X | ***<0.01*** | 0.48 | 0.98 | 0.83 | 0.91 |

However, in the table, we have highlighted some numbers that did not reach 50%. This means that for the corresponding dimensions, the word order prediction based on these parameters or the accuracy of predicting these parameters based on word order is relatively low. Although there is a correlation between these parameters and word order, their applicability may not be consistent in all cases. Therefore, further screening is required to ensure that the parameters, at each level and rounded to one decimal place, have FPR and RPR greater than or equal to 60%. The following patterns were found to be significantly correlated ($p < 0.01$) and exhibited high correlations in all four aspects:

Table 6: Significantly correlated and high-efficiency word order parameters

| Number | Patterns | p-value | $FPR^{OV}$ | $FPR^{VO}$ | $RPR^{VO}$ | $RPR^{OV}$ |
|---|---|---|---|---|---|---|
| **1** | acl/N | ***<0.01*** | 0.55 | 0.88 | 0.82 | 0.65 |
| **2** | advcl/V | ***<0.01*** | 0.70 | 0.79 | 0.85 | 0.59 |
| **3** | advcl/X | ***<0.01*** | 0.65 | 0.81 | 0.84 | 0.60 |
| **4** | aux/V | ***<0.01*** | 0.55 | 0.91 | 0.82 | 0.55 |
| **5** | case/N | ***<0.01*** | 0.71 | 0.87 | 0.87 | 0.71 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **6** | cop/N | ***<0.01*** | 0.74 | 0.94 | 0.90 | 0.74 |
| **7** | cop/A | ***<0.01*** | 0.75 | 0.92 | 0.92 | 0.75 |
| **8** | cop/X | ***<0.01*** | 0.74 | 0.92 | 0.90 | 0.74 |
| **9** | nmod/N | ***<0.01*** | 0.65 | 0.83 | 0.84 | 0.63 |

Due to the same dependency relationship shared by cop/N, cop/A, and cop/X, we need to decide how to select the word order within the same dependency relationship. After careful consideration, we choose to discard advcl/X and only select advcl/V because the predictive rate of advcl/A in $FPR^{OV}$ performs poorly and may interfere with the average value. As for cop/X, since it does not have the interference of outliers, it is reasonable to include cop/X in the core parameters for word order.

Based on the method outlined in Section 2.1, this paper examined the FPR and RPR or word order patterns.The study identified six cross-linguistically significant word order patterns that demonstrated strong correlations with surface word order at various levels.

Table 7: FPR and RPR for the Six Core Word Order features

| Number | Patterns | p-value | $FPR^{OV}$ | $FPR^{VO}$ | $RPR^{VO}$ | $RPR^{OV}$ |
|---|---|---|---|---|---|---|
| **1** | acl/N | ***<0.01*** | 0.55 | 0.88 | 0.82 | 0.65 |
| **2** | advcl/V | ***<0.01*** | 0.70 | 0.79 | 0.85 | 0.59 |
| **3** | aux/V | ***<0.01*** | 0.55 | 0.91 | 0.82 | 0.55 |
| **4** | case/N | ***<0.01*** | 0.71 | 0.87 | 0.87 | 0.71 |
| **5** | cop/X | ***<0.01*** | 0.74 | 0.92 | 0.90 | 0.74 |
| **6** | nmod/N | ***<0.01*** | 0.65 | 0.83 | 0.84 | 0.63 |

Compared to the BDT theory (Dryer, 1992), the six word order parameters identified in this study exhibit higher levels of replicability and operability. Therefore, we define them as the core word order features. These word order features will be utilized in Section 4 to measure the basic word order of world languages and explore the relationship between word order features and surface word order harmony.

# 4. Harmony between Featured Order and Surface Order

## 4.1 Surface Word Order and Featured Word Order

In this subsection, we focus on examining the relationship between word order features and surface word order. Therefore, operational definitions of surface word order and featured word order are required. It should be noted that both surface word order and featured word order need to consider two aspects: **nature** and **degree**. For example, Mandarin Chinese and French both have a dominnat VO word order, which is the same in nature. However, Mandarin Chinese has a VO frequency of 99.9%, while French has a frequency of 70.1%[5], indicating a difference in degree. Therefore, the definitions need to utilize the Simple Order Index (SOI) and the Parameterized Order Index (POI) proposed in

[5] The data is sourced from the following corpora: UD_Chinese-GSDSimp@2.12 and UD_French-GSD@2.12.

Section 4.2 as the foundation for their replicability.

**Surface Word Order:** Surface word order refers to the OV-VO word order category reflected directly by the frequency of simple texts in a language. Among multiple word orders observed in practice, the word order with the highest frequency is considered the surface word order type of that language. The degree of surface word order is described using the Simple Order Index (SOI).

**Featured Word Order:** Featured word order refers to the word order category derived from a language by calculating its word order featured. The overall word order category reflected by multiple word order features is considered the featured word order type of that language. The degree of featured word order is measured using the Parameterized Order Index (POI).

The calculation of SOI and POI will be introduced in Section 4.2.

## 4.2 Simple Order Index (SOI) and Parameterized Order Index (POI)

To facilitate quantitative analysis, we have defined two indices: the Simple Order Index (SOI) and the Parameterized Order Index (POI). What is need to be emphasized is in this paper, all the SOI and POI will be exhibited a normalized form, which helps to indicate the relative rank of word order. In a normalized form, -1 is the minim and 1 is the maxim. Namely -1 is most OV and 1 is most VO.

SOI is a straightforward measure of word order. It calculates the frequency of VO-OV patterns in a language corpus. The SOI is designed to be negative for purer OV languages and positive for purer VO languages. For example, based on the data from UD_Japanese-BCCWJ@2.11, all sentences in Japanese are OV, resulting in an SOI of approximately -1. This makes Japanese one of the most prototypical OV languages based on surface word order.

POI, on the other hand, is a word order index derived from the analysis of word order parameters. It can be summarized as the sum of the contributions of all word order parameters to the overall word order pattern. The contribution is determined by two factors: 1) frequency of occurrence, where higher frequency leads to a greater contribution for a particular parameter, and 2) RPR, where parameters with higher RPR have a greater contribution (not discussed further if unnecessary). For example, Vietnamese has significant contributions from parameters related to VO in terms of both frequency and RPR, resulting in a POI of approximately 0.98. From a featured word order perspective, Vietnamese is one of the most prototypical VO languages.

The following section provides a detailed explanation of the calculation methods for these indices.

### 4.2.1 Simple Order Index (SOI)

#### 4.2.1.1 Shannon entropy

In this paper, Shannon entropy (ENTRO) is used to measure the average uncertainty of symbols (Shannon, 1948), Levshina (2019) introduced it in token-based typology to describe the variation of word orders. It is calculated as follows: first, the probability of each symbol appearing needs to be known. Then, the product of each symbol's probability and

its corresponding information content is calculated, and these products are summed. Finally, the negative of the sum is taken. The information content is calculated using logarithm with base 2, where the information content is equal to -log2(probability of the symbol appearing). Therefore, the formula for Shannon entropy is as follows, where H(X) represents the Shannon entropy of random variable X, P(x) represents the probability of random variable X taking the value x, and log2 denotes the logarithm with base 2:

$$H(X) = -\Sigma\,[\,P(x) * log2(P(x))\,] \quad (2)$$

Furthermore, since all the discussed syntactic order patterns in this paper are binary, there are only two possibilities: XY or YX. Therefore, Shannon entropy can be expressed using the following formula:

$$H = -[\,P(XY) * log2\,(P(XY)) + P(YX) * log2(P(YX))] \quad (3)$$

Shannon entropy is a function that initially increases and then decreases. Therefore, it has a maximum value. According to the properties of the entropy function, if the probability of the maximum entropy value is the same under any condition, for binary syntactic order patterns, the Shannon entropy reaches its maximum value of 1 only when P(XY) = P(YX). When P(XY) = 1 or P(YX) = 1, the Shannon entropy reaches its minimum value of 0. The full Entropy data of patterns of languages can be found in supplementary files.

