A Corpus-driven Description of OV Order in Archaic Chinese

Qishen Wu[1] **Santiago Herrera**[1] **Pierre Magistry**[2] **Sylvain Kahane**[1]
[1]MoDyCo, Université Paris Nanterre
[2]ERTIM, INALCO, Paris
{qishen.wu, sherrera, skahane}@parisnanterre.fr, pierre.magistry@inalco.fr,

Abstract

This paper presents a quantitative study of Object-Verb (OV) order in Archaic Chinese based on a Universal Dependencies (UD) treebanks. Treating word order as a binary choice (OV vs VO), we train a sparse logistic-regression classifier that selects the most salient syntactic features needed for an accurate prediction to investigate the specific syntactic contexts allowing OV word order and to identify to what extent do these factors favor this order. The ranked features are understood as interpretable rules, and their coverage and precision as quantitative properties of each rule. The approach confirms earlier qualitative findings (e.g. pronoun object fronting and negation favor OV) and uncovers new contrasts in word order between different reflexive pronouns. It also identifies annotation errors that we corrected in the final analysis, illustrating how the quantitative models, combined with fine-grained corpus analysis, can improve treebank quality. Our study demonstrates that lightweight machine-learning techniques applied to an existing syntactic resource can reveal fine-grained patterns in historical word order and this can be reapplied to other languages.

1 Introduction

In this paper, we investigate Object-Verb (OV) word order in Archaic Chinese, following the historical linguistic periods of Chinese language proposed by Wang (1980). Wang defines Archaic Chinese as the historical stage extending from the Shang Dynasty (circa 1600 - 1046 BCE) and ending in the early Han Dynasty (206 BCE - 25 CE). This historical stage covers oracle bone inscriptions, bronze inscriptions, and classical texts from the pre-Qin period up to early Han period. Our study primarily focuses on representative texts from Eastern Zhou Dynasty (circa 771 - 256 BCE) to the early Han Dynasty. Our main goal is to systematically explore syntactic structures, particularly the conditions facoring or promoting OV order.

Most scholars currently agree that a stable SVO word order was already established by the time of Archaic Chinese. For instance Wang (1980) and Feng (2013) argued that during the Pre-Archaic period, the basic word order of Pre-Archaic Chinese was likely SOV, but as the language evolved, an SVO word order was established by the Archaic Chinese period. Ma (1898) highlight that object is placed after verbs in Archaic Chinese. More recently, Peyraube (1997) argue that Archaic Chinese was always a SVO language and OV word order is used only in specific syntactic contexts that are strictly constrained. Djamouri (2014) also states that the SVO word order had already become the dominant order in the Shang oracle bone inscriptions, based on a detailed study of 5,500 complete sentences from the Shang Dynasty(ca. 1600–1046 BCE), among which 94% followed the SVO order, while only 6% exhibited an SOV order.

Therefore, based on the research of these scholars, we can assume that Archaic Chinese is a stable SVO language. However, when exploring Archaic Chinese corpora, we find that although the OV word order is not dominant, it appears to be far from uncommon. In many syntactic contexts, OV word order is allowed and sometimes even preferred, which means that the VO order may remain productive in Archaic Chinese. In this regard, Yu (1981) highlight certain features of SOV language in Pre-Archaic Chinese, such as modifiers occurring before heads, which partially explain the possible origins of SOV word order in Archaic Chinese. Peyraube (1997) identifies four cases of OV word order in Archaic Chinese. Finally, Pan and Jiao (2023) conclude that OV order in Archaic Chinese can be roughly classified into unmarked object-fronting and marked object-fronting, after consulting previous studies on object-fronting in Archaic

Chinese by several different linguists. They also provide more detailed syntactic context in which OV word order can be used.

However, existing research remains predominantly qualitative in nature, mainly focusing on identifying potential syntactic environments and describing by hand the conditions that allow OV word order in Archaic Chinese. Besides, previous studies have suggested that there is no absolute grammatical rule enforcing OV order; instead, OV structures represent a possible syntactic choice within certain contexts. Therefore, given the observed variability in scientific statements about OV order, a quantitative exploration of syntactic tendencies becomes essential. Our aim is to undertake a more rigorous exploration, grounded in corpus data, to ascertain the specific manifestations of OV word order within Archaic Chinese corpora. In this paper, building upon previous theoretical research on OV word order, we aim to clarify the following research question: What specific syntactic contexts systematically permit or prefer OV order in Archaic Chinese and to what extent do these factors favor the OV order? To address this question, we adopt quantitative methods from computational linguistics, analyzing a syntactically annotated treebank to precisely identify and quantify the syntactic conditions that lead to object fronting.