4.2.1.1 Simple Order Index (SOI): process and result

To describe the text frequency and variability of simple word order preferences for VO-OV, the following calculation format is used:

$$SOI = \begin{cases} (-1) * \big(1 - ENTRO(OV)\big), & P(OV) > P(VO) \\ 1 * \big(1 - ENTRO(VO)\big), & P(VO) > P(OV) \end{cases} \quad (4)$$

The simple word order variation is determined by the multiplication of two factors:

(i) Dominant Word Order: If the corpus correspond to a higher value for Verb-Object (VO) compared to Object-Verb (OV), it takes a value of 1; otherwise, it takes a value of -1.

(ii) Scoring coefficient: The scoring coefficient is calculated as (1 minus the entropy value of OV/VO). The purer the occurrence is, the higher the coefficient; conversely, the coefficient is lower.

Below is an example of calculating the SOI for various syntactic constructions in Icelandic:

Table 8: A SOI illustration of Icelandic

| Features | OV/VO | Entropy | Dominance | Coefficient | SOI |
|---|---|---|---|---|---|
| **Obj/V** | 7792/**34590** | 0.688 | 1(VO) | 0.312 | 0.312 |
| **The SOI of Icelandic** | | | | | 0.312 |

The following table presents the performance of SOI across 15 language sample:

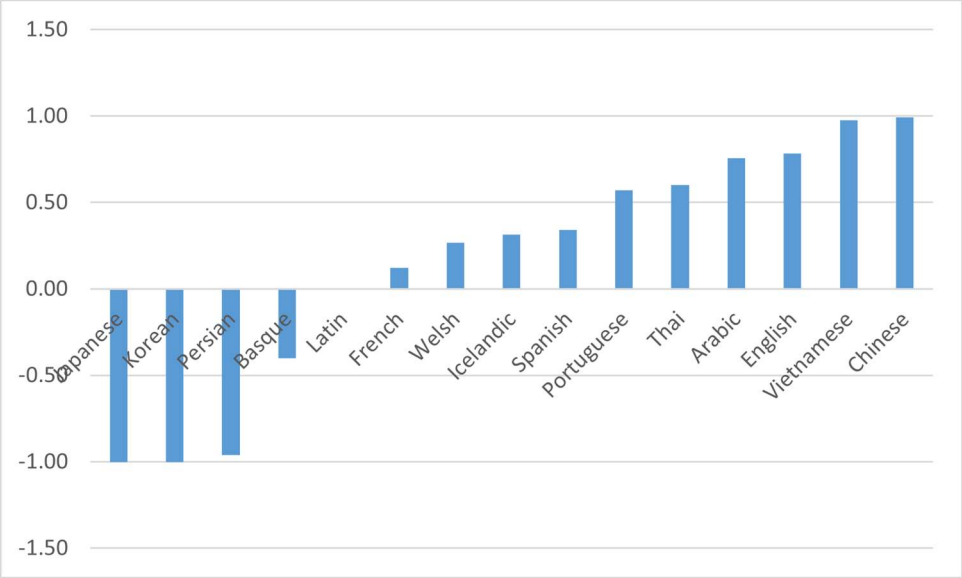


figure 3: SOI of 15 language samples

Figure 3 demonstrates that based on the simple text frequency, Japanese and Korean tend to have an OV (Object-Verb) order, while Vietnamese and Mandarin lean towards a VO (Verb-Object) order. Latin, German, and Czech exhibit a relatively flexible word order.

However, relying solely on SOI can lead us into biases toward surface word order. To avoid such biases, it is necessary to combine SOI with POI to conduct a two-dimensional measurement of word order and potentially uncover the true nature of word order. Section 4.2 further explores the relationship between SOI and POI to reveal the constraints between surface word order and the underlying word order feature.

### 4.2.2 Parameterized Order Index (POI): process and result

Parameterized Order Index（POI）describes the featured word order of world languages based on the six core parameters identified in Section 2. The calculation method for the POI, considering the dominance of each parameter, is as follows:

（5）

$$POI_{j/i} = \begin{cases} (-1) * RPR^{ov}\left(feature_j\right) * \left(1 - ENTRO\left(feature_j\right)\right), P\left(feature_j^{ov}\right) > P\left(feature_j^{vo}\right) \\ 1 * RPR^{vo}\left(feature_i\right) * \left(1 - ENTRO\left(feature_i\right)\right), P\left(feature_i^{vo}\right) > P\left(feature_i^{ov}\right) \end{cases}$$

$$POI = \sum POI_{j/i}$$

Below is a detailed introduction to the calculation methods of POI:

In other words, each word order feature has a POI, and the total POI for a language is the sum of all the word order features' POIs. For each POI, it is calculated by multiplying

three values:

(i) Dominant Word Order: If the feature corresponds to a higher value for Verb-Object (VO) compared to Object-Verb (OV), it takes a value of 1; otherwise, it takes a value of -1.

(ii) Corresponding RPR: If the dominant word order is 1, namely the order of OV, the RPR-VO is used; otherwise, the RPR-OV is used.

(iii) Score Coefficient: The score coefficient is calculated as (1 - entropy value of this feature). The purer the feature, the higher the coefficient; otherwise, it is lower.

The first two factors correspond to the retrorse predictive rate (RPR) of the word order features in this study, as all the work in this section uses participant predictions of word order. Therefore, the RPR is used for all the features, and in VO languages, RPR-VO is used, while in OV languages, RPR-OV is used. The third factor takes into account the issue of feature mixing. If a language's word order feature is ambiguous in frequency performance, its final POI will be low. It can only receive high index when the frequency is skewed sharply towards a particular word order.

Below is an illustration of calculating the POI for various syntactic constructions in Icelandic:

Table 9: A POI illustration of Icelandic

| Features | Feature $^{ov}/_{vo}$ | Entropy | Dominance | $RPR^{VO}$ | $RPR^{OV}$ | Coefficient | POI |
|---|---|---|---|---|---|---|---|
| **acl/N** | 14/**404** | 0.211 | 1(VO) | ***0.823*** | 0.640 | 0.788 | 0.649 |
| **advcl/V** | 3562/**8145** | 0.886 | 1(VO) | ***0.854*** | 0.593 | 0.114 | 0.094 |
| **aux/V** | 1327/**16432** | 0.383 | 1(VO) | ***0.82*** | 0.550 | 0.617 | 0.508 |
| **case/N** | 151/**35148** | 0.028 | 1(VO) | ***0.87*** | 0.708 | 0.972 | 0.800 |
| **cop/X** | 7808/**20310** | 0.852 | 1(VO) | ***0.90*** | 0.737 | 0.148 | 0.122 |
| **nmod/N** | **577**/75 | 0.515 | -1(OV) | 0.84 | ***0.625*** | 0.48 | -0.3144 |
| The POI of Icelandic (unnormalized/normalized) | | | | | | | 1.86(0.44) |

We can observe that in the individual POIs, nmod/N tends to favor OV, while the other options tend to favor VO. Therefore, the calculation for nmod/N's PRP selected RPR-OV and exhibited a negative POI, while the other features selected RPR-VO and exhibited a positive POI. In addition, the VO frequency and OV frequency of cop/X and advcl/V are relatively equal, which leads to a very high Entropy, resulting in lower scoring coefficients. Consequently, the absolute value of the POI is relatively small. Ultimately, we obtain a POI of *1.86* for Icelandic, and the normalized POI is 0.44, indicating that Icelandic exhibits a moderately OV word order, locating in the middle of world languages.

The following are the POIs for 15 language, which can inform a basic tendency of the POI of world languages:

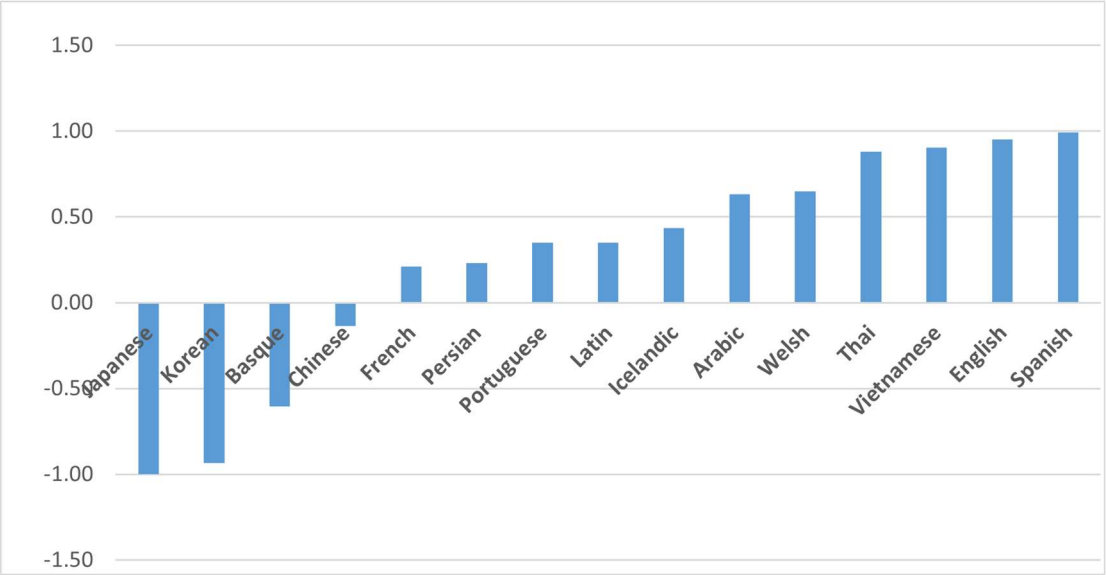


figure 4: POI of 15 language samples

In Figure 4, the 15 language samples show that Japanese and Korean have a strong preference for OV word order, which is as same as SOI implies. However, Spanish lean towards VO and Mandarin Chinese and Persian exhibit a mixed word order, which is different from the information Provided by SOI. Then the next step is to expore the interaction between SOI and POI, which is in section 4.3.

To Addition, since the POI only reflects the overall values of word order parameters for each language and does not provide a detailed analysis of the individual conditions of the six parameters. This is merely a macro-level description of the tendencies of world languages based on their featured word orders, and further research is needed for a more granular analysis at the typological level, which will be argued in section 5.

## 4.3 Cross-Linguistic harmony between Featured Word Order and Surface Word Order

Based on the measurement methods described in sections 3.1 and 3.2, we obtained two-dimensional word order measurements, namely, the SOI &POI (Normalized) for 78 languages.

Figure 5 displays the binary VO-OV word order information in two dimensions and presents a regression curve with a 95% confidence interval. The results reveal a significant linear correlation between the POI and SOI of world languages ($p < 0.01$, $R^2 = 0.430$). In other words, if a language tends towards an OV or VO word order based on featured factors, its surface word order also tends to lean towards OV or VO.

The horizontal axis represents the distribution of languages as OV or VO types based on simple frequency analysis from corpora. The vertical axis, ranging from bottom to top, reflects the degree of OV-VO measured by the POI, which integrates the reverse prediction rate (RPR) of word order features and the distribution of each feature within the language corpora, resulting in a word order index.