2 Related Work

OV order in Archaic Chinese Former researches has showed that the use of OV word order is typically governed by strict syntactic contexts or specific semantic purposes. Most of these studies have adopted qualitative approaches to explore the occurrences and cases of OV word order in Archaic Chinese, summarizing and analyzing specific cases based on corpora and textual examples. As early as the end of the 19th century, Ma (1898) has noted that in Archaic Chinese, when a verb is preceded by a negation element or when the direct object is an interrogative pronoun, the corresponding direct object must be fronted. Since the 1980s, numerous scholars have analyzed the occurrences of the demonstrative pronoun *shì* 是(this) and the third-person pronoun *zhī* 之(3-person) in OV order when they appear between the direct object and the verb. For instance, Wang (1989) and Xiang (2010) argue that in OV word order, they function as resumptive pronouns referring to the object NP. In contrast, Han (1996) interprets *shì* 是(this) and *zhī* 之(3-person) as grammatical markers in OV constructions. Towards the end of the 20th century, Peyraube (1997) summarized the instances of OV word order in Archaic Chinese, identifying four main types and the following four examples are taken from Peyraube (1997):

1. The object is an interrogative pronoun;

   (1) 子 何 言
   zi hé yán
   you what say
   What do you say?

2. The object is the demonstrative pronoun *shì* 是(this);

   (2) 子 孫 是 保
   zi sūn shì bǎo
   son grandson this preserve
   The future generations (will) preserve this.

3. The object is a pronoun in negative sentences;

   (3) 不 吾 知 也
   bù wú zhī yě
   negation 1PRON understand final-part
   (You) don't know me.

4. The object is a noun phrase (NP) followed by a preverbal object marker *shì* 是(this) or *zhī* 之(this).

   (4) 四 方 是 維
   sì fāng shì wéi
   four region objet-marker unite
   (You should) unite the four region.

Very recently, Pan and Jiao (2023) categorize OV constructions into marked and unmarked types. Unmarked OV constructions include the fronting of WH-pronominal objects; in declarative sentences, the fronting of the demonstrative *shì* 是(this); in negative sentences, the fronting of demonstrative pronouns *zhī* 之(this), *shì* 是(this) and *cǐ* 此(this), personal pronouns, and noun phrases (NPs). In addition, marked OV constructions are further subdivided based on the type of marker into *wéi* 唯(only)-type, *shì* 是(this)/ *zhī* 之(this)-type, and

wéi 唯(only)… *shì* 是(this)/ *zhī* 之(this)-type constructions. For each type, the authors provide specific examples from the corpus and detailed analyses.

Grammar Study and Computational Linguistics In recent decades, the intersection of grammar study and computational linguistics has gained increasing attention, leveraging computational tools to deepen the understanding of linguistic structures and phenomena. This interdisciplinary approach has significantly transformed linguistic research. Vlachos and Craven (2010) parse biomedical text to extract features based on syntactic dependency relations, then they feed a sparse Bayesian logistic-regression model with extracted features to classify speculative language. This method improves the model's ability to recognise phrases that express uncertainty. The development of annotated corpora, such as dependency treebanks of Universal Dependencies (UD) frameworks (de Marneffe et al., 2014; Nivre et al., 2020; de Marneffe et al., 2021) which includes treebanks for more than 160 languages, have provided robust datasets for quantitative and comparative analyses. By analyzing data from the Universal Dependencies project, Gerdes et al. (2021) introduces a quantitative approach to linguistic typology, which move beyond traditional implicational universals to identify quantitative patterns in word order typology. This method allows for a more nuanced understanding of syntactic structures across languages, highlighting statistical tendencies rather than absolute rules. Levshina (2019), promoting a token-based typology, measures word-order variability with Shannon entropy calculated from Universal Dependencies data for about 60 languages. They find that languages fall into three clusters: high-entropy morphologically rich VO/flexible orders, mid-entropy analytic VO languages, and low-entropy OV languages. Chaudhary et al. (2022) developed a framework to assist linguists in the extraction of comprehensible syntactic rules, specifically focusing on morphological agreement, case marking, and word order. This system was validated across multiple languages, demonstrating its capability to generalize and apply linguistic rule extraction effectively in diverse language contexts. More recently, Herrera et al. (2024) introduces a novel method for inferring and mining, in a more exploratory design, detailed syntactic rules from treebanks: by employing sparse logistic regression enhanced with a richer feature search space, they effectively identify significant grammatical patterns, particularly for agreement and word order in Spanish, French, and Wolof, successfully uncovering both well-known and underexplored syntactic tendencies and rules.

3 Corpus and Method

In our study, we use Kyoto University's UD syntactic treebank of Classical Chinese in its version 2.15 (Yasuoka, 2019; Yasuoka et al., 2022). This corpus comprises an extensive dataset of 86,239 sentences. Our study focuses on the Archaic Chinese subset of the corpus, which includes three Confucian texts *"Lún yǔ"* 論語, *"Lǐ jì"* 禮記, *"Mèng zǐ"* 孟子, one classical poetic text *"Chǔ cí"* 楚辭 and one historical text *"Zhàn guó cè"* 戰國策. This sub-corpus consist of a total of 55,632 sentences, providing a rich resource for in-depth linguistic analysis.

Our aim, as stated above, is to identify quantitative and gradient rules or tendencies that favor OV word order in Archaic Chinese. For that, it is essential first to define what constitutes for us a quantitative grammar rule. Inspired by correspondence rules of the Meaning-Text Theory (Melcuk, 1988) and by Chaudhary et al.'s (2022) work, Herrera et al. (2024) formalise a grammatical rule with three elements or patterns: **the scope** S, which is the domain within which the specific grammatical phenomena under investigation may occur. In our case this consists of occurrences of verb with an object; **the target linguistic phenomenon** Q that has to be predicted, which is for us the object preceding the verb; and **the predictor pattern** P, in our study, the syntactic context that allows object fronting. Consequently, our aim is to investigate what are the syntactic contexts P that allow object fronting in Archaic Chinese, and to what extent ($\alpha\%$) of object fronting is likely to occur under each of these conditions.

$$S \implies (P \overset{\alpha\%}{\Longrightarrow} Q)$$

This formalization captures both the probabilistic nature of grammar and the overlapping relationships between grammar factors, making it highly adaptable to diverse linguistic frameworks and phenomena. By associating features with specific syntactic contexts, this approach offers a quantitative yet interpretable method for modelling grammar.

We then adopted the method described in Herrera et al. (2024) which use a linear classifier to extract the most salient features that predict the

linguistic phenomena. This method has been tested by using syntactic treebank corpora of English, French, Spanish and Wolof, demonstrating its applicability across different languages. This particular method was selected due to its inherent tendency to favor an exploratory approach.

More specifically, to identify the syntactic contexts that most strongly predict OV order, the authors use sparse logistic regression classifier to distinguish between OV and VO constructions based on the syntactic features extracted from the treebank. The feature space employed by the classifier is, as a matter of implementation, determined by manual specification. In our case, it consists for each node defined in the scope, i.e. the verb and the object, the following UD features: part-of-speech tags, dependency relations, morphological features (such as pronoun type), and clause-level modifiers (like presence of negation or sentence particles)[1]. We did not include lexical forms in the initial model to generalize across surface variation, though this remains a direction for future work.[2]

The classifier is tasked with estimating the probability of object fronting or not given a set of features. Once the model is trained, we examine the features that most heavily influence its decision-making process, specially the features corresponding to the syntactic conditions under which OV order is most likely to occur.

The authors use L1-regularisation to rank the most informative features for predicting syntactic phenomena. Specifically, they train the model for $k+1$ regularization strengths α_i, $0 \leq i \leq k$, which controls sparsity through the regularization strength parameter α. When α is large, only the most relevant features are retained; as α decreases, additional features gradually enter the model as their associated weights become non-zero. The k is set to 100 by default, with α_0 set to 0.01 and α_k set to 0.001. This built-in feature selection keeps only the most informative syntactic factors and suppresses noisy features. Because the surviving features and their weights are directly inspectable, the model is far more transparent than neural models and therefore well suited to recognising and interpreting the grammatical patterns that govern object fronting in Archaic Chinese. To test the statistical significance of each grammatical rule, the authors applied a G-test comparing the observed and expected distributions different features. The G statistic approximates a χ^2 distribution, and p-values were computed accordingly. Features with $p < 0.01$ were considered statistically significant, indicating that the observed association between the grammatical feature and the target linguistic phenomenon is unlikely to occurre by chance.

The authors also compute some statistical measures for each extracted pattern to understand its behaviour within the corpus. These measures are coverage and precision and are calculated as follows:

$$\text{Coverage} = \frac{\#(S \wedge P \wedge Q)}{\#(S \wedge Q)},$$

$$\text{Precision} = \frac{\#(S \wedge P \wedge Q)}{\#(S \wedge P)}$$

The **coverage** indicates among all OV occurrences, how likely the specific grammatical phenomenon occurs, whereas the **precision** measures among all occurrences exhibiting this selected feature, how likely they follow the OV order. For example, consider the grammatical feature "the object is an interrogative pronoun". If its coverage is 37%, this means that interrogative pronouns account for 37% of all OV occurrences of our corpus. If its precision is 73%, it indicates that among all instances where an interrogative pronoun serves as an object, 73% of these are fronted. High coverage suggests that the feature is common among OV constructions, whereas high precision implies that the feature strongly predicts OV word order.