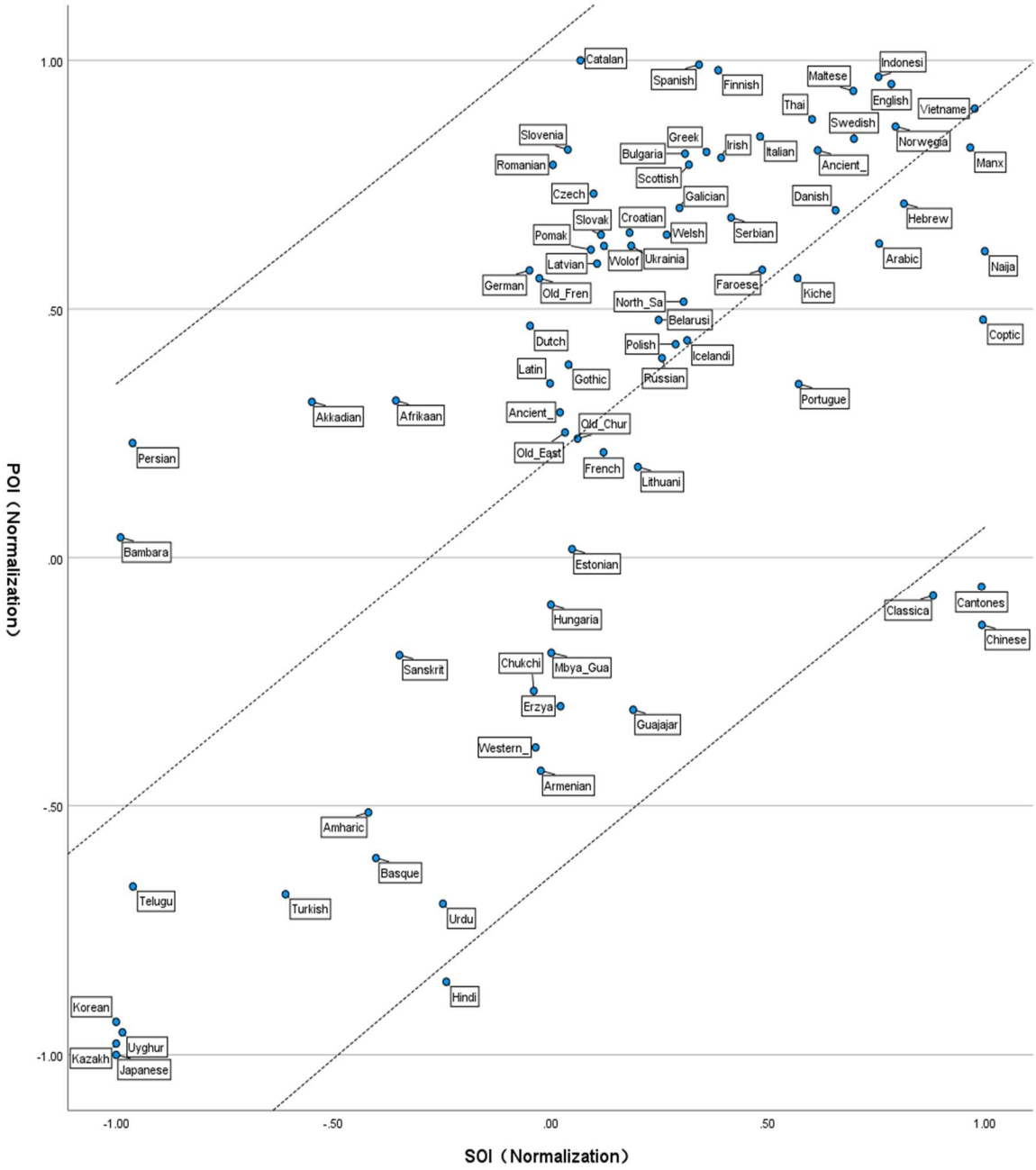


Figure 5: Word Order Measurements for 78 Languages

Figure 5 aligns with the fundamental conclusions of typological research. Looking at individual languages, for example, Japanese and Korean are prototypical OV languages, exhibiting high OV scores in both SOI and POI measurements. These languages are often referred to as strict subcategories of verb-final languages (Greenberg, 1963). English, Vietnamese, and Thai, on the other hand, are typical VO languages, positioned in the upper-right quadrant.

Furthermore, it is evident that the majority of languages roughly fall within the confidence interval of the linear correlation between POI and SOI. This implies that human languages do not permit strong violations of SOI and POI, meaning a strong divergence between

surface word order and word order features is not allowed. Namely, the greater the deviation from harmony, the less likely such a language is to exist.

## 4.3 The cross validation of Z-pattern and PCCH

The distribution of SOI and POI presented in Section 4.2 not only demonstrates the cross-linguistic harmony of world languages in terms of surface word order and syntactic features, but also bears significant resemblance in distribution to Gerdes et al.'s (2021) quantitative typology. Moreover, it offers evidence from a token-based perspective for Hawkins' (1983) proposed Principle of Cross-Category Harmony (PCCH), thereby to some extent reflecting the alignment of this study with other research endeavors and facilitating cross-validation.

### 4.3.1 SOI&POI follow a distribution of loose Z- Pattern

Gerdes et al (2021:23) examined several two-dimensional distributions related to word order implicational universal . In cases where world languages exhibit a distribution resembling a "Z" shape in the two-dimensional space, such distribution patterns are categorized as “z-patter. In Figure 6, on the left is a distribution of all dependent vs. complements of an adposition diagram, on the right is the z-pattern schma.

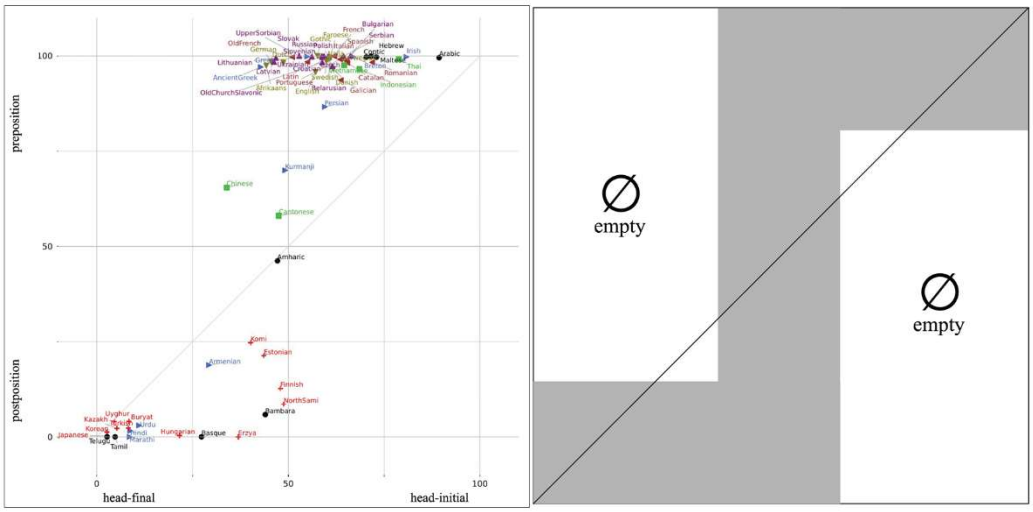


Figure 6： Head-final/initial-adposition the Z-pattern (Gerdes et al 2021:23, original tittle :Figure 24: All dependent vs. complements of an adposition diagram; Figure 25: The Z-pattern)

The interpretation of the Z-pattern can be conducted from the perspectives of both the X-coordinate and the Y-coordinate (Gerdes et al 2021:24). In this means, another explaniation of figure 5 can be argued as following:

From the X-coordinate perspective, when a language strongly tends toward a surface order of VO or OV, their order of features also exhibits a tendency toward VO or OV, or at least maintains a relatively neutral order of linguistic features. However, significant deviations between the surface order and the order of linguistic features are not allowed.

From the Y-coordinate perspective, when a language strongly leans towards a preferred order of either VO or OV, their order of linguistic features also demonstrates a tendency toward VO or OV, or at least applies a mixed surface order. Nevertheless, substantial

contradictions between the surface order and the order of linguistic features are not permissible.

However, a noteworthy phenomenon emerges. The distribution in this study is, in fact, even more scattered than the z-pattern in Figure 6. This is evident in the language distribution along the axis positioned at 0 comparing to the z-pattern inn figure 6. In the z-pattern, both symmetric "empty" positions imply the impossibility of certain language occurrences. In contrast, within the distribution of Figure 5, aside from the voids in the upper-left and lower-right corners, language is distributed across other regions. This paper argues that this distribution is more relaxed. Moreover, in comparison with the "fatter z-pattern" proposed by Gerdes et al. (2021), it appears to be more dispersed. Coined here as a "loose z-pattern," Figure 6 depicts the distribution of Figure 5 with language labels removed, showcasing an even more lenient Z-pattern. The voids in the distribution are more concentrated.