4 Results

In this section, we analyze the syntactic factors of OV word order selected by the linear model. Among 55,632 sentences in the corpus, a total of 783 instances of OV word order were identified. The selected syntactic features significantly influencing OV word order are shown in the Table 1.[3]

The linear classifier identified nine grammatical factors that influence OV structures to varying

[1] features appearing fewer than 5 times in the corpus were excluded to reduce noise

[2] To extract more general features, we initially did not include orthographic forms in our analysis. However, based on the preliminary results, we may consider incorporating orthographic forms as an influencing feature in future research.

[3] We implemented our approach using the code provided by Herrera et al. (2024) in their paper. All results and data are included as supplementary material in paper submission portal.

	pattern P	occurrences of P	occurrences of Q	decision	coverage	precision	α
1	Object is a pronoun	5509	601	OV	76.8	10.9	0.012
2	Object is an interrogative pronoun	395	292	OV	37.3	73.9	0.007
3	Object is Third-person pronoun	3961	39	VO	10.7	99.0	0.005
4	Verb has an adverbial clause modifier	2183	257	OV	32.8	11.8	0.004
5	Verb has an expletive	210	129	OV	16.5	61.4	0.003
6	Object is a reflexive pronoun	209	89	OV	11.4	42.6	0.001
7	Verb has Degree "Equ"	801	91	OV	11.6	11.4	0.001
8	Verb has a sentence particle	3783	155	OV	19.8	4.1	0.001
9	Verb has a negative modifier	2707	83	OV	10.6	3.1	0.001

Table 1: Top features selected by the classifier favoring OV order.

degrees[4]. Based on the α values of each grammatical feature, we can observe that pronominal objects, as the most prominent feature, are identified first by the model. The next most significant feature is interrogative pronominal objects. The third significant grammatical feature—third-person pronominal objects—is somewhat special, as it is identified by the model as indicative of VO order[5]. The fourth feature selected is verbs with an adverbial clause modifier, while the fifth is verbs with an expletive. The final four grammatical features are distinguished at an α value of 0.001, with no clear difference in their level of relevance. Based on this ordering, we can already see that pronouns and pronoun-related features play a significant role in object fronting. We then describe the top nine factors in detail.

Firstly, the data indicates that the first and the most salient factor involves objects that are pronouns. The coverage for fronted pronouns is as high as 76.8%, highlighting a strong focus on pronouns in OV word order in Archaic Chinese. However, the precision for fronted pronouns is only 10.9%, indicating that only 10.9% of clauses with pronouns as objects exhibit fronting. This suggests that while pronoun fronting is a prominent feature of OV word order, OV structures themselves are not dominant in Archaic Chinese, as most verbs with pronoun objects do not exhibit this pattern. The second factor reveals that interrogative pronouns as fronted objects are particularly significant in the corpus. The coverage shows that interrogative pronouns account for 37.3% of all fronted objects, indicating their prominence. Since interrogative pronouns are a subset of pronouns, this finding can be seen as a refinement of the first rule, which shows that approximately half of the fronted pronouns are interrogative pronouns. Furthermore, the precision for this grammatical phenomenon is 73.9%, meaning that in cases where interrogative pronouns serve as objects, they are fronted in the vast majority of instances. This factor has been selected in the second place also highlighting a strong syntactic tendency for the fronting of interrogative pronouns in Archaic Chinese. These two factors are also consistent with Peyraube's (1997) and Pan and Jiao's (2023) research on object fronting in Archaic Chinese.

And the third rule reveals a grammatical situation where VO word order is clear favored, i.e. when the object is a third-person pronoun[6]. Coverage indicates that among all post-verbal objects, third-person pronouns account for 10.7%. However, this is an extremely precise rule: 99.0% of third-person pronouns used as objects are in fact postposed. At the same time, the extremely high precision of this rule and the fact that third-person pronouns used as fronted objects appear only 39 times in the corpus push us to consider whether these cases are very special or possibly due to annotation oversights or errors. Examining the occurrences, among the 3,961 occurrences of third-person pronouns used as objects, only two were *qí* 其(3-person), while the rest were *zhī* 之(3-person). Furthermore, all 39 fronted instances are *zhī* 之(3-person), each occurring as a fronted object in negative clauses, as shown in the example 5. This indicates that for the position of third-person pronoun objects, the decisive factor is actually another rule, namely, the influence of negation on object fronting. Moreover, in our corpus (which to some extent reflects features of Classical Chinese), the grammatical feature "third-person pronoun as object" can effectively serve as a criterion for VO word order.