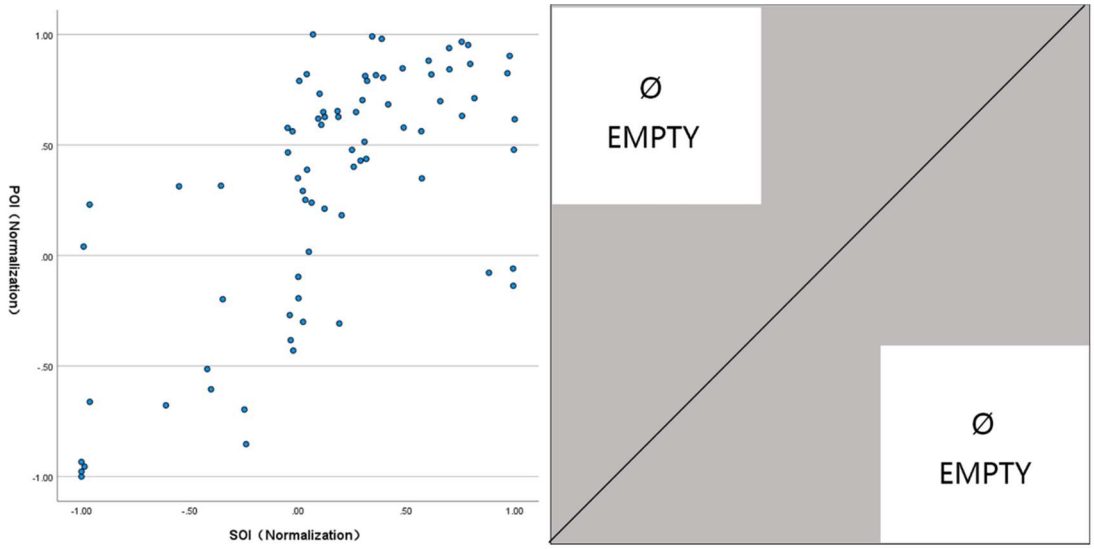


Figure 6: SOI-POI and the loose Z-pattern

This discrepancy in distributions might be attributed to two factors: firstly, Gerdes et al. (2021) did not solely focus on the "Object" dependency relation, but rather took all core-dependent syntactic orders as the horizontal axis, investigating their distributions concerning other orders (such as adpositions). In other words, their horizontal axis (the syntactic order of all dependency relation directions/structure' order) encompasses the vertical axis (the syntactic order of a specific dependency relation direction/structure's order), potentially leading to an augmented positive correlation between the two.

Secondly, Gerdes et al. (2021) primarily examined individual syntactic features, whereas this study combined multiple syntactic features for analysis. This approach might underscore the greater flexibility in human language when multiple features coexist. Such violations could arise from the accumulation of subtle violations over multiple instances. Regardless of how they accumulate, these deviations cannot surpass a certain threshold. The region defined by this threshold represents an area of significant violation of surface constraints, which is unlikely to accommodate language distribution.

### 4.3.2 SOI&POI as a token level validation of PCCH

This study also provides quantitative validation in support of Hawkins' (1983) Principle of Cross-Categorical Harmony (PCCH). Hawkins posits that the harmony between VO and OV can be understood through the lens of the PCCH, which incorporates distributed exceptions of HDT. The PCCH theory posits that as the sequence conflicts between operand items and operand objects of differing word orders increase, the number of example languages demonstrating harmony decreases. However, this principle is generally a type-level, that is, the typological without corpus evidence, but not token-level.

This study provides token-level evidence of PCCH. The VO-OV word order corresponds to the same tendency in the syntactic feature of VO-OV for the overall sentence constituents. The farther this tendency deviates from a harmonious pattern, the fewer languages exhibit such a tendency. A clearer perspective emerges through the examination of standardized residuals. Residuals represent the disparity between observed and predicted values. In the context of this study, residuals can be interpreted as indicating the extent of divergence between surface word order and featured word order.

Table 10: the Statistics of Standardized residuals

| Standardized residuals | Number of languages | Cumulative Number | Percentage | Cumulative percentage |
|---|---|---|---|---|
| ±0~0.5 | 33 | 33 | 42.31% | 42.31% |
| ±0.5~1 | 28 | 61 | 35.90% | 78.21% |
| ±1~1.5 | 10 | 71 | 12.82% | 91.03% |
| ±1.5~2 | 2 | 73 | 2.56% | 93.59% |
| ±2~2.5 | 1 | 74 | 1.28% | 94.87% |
| ±2.5~3 | 4 | 78 | 5.13% | 100.00% |

In the overarching trend, as the standardized residuals expand—indicating a gradual escalation in the breach of cross-category harmony—the count of languages steadily diminishes. Notably, over 90% of languages exhibit residuals measuring less than 1.5. However, within the segment encompassing the most substantial residuals, ranging from ±2.5 to 3, a conspicuous anomaly emerges, with a total of four languages falling into this category: Persian, (Mandarin) Chinese, ancient Chinese, and Cantonese.

Consequently, this validates the study's second inquiry: a harmonious correlation indeed exists between a language's surface word order and its featured word order. The attributes of word order wield a pivotal restrictive influence on surface word order. The degree of deviation from harmony inversely corresponds to the probability of the existence of such a language.

## 4.4 A Classification System of Featured and Surface Word Order

Based on the aforementioned observations, this study reveals a constraining relationship between surface word order and featured word order. The subsequent question is whether such a constraining relationship has an impact on the typological classification of word order. Under the two-dimensional classification of word order, the conventional categorization of languages into VO, OV, mixed, or free word order can no longer be

considered accurate from a typological perspective, especially a token-based one. For example, if we classify Mandarin Chinese as a VO language, it fails to explain its mixed word order. Conversely, if we include Mandarin Chinese under the category of mixed word order, it fails to accurately describe its strong tendency towards a surface basic word order of VO.

The fundamental contradiction lies in the fact that previous classifications of VO-OV word order were one-dimensional. Qualitative studies by Li & Thompson (1974, 1975) and Sun & Givon (1985:329), as well as the series of studies by Jin Lixin et al. and Liu (2010) that followed a quantitative approach, could only determine whether a language is VO or OV, or to what extent it leans towards VO or OV, without specifying whether this VO or OV is related to surface word order (textual frequency) or featured factors (grammatical configurations).

Therefore, based on the data and discussions presented above, this paper proposes the following two-dimensional classification model for word order:

Table 11: Qualitative Classification of Two-Dimensional Word Order

| | OV surface order | mixed surface order | VO surface order |
|---|---|---|---|
| VO featured order | VO featured、OV surface（**Empty**） | VO featured、mixed surface（Catalan） | VO featured、surface（Vietnamese） |
| mixed featured order | mixed featured、OV surface（Persian、Bambara） | mixed featured、mixed surface（Hungarian） | mixed featured、VO surface（Mandarin Chinese, Cantonese） |
| OV featured order | OV featured、OV surface（Japanese, Korean） | OV featured、mixed surface（Hindi） | OV featured、VO surface（**Empty**） |

It is worth noting that such patterns can be either qualitative (such as table 7 above) or quantitative (such as the z-pattern in Figure 6), the grey background of table 7 can be supposed a mark of “possible distribution”. From a qualitative perspective, we establish two dimensions: "featured word order" and "surface word order." Each dimension has three levels: "VO," "OV," and "mixed." Thus, the world's languages can be divided into nine logically possible subcategories. Similar to the discussion in Section 3, due to the interdependence between featured and surface word orders, there are hardly any languages in the world where the surface word order is opposite to the featured word order. The featured and surface word orders of a language are at least not noticeably conflicting.

Typical (S)OV languages, such as Korean and Japanese, are the most harmonious OV languages, where both the featured and surface word orders are OV. On the other hand, Vietnamese is the most harmonious VO language, with both the featured and surface word orders being VO. Of course, there are also languages where at least one of the featured and surface word orders is mixed. Mandarin Chinese is the most typical example of a

language with a "mixed featured word order and surface word order VO." From this perspective, the mixed-word-order hypothesis for Mandarin Chinese is a reasonable conclusion, but it lacks focus on the surface word order.

The next section analyzes several representative patterns of word order harmony, further exploring different featured word order configurations in different languages, especially in harmonious and weakly harmonious languages. This may help us further understand the interaction between surface word order and featured word order, as well as the internal configuration patterns of the word order featured.