[4]All nine factors are statistically significant, with p-values below 0.01 (detailed results are provided in the supplementary material).

[5]In this case, coverage and precision are calculated for $\neg Q$: Coverage $= \frac{\#(S \wedge P \wedge \neg Q)}{\#(S \wedge \neg Q)}$, Precision $= \frac{\#(S \wedge P \wedge \neg Q)}{\#(S \wedge P)}$

[6]coverage and precision are calculated for ¬Q

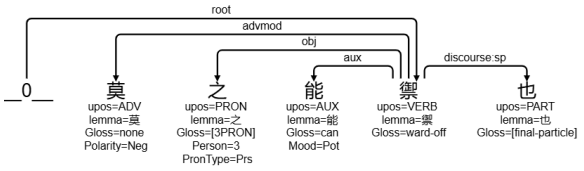

Figure 1: Fronting of *zhī* 之(3-person) in negative sentence.

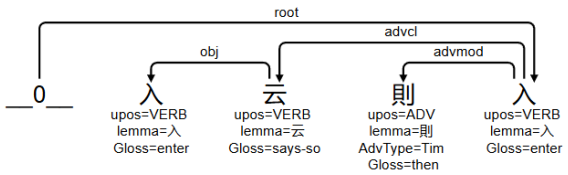

Figure 2: Fronting of object of *yún* 云(say).

(5) 莫 之 能 禦 也
mò zhī néng yù yě
none 3PRON can ward-off final-particle
None can resist it.

The next rule indicates that in adverbial clauses, the object is fronted in 11.8% of cases. Previous qualitative studies on Archaic Chinese have not specifically analyzed object fronting in adverbial clauses. Therefore, we conducted more detailed analyses of this grammatical context in our corpus. After ruling out several clear annotation mistakes, we found that there are 234 instances where the object is a pronoun, in which interrogative pronouns *hé* 何(what) accounts for 47.5% (122/257) and the demonstrative pronoun *shì* 是(this) accounts for 37.4% (96/257). Additionally, the demonstrative pronoun *shì* 是(this) in this context always appears in the fixed structure *shì yǐ* 是以(because of). Therefore, we can conclude that this rule may not seem to hold genuine grammatical significance, and it is other correlated grammatical features that actually constrain OV word order.

We then individually examined the remained 23 instances where the fronted object was a noun. We then discovered two instances of incorrect part-of-speech tagging for the interrogative pronoun *hé* 何(what) and one instances that the noun object is modified by the interrogative word *hé* 何(what) to form a wh-phrases *hé gù* 何故(why), as illustrated in the example 6. The remaining cases involve noun-object fronting in non-negative clauses and non wh-phrases, which drew our attention. We found that in the remaining 20 example clauses, *yǐ* 以(use) appeared as a verb 17 times, while *yún* 云(say) appeared three times. Among them, *yún* 云(say) occurred in a well-structured clause (in example 7), and perhaps the structure and rhythm of the clause prompt the fronting of the object. The case for *yǐ* 以(use) is different. First, *yǐ* 以(use) appeared multiple times as a verb in the examples covered by this rule, forming fixed expressions such as *hé yǐ* 何以(how) and *shì yǐ* 是以(because of). Additionally, in these instances, there was also the phenomenon of a noun object being fronted. This might suggest that when *yǐ* 以(use) is used as a verb in instrumental constructions with nouns, it may, to some extent, also encourage the fronting of its object. However, these instances are too rare in our corpus to draw a definitive conclusion. We thus offer only a tentative hypothesis here, and a more detailed discussion will require further targeted research in the future with more specific data.

(6) 墜 何 故 以 東 南 傾
dì hé gù yǐ dōng nán qīng
earth what reason use east south overturn
Why does the earth tilt to the southeast.

(7) 入 云 則 入， 坐 云 則 坐，
rù yún zé rù， zuò yún zé zuò，
enter say then enter，sit say then sit，
食 云 則 食
shí yún zé shí
food say then feed
when told him to come into his house, he came; when told him to be seated, he sat; when told him to eat, he ate.

The fifth rule indicates that when a verb is modified by an expletive, 42.6% of its object is fronted, which means the 42.6% object is fronted with a marker. By examining coverage, we can see that fronted objects carrying such markers are not very common among all instances of object fronting. However, the precision shows that when a verb does have an expletive modifier, the object is more likely to be fronted. In light of this grammatical situation, we also examined examples from our corpus. After ruling out some clear annotation errors, we found that the primary word functioning as an expletive to modify the verb is *zhī* 之(3-person). There are also a few instances of *shì* 是(this) (12 occurrences, 4 of which are OV) and *sī* 斯(this) (2 occurrences, both VO). However, when we analyze the instances of *zhī* 之(3-person) in the corpus, we discover some subtle annotation issues. Because *zhī* 之(3-person) is a very commonly used syntactic

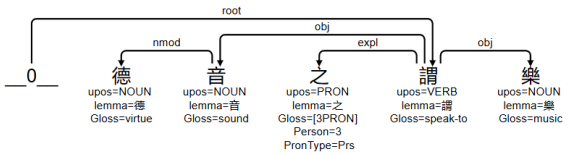

Figure 3: NP before *zhī* 之(3-person) considered as object of verb.

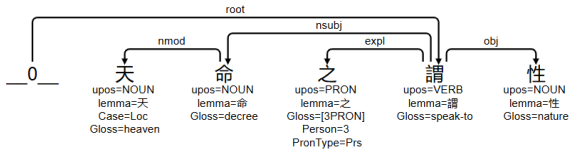

Figure 4: NP before *zhī* 之(3-person) considered as subject of verb.

word in Archaic Chinese with a wide range of grammatical functions, two occurrences containing *zhī* 之(3-person) that appear structurally similar can nevertheless be annotated quite differently in the corpus, as shown by the example 8 and 9. This illustrates that when performing quantitative analyses on an already annotated corpus, the choices made in annotation directly affect the analysis results. Therefore, although our quantitative methods are relatively swift and can straightforwardly provide coverage and precision for a consice set of grammatical phenomena within the corpus, we should still analyze specific corpus instances, as we have done above.

(8) 德 音 之 謂 樂
dé yīn zhī wèi yuè
virtue sound 3PRON use speak-to
The sound of benevolence is what makes true music.

(9) 天 命 之 謂 性
tiān mìng zhī wèi xìng
heaven decree 3PRON speak-to nature
A person's natural endowment is called "nature".

The sixth rule indicates that when the object is a reflexive pronoun, the object is fronted in 42.6% of cases. This rule does not have high coverage, but its precision is relatively high. Regarding the fronting of reflexive pronouns as objects, earlier qualitative research did not provide much discussion on this topic. Therefore, we conducted further research on reflexive pronouns in our corpus. In the corpus, there are two reflexive pronouns, *zì* 自(self) and *jǐ* 己(self), with the object *zì* 自(self) appearing 100 times and the object *jǐ* 己(self) appearing 109 times. However, these two reflexive pronouns differ significantly in their positioning when used as objects. After verifying the corpus and its annotations, we have found that all instances in which *zì* 自(self) was annotated as a post-verbal object turned out to be misannotations, all sharing the same structure: *shǐ/yǐ* 使/以(to make/ to use) + *zì* 自(self) + V. In this structure *zì* 自(self) was considered as an object of *shǐ/yǐ* 使/以(to make/ to use), but in fact *zì* 自(self) is the fronted object of the verb following it, and there are 15 such misannotations. Once corrected, *zì* 自(self) consistently appears as a fronted object in the corpus, regardless of whether the context is negative. By contrast, *jǐ* 己(self) as an object tends to be post-verbal, with only four instances of fronting, all of which occur in negative contexts. In comparison, when used as an object, *zì* 自(self) seems more like a reflexive pronoun specifically dedicated to fronting, whereas *jǐ* 己(self) does not exhibit this tendency.

The seventh rule is a more finely specified grammatical feature indicating the degree expressed by the verb, with "Equ" denoting "equal." In the corpus, the only verbs annotated in this manner are *rú* 如(be-like) and *ruò* 若(be-like). Moreover, upon excluding potential annotation errors and clause structures requiring further analysis, we find that the fronted objects under this rule are all interrogative pronouns. The eighth rule likewise shows a similar pattern. It states that when the verb is modified by a discourse element, specifically a sentence particle, 4.1% of object is fronted. However, both the coverage and precision of this rule are quite low. Upon reviewing the corpus data, we found that in those instances identified by the model, the genuinely decisive grammatical features for OV word order are actually interrogative-pronoun objects, demonstrative-pronoun objects, and object fronting in negative clauses. The classifier identifies this rule primarily because sentence particles occur very frequently in the corpus (nearly one-fifth of OV occurrences contain sentence particles) so the model treats this grammatical phenomenon as a distinct rule. Therefore, in connection with rules 4, 7, and 8, we have identified an aspect of the linear model that requires further refinement. Some grammatical factors identified by the model as influencing OV word order may actually be coincidental by products of large-scale data, and in reality, other grammatical factors are what truly

determine OV word order. We thus need to examine actual corpus instances to validate the model's results.