# 5. Configurations of Word Order: Representative Examples

## 5.1 Configuration of Feathers of Strongly Harmonious Languages

We refer to languages with a high degree of consistency between surface word order and word order features as strongly harmonious languages. Strongly harmonious languages can be categorized into VO-strongly harmonious, OV-strongly harmonious, and mixed-strongly harmonious languages.

The Vietnamese language serves as a typical representative of VO-strongly harmonious languages. Vietnamese exhibits an extreme bias toward the VO word order, reaching 99.4% in terms of surface word order. In its word order features, there are five features that lean towards VO, namely aux/V, case/N, cop/X, acl/N, and nmod/N. There is one feature that leans towards OV, which is advcl/V. The specific information is shown in the table below (in table 12~16, the POI and SOI are all normalized):

Table 12: Vietnamese as a Representative of VO-Strongly Harmonious Language

| Features | acl/N | advcl/V | aux/V | case/N | cop/X | nmod/N |
|---|---|---|---|---|---|---|
| dominance | VO | OV | VO | VO | VO | VO |
| coefficient | 99.90% | 29.23% | 93.20% | 96.21% | 85.41% | 92.22% |
| POI (Per feature) | 0.82 | -0.19 | 0.77 | 0.79 | 0.70 | 0.76 |
| POI | 0.90 | | SOI | | 0.98 | |

From the perspective of specific score contributions, the contribution of each VO feature is very high, while the reverse contribution of OV features is relatively small. Among all VO features, Vietnamese obtained at least 80% of the scores. The low contribution of OV features may be due to the relatively small RPR weight of advcl/V in reverse testing OV languages, and the Vietnamese language does not exhibit an extreme OV tendency in advcl/V.

Japanese is the most representative OV strong harmony language. From the surface word order perspective, the frequency of OV in Japanese reaches 100%. From the perspective of word order features, all word order features in Japanese show an OV tendency, and the score coefficients of all word order features are very high, close to 100%. This makes Japanese an extremely OV strong harmony language.

Table 13: Japanese as a Representative of OV-Strongly Harmonious Language

| Features | acl/N | advcl/V | aux/V | case/N | cop/X | nmod/N |
|---|---|---|---|---|---|---|
| dominance | VO | OV | VO | VO | VO | VO |
| coefficient | 97.57% | 84.20% | 99.44% | 99.82% | 99.90% | 88.47% |
| POI (Per feature) | -0.63 | -0.54 | -0.64 | -0.65 | -0.65 | -0.57 |
| POI | -1.00 | | SOI | | -1.00 | |

The representative of mixed strong and harmonic languages is Hungarian. In terms of surface word order, Hungarian is a typical OV-VO language with nearly equal frequencies, resulting in an SOI close to 0. From the perspective of word order features, Hungarian exemplifies two basic patterns of mixed feature configurations: (1) neutral features and (2) feature mixture.

Neutral features refer to cases where the word order can be loosely classified as either VO or OV based on absolute frequencies, but due to their low coefficient scores, their overall word order scores tend toward 0. Examples of neutral features in the table below include advcl/V and aux/V.

Feature mixing refers to the phenomenon where two or more simultaneously strong and opposite word order features, although individually scoring relatively high, cancel each other out when calculating the result of the word order feature. For example, consider the two word order features advcl/V and case/N in the table below. While the former favors VO and the latter favors OV, overall, due to providing contradictory evidence, the word order feature tends to lean towards a mixed word order.

Table 14: Hungarian as a Representative of mixed-Strongly Harmonious Language

| Features | acl/N | advcl/V | aux/V | case/N | cop/X | nmod/N |
|---|---|---|---|---|---|---|
| dominance | VO | OV | VO | VO | VO | VO |
| coefficient | 99.90% | 5.05% | 7.06% | 98.34% | 15.86% | 43.00% |
| POI (Per feature) | 0.82 | 0.04 | -0.05 | -0.64 | -0.1 | -0.28 |
| POI | -0.10 | | SOI | | 0.00 | |

From the three examples above, we can observe some characteristics of strong harmonious languages. Non-mixed strong harmonious language configurations exhibit a basic bias towards a consistent word order, with each harmonious word order feature scoring sufficiently high, indicated by their frequency in the text. On the other hand, mixed strong harmonious languages demonstrate two types of feature configurations: either individual word order feature have lower scoring coefficients or multiple opposing word order feature with higher scores cancel each other out, resulting in the ultimate leaning towards a mixed word order.

## 5.2 Configuration of Feathers in Weakly Harmonious Languages

Weakly harmonious languages refer to languages where the surface word order and the manifestation of word order features are inconsistent. There are two typical examples of such languages: Persian, a weakly harmonious language with an underlying OV word order, and Mandarin Chinese, a weakly harmonious language with an underlying VO word order.

In Persian, the surface word order appears as an OV language with a frequency of 99%. However, when considering the word order features, Persian strongly leans towards a VO word order in the scores of the acl/N feature, while the cop/X feature strongly favors an OV word order. Other word order features predominantly exhibit a mixed state. These feature configurations result in a weakly harmonious between the word order features and the surface word order in Persian, making it one of the most representative examples of weakly harmonious languages.

Table 15: Persian as a Representative of OV, mixed featureLanguage

| Features | acl/N | advcl/V | aux/V | case/N | cop/X | nmod/N |
|---|---|---|---|---|---|---|
| dominance | VO | OV | VO | VO | VO | VO |
| coefficient | 99.90% | 0.02% | 0.01% | 25.49% | 98.69% | 81.16% |
| POI (Per feature) | 0.82 | 0.00 | 0.00 | 0.21 | -0.64 | 0.67 |
| POI | 0.23 | | SOI | | -0.96 | |

Another typical example is Mandarin Chinese. The surface word order of Mandarin Chinese is 99% VO. However, when analyzing the word order features, on one hand, the scoring coefficients of case/N and aux/V are relatively low. On the other hand, among the word order features with higher scoring coefficients, acl/N and cop/X lean towards VO, while nmod/N leans towards OV, thus neutralizing each other. This results in Mandarin Chinese having a mixed type of word order features.

Table 16: Mandarin Chinese as a Representative of VO, OV, mixed feature Language

| Features | acl/N | advcl/V | aux/V | case/N | cop/X | nmod/N |
|---|---|---|---|---|---|---|
| dominance | VO | OV | VO | VO | VO | VO |
| coefficient | 8.10% | 94.43% | 0.26% | 0.43% | 99.90% | 98.18% |
| POI (Per feature) | 0.07 | -0.61 | 0.00 | 0.00 | 0.82 | -0.64 |
| POI | -0.14 | | SOI | | 0.99 | |

Through the above case analyses, we further examine the constraints between surface word order and word order features, investigate the characteristics of harmony between feature and surface, and analyze the features of weakly harmonious languages. Particularly, we explore the internal situations of mixed word order features, identifying the presence of internal mixing within features as well as the neutralization patterns between features. In languages with mixed word order features, these two patterns often coexist.