Finally, let us turn to the ninth rule, which indicates that when a verb is modified by a negative element, 3.1% of the object is fronted. This rule aligns with previous qualitative research, but since its coverage and precision are both low, we also need to analyze the corpus data. In our corpus, the negative elements that can modify a verb include *bù* 不(not), *mò* 莫(none), *wèi* 未(not-yet), *fú* 弗(not), *wú* 無(not-have), *fēi* 非(not), *wù* 勿(don't), and *wú* 毋(don't). The fronted objects under this rule include *zhī* 之(3-person), which is only fronted in negative clauses as mentioned in the third rule, the reflexive pronouns *zì* 自(self) and *jǐ* 己(self) (mentioned in the sixth rule), the first-person pronouns *wǒ* 我, *wú* 吾, *yǔ* 予, *yú* 余, and certain nouns or noun phrases some ending with *zhě* 者(the person/thing that ...). We can see that although this rule has low coverage and precision, it differs from the scenarios in rules 4, 7, and 8, as the instances in question are not entirely determined by other grammatical factors, and the negative element exerts a considerable influence on fronting the 3-person pronoun object *zhī* 之(3-person). The low coverage and precision of this rule may be due to the rarity of such object-fronting cases or because the dataset we used is not sufficiently comprehensive.

5 Conclusion

After carefully analyzing the model-generated results and validating them against the corpus, we conducted a detailed exploration of our research questions and obtained satisfactory outcomes: we have provided a corpus-driven quantitative analysis of OV word order, addressing the previously unexplored quantitative dimension of object fronting in Archaic Chinese.

First, we can observe that, the syntactic features automatically selected by the linear model align with the grammatical rules summarized by traditional linguistic studies: we see the importance of pronouns in object fronting in Archaic Chinese, especially interrogative pronouns, the tendency for object fronting in negative clauses, and the different roles of *zhī* 之(3-person) as various constituents in object-fronting constructions. Secondly, the quantitative generalizations from the regression model also helped us identify a new syntactic feature with significant influence on object fronting: the strong tendency for fronting when *"zì"* 自(self) when used as an object. Besides rather than merely identifying OV occurrences in Archaic Chinese, our results quantitatively demonstrate how common different object-fronting phenomena are, their productivity, and the strength of their tendencies.

Overall, this study demonstrates the significant potential of integrating quantitative computational methods into historical linguistic research, opening new avenues for systematic exploration of syntactic variation and grammatical structures in ancient languages.

6 Limitations and Future Research

However, our current approach also has certain limitations.

Firstly, corpus-based research is constrained by the limitations of the corpus itself. When we look closely at the specific content of the corpus, we still find issues arising from annotations: we discovered inconsistent annotations of identical sentence structures involving *zhī* 之(3-person), as well as part-of-speech tagging errors for *hé* 何(what). Moreover, due to the nature of the corpus texts, our current conclusions may be limited by the types of texts included in the corpus.

Secondly, recognizing that the phenomenon of object fronting may be influenced by multiple syntactic factors, such as the stronger tendency for pronoun objects to be fronted in negative clauses, we also attempted to select combinations of different grammatical features. Some meaningful syntactic feature combinations has also been distinguished, such as "object is a personal pronoun; verb has a negation modifier." Although the coverage of this combination was only 7.3%, its precision reached 41.3%, indicating that personal pronouns have a strong tendency to be fronted in negative clauses. However, in this case, we encountered many redundant grammatical combinations. For example, the model selected combinations such as "object is a pronoun; object is an interrogative pronoun" however, this rule essentially indicates that the object is an interrogative pronoun. Yet the model treats "object is an interrogative pronoun" and "object is a pronoun; object is an interrogative pronoun" as separate rules for comparison, which to some extent reduces the weight assigned to other potentially meaningful factors. This may be due to redundancy in the corpus annotation, or it may result from a lack of constraints when selecting features space.

Nevertheless, by combining the quantitative results of the linear classifier with detailed corpus analysis, we are also working toward improving the quality of corpus annotation. Through this integration, we have identified inconsistencies in syntactic annotation and found that attempts to extract combined grammatical rules reveal redundancy in the annotation information. Therefore, in future work, we will first use the current findings to optimize corpus annotation—correcting errors and inconsistencies—to improve the reliability of single-feature factor extraction. At the same time, we will attempt to reduce redundant information in order to support more accurate quantitative analysis of composite grammatical factors. We also note that OV constructions in Archaic Chinese are strongly influenced by lexical factors. Therefore, in future research, we plan to add orthographic features into our analysis.

Acknowledgments

This work was supported by the China Scholarship Council (CSC). This work is also supported by the Université Paris Nanterre. We gratefully acknowledge the contribution of Ms.Yuxin Zhang during the finalization of the camera-ready version. We also thanks beloved cat *Xiaoguai*, and guinea pigs *Coco* and *Chocolat*, whose quiet companionship provided much comfort throughout the writing process. Finally, We sincerely appreciates the anonymous reviewers for their valuable comments and suggestions, which helped improve the quality of this work.

References

Aditi Chaudhary, Zaid Sheikh, David R Mortensen, Antonios Anastasopoulos, and Graham Neubig. 2022. AUTOLEX: An Automatic Framework for Linguistic Exploration. *arXiv preprint*.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland. European Language Resources Association (ELRA).

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Redouane Djamouri. 2014. Duì shànggǔ hànyǔ fǒudìng jù lǐ dàicí bīnyǔ wèizhì de jìnyībù tǎolùn (對上古漢語否定句裏代詞賓語位置的進一步討論) [further discussion on positions of the object pronoun in negative sentences in archaic chinese]. *Lìshǐ Yǔyánxué Yánjiū [Research on Historical Linguistics]*, (02):47–57. [in Chinese].

Shengli Feng. 2013. *hàn yǔ yùn lù jù fǎ xué (*漢語韻律句法學*) [Prosodic syntax in Chinese (revised edition)]*. Shangwu Yinshuguan, Bei jing. [in Chinese].

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2021. Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa: a journal of general linguistics*, 6(1).

Xuezhong Han. 1996. Xiānqín fǒudìng jù zhōng "fǒu+dàibīn+dòng" jiégòu de yǔfǎ tèdiǎn (先秦否定句中"否+代賓+動"結构的語法特點) [grammatical properties of "negator + pronominal object + verb" construction in negative sentences in pre qin period]. *Běijīngdàxué xuébào(zhéxué shèhuì kēxué bǎn)[Journal of Peking University (Philosophy and Social Sciences)]*, (06):103–106. [in Chinese].

Santiago Herrera, Caio Corro, and Sylvain Kahane. 2024. Sparse logistic regression with high-order features for automatic grammar rule extraction from treebanks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15114–15125, Torino, Italia. ELRA and ICCL.

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

Jianzhong Ma. 1898. *ma shi wen tong(*馬氏文通*)[Ma's Grammar: The First Systematic Grammar of Chinese]*. Shangwu Yinshuguan, Bei jing. [in Chinese].

Igor A. Melcuk. 1988. *Dependency syntax: theory and practice*. SUNY series in linguistics. State University Press of New York, Albany.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Victor Junnan Pan and Yihe Jiao. 2023. Object-Fronting in Archaic Chinese. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Alain Peyraube. 1997. On word order in Archaïc Chinese. *Cahiers de linguistique - Asie orientale*, 26(1):3–20.

Andreas Vlachos and Mark Craven. 2010. Detecting speculative language using syntactic dependencies

and logistic regression. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, pages 18–25, Uppsala, Sweden. Association for Computational Linguistics.

Li Wang. 1980. *hàn yǔ shǐ gǎo (*漢語史稿*) [Lectures on the history of the Chinese language]*, 3 edition. Zhonghua Shuju, Bei jing. [in Chinese].

Li Wang. 1989. *Hànyǔ yǔfǎ shǐ (*漢語語法史*) [The history of Chinese grammar]*. Di 3 juan. Shangwu Yinshuguan, Bei jing. [in Chinese].

Xi Xiang. 2010. *Jiǎnmíng hànyǔ shǐ (*簡明漢語史*) [A Concise History of Chinese (Revised Edition)]*. Shangwu Yinshuguan, Bei jing. [in Chinese].

Koichi Yasuoka. 2019. Universal Dependencies Treebank of the Four Books in Classical Chinese. *DADH2019: 10th International Conference of Digital Archives and Digital Humanities*, pages 20–28.

Koichi Yasuoka, Christian Wittern, Tomohiko Morioka, Takumi Ikeda, Naoki Yamazaki, Yoshihiro Nikaido, Shingo Suzuki, Shigeki Moro, and Kazunori Fujita. 2022. Designing universal dependencies for classical chinese and its application. *Journal of Information Processing*, 63(2):355–363.

Min Yu. 1981. Dàojù tàn yuán (倒句探源) [tracing the object-predicate construction]. *Yǔyán yánjiū*, (00):78–82. [in Chinese].