# 6. Discussion: Word Order of Mandarin Chinese and a Preliminary explanation of Motivation

Based on the data analysis in the previous five sections, we have gained further understanding of the universality of word order in world languages. Firstly First, cross-

linguistically, the surface order and the order features are harmonious across languages, and the order measured by the order features and the surface order is largely consistent (strongly harmonious languages), or at least not in conflict (weakly harmonious languages). Secondly, in languages with mixed word order features, there are two typical manifestations: internal neutrality within features and neutralization between features.

Based on the above understanding, the most accurate summary of the typological status of Mandarin Chinese word order should be a typical surface VO word order, a weakly harmonious language with mixed featured word order. Specifically, Mandarin Chinese tends to be VO in terms of simple text frequency. However, it exhibits featured mixing: there are internally mixed neutral features such as aux/V and case/N, as well as mutual neutralization between features such as acl/N with a reverse VO tendency, cop/X, and nmod/N with a reverse OV tendency. This makes the word order features of Mandarin Chinese also mixed featured word order, as shown in the diagram below:

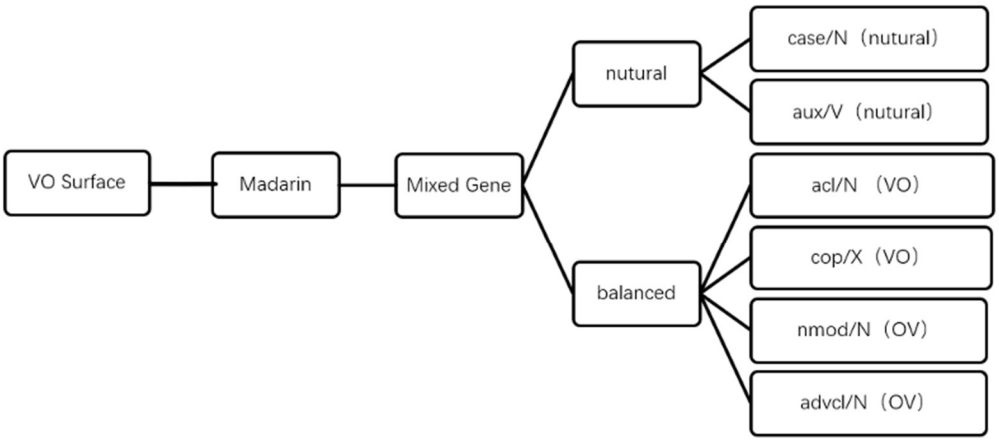


Figure 7: the configuration of Word order in Chinese Mandarin

The typological characteristics of Mandarin Chinese word order may be driven by the following factors:

Firstly, considering the strong VO surface word order, this may be attributed to the influence of prototypes and the encoding of transitive events in Mandarin Chinese basic sentence order. Psychological experiments provide evidence that when encoding prototypical transitive events, although human languages basically default to encoding them in the SOV word order, when the animacy of subject and object is similar, "role conflict" may arise, leading to the use of case marking or a shift towards SVO word order to distinguish the subject and object (Hall et al., 2013). This aligns with the traditional linguistic notion of case strategy and word order strategy (Siewierska & Bakker, 2008), namely, a language coding a prototypical event either by assigning case mark to NPs or use a AVP(AVP) word order. Mandarin Chinese does not have obligatory case marking and extensively uses SVO(AVP) word order to distinguish agent and patient relationships, which is consistent with the findings of relevant studies in psychology and linguistics. However, this tendency is weakened in intransitive structures, where OV syntactic configurations are more common.

Secondly, regarding the mixed featured aspect, this paper suggests that the "language contact hypothesis" proposed by Jin Lixin (2016) may be one possible explanation. Looking at the linguistic environments including Mandarin Chinese, Hungarian, and Persian, there is rich language contact. Mahootian (2018) explored that in Iran, there are 54-76 languages spoken within its borders, as well as bilingualism in multilingual environments. Kontra (2001) discovered from a sociolinguistic quantitative modeling perspective that more than a thousand years of language contact between Hungarian and Slavic languages has had a significant influence on Hungarian. These language contacts have resulted in a large number of loanwords and have also allowed the grammar systems to evolve featured word order features that accept two word orders, leading to the same VO-OV mixed word order as Persian, Hungarian, and Mandarin Chinese.

# 7. Concluding remarks

This paper quantitatively examines the syntactic structures related to VO and OV word orders using corpus-based methods, revealing a strong correlation between word order and featured factors in human languages. Based on this, an investigation is conducted on the typology of word order across world languages. The study uncovers the cross-linguistic harmony between word order features and surface word order, as well as the patterns of word order feature configurations. Building upon the measurement of word order in world languages, this paper defines the word order status of Mandarin Chinese as a weakly harmonious language with a surface VO order and a mixed featured profile. From a typological perspective, this word order type aligns with scholarly theories and empirical evidence regarding case configuration patterns, prototypical transitive clauses, and the hypothesis of language contact.

As a preliminary attempt to measure word order from two dimensions, this study has several limitations. First, it only focuses on the measurement method based on VO-OV typology, and does not introduce other word order parameters or grammatical structures, such as morphology, phonology, etc. Second, there is a lack of dialect and oral corpus, especially for Chinese, a language with complex word order variation both in geographical dialects (cf. Liu 2001a, Yiu 2014) and in oral grammar (cf. Gao 2008, Peck 2022), and the problem of the representativeness and comparability of text types (cf. Croft 2003: 112). are not excluded. Third, the measurement method starts from the harmony of VO-OV features but does not involve the comparison between other existing or potential models of language structures including word orders (e.g., Cysouw, Albu & Dress 2008, Dediu 2011, Parkvall 2008, Maslova 2002, 2004).

Furthermore, regarding the encoding motivations of the mixed-feature word order proposed in this article, namely, "prototype transitivity coding" and "language contact," further validation is required. Such validation necessitates additional means of data analysis, evidence from psychological experiments, or the establishment of new annotated corpora. Particularly in the examination of the "language contact" factor, if the existing annotated corpora can possess improved geographically-based continuity, in the most ideal scenario, if future annotated corpora could form a "dialect continuum" within a certain

range, and analyze it in a perspective of that dialects organized in a continuum without sharp boundaries (Heeringa 2001). This would provide significant assistance for verifying the hypothesized motive of "language contact leading to feature mixing" and any potential cross-linguistic commonalities and variations related to geographical factors.

In short, this paper proposes a two-dimensional word order measurement method, and measures the world languages from the two levels of surface word order and feature word order. It finds the cross-category harmony between surface word order and feature word order, proposes a new classification model for word order types, and points out that Chinese is a language with mixed features of surface VO. The research conclusion of this paper needs more microscopic and detailed investigation, as well as more proof or falsification of linguistic facts.

## *Supplementary Files*

This paper provides following Supplementary Files at OSF as following.
https://osf.io/eg83a/?view_only=352763e72a4c45bd905b9e5841bf5019

(1) 1-Corpus and size.pdf

This file provides a full list of the corpus used in this paper.

(2) 2-Frequency of word order patterns.csv

This file provides a frequency of word order patterns in this study, but all the word order patterns that occurs less than 10 have been excluded.

(3) 3-dominant (feature) order.csv

This file provides a judgment of dominant VO-OV order in languages or a dominant VO-OV matched order or word order patterns in languages with a threshold value of 50%, the OV or OV matched order is marked as “1”, and VO as”2”, and “3” stands for a non-dominance.

(4) 4-RPR&FPR in two theories.pdf

This file provides a list of RPR&FPR in two theories.

(5) 5-Entropy of word order patterns.csv

This file provides a full list of all the Entropy of word order patterns of languages.

(6) 6-POI&SOI.pdf

This file provides a unnormalized and normalized POI&SOI of world languages.

# References


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

Bickel, B. 2010. Capturing particulars and universals in clause linkage: A multivariate analysis. In I. Bril (Ed.), Clause-hierarchy and clause-linking: The syntax and pragmatics interface (pp. 51-101). Amsterdam: Benjamins. DOI:10.5167/uzh-48989

Bickel, B. 2015. Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. In B. Heine & H. Narrog (Eds.), The Oxford handbook of linguistic analysis (2nd ed., pp. 901-923). Oxford: Oxford University Press. DOI:10.5167/uzh-109110

Croft, W. 2003. Typology and universals (2nd ed.). Cambridge: Cambridge University Press.

Cysouw, M., Albu, M., & Dress, A. 2008. Analyzing feature consistency using dissimilarity matrices. STUF, 61, 263–279. https://doi.org/10.1080/10618600.2017.1305278

Dryer, M. S. 1992. The Greenbergian Word Order Correlations. Language, 68(1), 81-138. http://dx.doi.org/10.1353/lan.1992.0028

Dryer, M. S. 2009. The Branching Direction Theory of Word Order Correlations Revisited. In S. Scalise, E. Magni, & A. Bisetto (Eds.), Universals of Language Today (Vol. 76, pp. 185–207). Springer Netherlands. DOI: 10.1007/978-1-4020-8825-4_10

Dryer, M. S. 2013. Determining Dominant Word Order. In M. S. Dryer & M. Haspelmath (Eds.), WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533

Matthew S. Dryer. 2013. Order of Object and Verb.In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/83, Accessed on 2023-08-25.)

Matthew S. Dryer. 2013. Order of Adposition and Noun Phrase. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/85, Accessed on 2023-08-25.)

Mahootian, S. 2018. Language contact and multilingualism in Iran. In A. Sedighi & P. Shabani-Jadidi (Eds.), The Oxford Handbook of Persian Linguistics (online ed.). Oxford Academic. https://doi.org/10.1093/oxfordhb/9780198736745.013.14

Maslova, E. 2004. Dynamics of typological distributions and stability of language types. Voprosy Jazykoznanija, 5, 3–16.

Maslova, E., & Nikitina, T. 2008. Stochastic universals and dynamics of crosslinguistic distributions: The case of alignment types. Retrieved from http://www.anothersumma.net/Publications/ProbabilityPubl.html

Gao, Q. 2008. Word Order in Mandarin: Reading and Speaking. In M. Chan & H. Kang (Eds.), Proceedings of the 20th North American Conference on Chinese Linguistics 2 (pp. 611-626). Columbus, OH: The Ohio State University.

Gerdes, K., Kahane, S., & Chen, X. 2021. Typometrics: From implicational to quantitative universals in word order typology. Glossa: A Journal of General Linguistics, 6(1), 17.https://doi.org/10.5334/gjgl.764

Greenberg, J. H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), Universals of Language (pp. 73-113). Cambridge: MIT Press. https://doi.org/10.1515/9781503623217-005

Hashimoto, M. 1976a. Language Diffusion on the Asian Continent. Computational Analyses of Asian and African Languages, 3(4).

Hawkins, J. A. 1983. Word order universals. Academic Press.

Heeringa, W., & Nerbonne, J. 2001. Dialect areas and dialect continua. *Language Variation and Change, 13*(3), 375-400. doi:10.1017/S0954394501133041

Levshina, N. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. Linguistic Typology, 23(3), 533–572. https://doi.org/10.1515/lingty-2019-0025

Lehmann, W. P. 1973. A structural principle of language and its implications. Language, 49, 42-66. https://doi.org/10.2307/412102

Lehmann, W. P. 1978. The great underlying ground-plans. In W. P. Lehmann (Ed.), Syntactic typology (pp. 3-55). Austin: University of Texas Press.

Li, C. N., & Thompson, S. 1974. Historical change of word order: A case study in Chinese and its implications. In J. M. Anderson & C. Jones (Eds.), Historical linguistics (pp. 199-217). Amsterdam: North-Holland.

Li, C. N., & Thompson, S. 1975. The semantic function of word order in Chinese. In C. N. Li (Ed.), Word order and word order change (pp. 163-95). Austin: University of Texas Press.

Liu, D. 2001a. Hanyu fangyan de yuxu leixing bijiao [A typological comparison of word order among Chinese dialects]. Contemporary Research in Modern Chinese, 2, 24–38.

Liu, H. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. Lingua, 120(6), 1567–1578. https://doi.org/10.1016/j.lingua.2009.10.001

Li-xin, J. 2016. Typological Evidence of Mixed Word Orders in Mandarin and Its Motivation. Chinese Language Learning, (3), 3-11.

Shannon, C. E. 1948. A mathematical theory of communication. Bell System Technical Journal, 27(3), 379–423.

Siewierska, A., & Bakker, D. 2008. Case and alternative strategies: Word order and agreement marking. In A. Malchukov & A. Spencer (Eds.), The Oxford Handbook of Case (pp. 290-303). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199206476.013.0020

Sun, C.-F., & Givón, T. 1985. On the So-Called Sov Word Order in Mandarin Chinese: A Quantified Text Study and Its Implications. Language, 61(2), 329–351. https://doi.org/10.2307/414148

Parkvall, M. 2008. Which parts of language are the most stable? STUF, 61, 234-250. https://doi.org/10.1524/stuf.2008.0023

Peck, J. 2022. Semantic and Pragmatic Conditions on Word Order Variation in Chinese. In C. Huang, Y. Lin, I. Chen, & Y. Hsu (Eds.), The Cambridge Handbook of Chinese

Linguistics (pp. 444-466). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108329019.025

Vennemann, T. 1973. Explanation in syntax. In J. P. Kimball (Ed.), Syntax and semantics (Vol. 2, pp. 1–50). New York: Seminar Press.

Vennemann, T. 1974a. Analogy in generative grammar: The origin of word order. In Proceedings of the Eleventh International Congress of Linguists (1972) (pp. 79-83). Bologna.

Vennemann, T. 1974b. Theoretical word order studies: Results and problems. Papiere zur Linguistik, 7, 5-25.

Vennemann, T. 1976. Categorial grammar and the order of meaningful elements. In A. Juilland (Ed.), Linguistic studies offered to Joseph Greenberg on the occasion of his sixtieth birthday (pp. 615-634). Saratoga, CA: Anma Libri.

Vennemann, T., & Harlow, R. 1977. Categorial grammar and consistent basic VX serialization. Theoretical Linguistics, 4, 227-254.

Wadley, S. A. 1996. Altaic Influences on Beijing Dialect: The Manchu Case. Journal of the American Oriental Society, 116(1), 99–104. https://doi.org/10.2307/606376

Yan, J. & Liu, H. 2023. Basic word order typology revisited: a crosslinguistic quantitative study based on UD and WALS. *Linguistics Vanguard*. https://doi.org/10.1515/lingvan-2021-0001

Yiu, C. Y. 2014. Typology of Word Order in Chinese Dialects: Revisiting the Classification of Min. Language and Linguistics, 15(4), 539–573. https://doi.org/10.1177/1606822X1453205