Supervised Models on Observing Chinese Verb-Dependency Patterns

Junehyuk Yoo
University of Illinois at Urbana-Champaign
jy58@illinois.edu

Abstract

This paper investigates clause-level word order in Mandarin Chinese using supervised models trained on a Universal Dependencies (UD) treebank. Word orders such as SVO, SOV, SV, VO, OV, and V were modeled, comparing a simple majority baseline with spaCy's pretrained dependency parser. The models were tested against expert UD annotations to assess their ability to recover Mandarin's syntactic patterns. Results show that while the baseline captures the most frequent structure, spaCy offers marginal improvements, particularly on less common constructions. This paper will highlight both the challenges of automatic parsing in Mandarin and the potential of computational models to assist in corpus annotation of variable word order patterns.

1 Introduction

Mandarin Chinese is traditionally described as a language with a fixed subject-verb-object (SVO) word order. However, a number of alternative constructions, such as the *bǎ* (把) and *bèi* (被) constructions, systematically deviate from this pattern. Figure 1 shows an example of these constructions. These variations raise the question of how strictly SVO Mandarin actually is in usage, and whether computational models can detect these word order differences.

This paper investigates these issues using a Universal Dependencies treebank for traditional Mandarin Chinese. We compare a simple frequency-based baseline model with the pretrained spaCy dependency parser, focusing on their ability to identify and classify clausal word order patterns (SVO, SOV, VO, etc.). Our goal is to assess both the distribution of non-canonical structures in corpus data and the extent to which automatic parsers capture this gradient variation.

SVO
弟弟 罵了 小狗。
brother scold puppy
(The younger brother scolded the puppy.)

BA Construction (SOV)
弟弟 把 小狗 罵了。
brother BA puppy scold

BEI Construction, Long Passive (OSV)
小狗 被 弟弟 罵了
Puppy BEI brother scold
(The puppy was scolded by the younger brother.)

BEI Construction, Short Passive (OV)
小狗 被 罵了
Puppy Bei scolded
(The puppy was scolded.)

Figure 1: An example of a sentence written in canonical form and with BA and BEI (long and short passive) constructions.

2 Materials and Methods

2.1 UD Treebank Corpus

The corpus used in this study was taken from UD's treebank. The specific corpus is the Traditional Chinese Universal Dependencies Treebank (GSD treebank) annotated and converted by Google. Each file contains sentences in CoNLL-U format, with annotated parts of speech and dependency links for all tokens.

The Chinese UD treebank distinguishes a wide range of dependency relations, including core arguments (nsubj, obj), passive markers (nsubj:pass, aux:pass), and BA/BEI constructions (obl:agent, obl:patient). Additional relations such as temporal modifiers (obl:tmod), copular clauses (cop), and verb compounds (compound:vo, compound:vv) are also present, though the focus of the analysis is on verbal clauses with clear subject, verb, object relations [1].

[1] UD Treebanks: https://universaldependencies.org/zh/index.html

2.2 Gold-label Extraction Method

Gold labels for the word order were derived directly from UD annotations. For each sentence in the corpus, all verbs were identified, and their associated subjects and objects (if present) were extracted using the UD dependency annotations. Word order labels (e.g., SVO, SOV, VO) were then determined by comparing the indices of subject, verb, and object tokens within each clause. When a sentence contained multiple clauses, each clause was processed independently to yield its own label.

2.3 Baseline Frequency Model

A simple majority-frequency baseline was implemented as a basic comparison point. This model builds a frequency distribution over the possible word orders observed in the training corpus (e.g., SVO, SOV, VO, OV, V). At prediction time, the model outputs the most common order overall unless a heuristic override applies. Specifically, two heuristics were incorporated for constructions characteristic of Mandarin: (1) sentences containing the particle *bǎ* (把) are labeled as SOV, and (2) sentences containing the passive marker *bèi* (被) are labeled OVS if a subject is present, OV otherwise.

2.4 spaCy Model

The performance of the popular pretrained spaCy model (`zh_core_web_trf`[2]) was also evaluated. Unlike the baseline, which relies only on frequency counts and heuristics, the spaCy model is a transformer-based dependency parser trained on large-scale annotated corpora. It outputs POS tags and dependency relations that can be used to infer clause-level word order patterns.

A key challenge arises from the fact that spaCy's POS tagset and dependency labels are not fully aligned with UD's. For example, UD includes distinct categories such as `AUX`, `INTJ`, and `SYM`, whereas spaCy collapses these into broader categories. Although not necessarily relevant in this paper, it is something to keep account of. In cases where spaCy failed to identify a verb in a clause, the result was recorded as `N/A`, indicating that no valid wordorder label could be extracted.

This approach allows us to measure how well a widely used pretrained parser can capture less common Mandarin word order constructions (e.g., SOV in BA constructions, OVS in BEI passives), compared to a simple frequency-based baseline.

[2] spaCy models: https://spacy.io/models/zh

3 Results

Tables 1 and 2 detail the classification results for the baseline frequency model and the pretrained spaCy model, respectively. The baseline achieved an overall accuracy of 0.29, while spaCy reached 0.38. The baseline's simple heuristics allowed it to at least acknowledge the presence of BA/BEI constructions, though it rarely labeled their clause orderings correctly. In contrast, the spaCy model entirely failed to identify SOV, OV, and OVS patterns, suggesting difficulties in capturing long and short passive structures headed by BEI and BA respectively.

More broadly, the baseline defaulted to predicting majority classes such as "V," leading to high recall but at the cost of precision and near-total failure on other structures. SpaCy distributed its predictions across a wider range of classes and performed better on common patterns like SV, SVO, and VO, but it too showed poor generalization to low-frequency constructions. This discrepancy reflects both the class imbalance in the UD corpus and the challenges of aligning pretrained models with Mandarin-specific dependency conventions.

Label	**Precision**	**Recall**	**F1-score**	**Support**
OV	0.00	0.09	0.01	34
OVS	0.00	0.00	0.00	0
SOV	0.04	0.38	0.07	32
SV	0.00	0.00	0.00	3423
SVO	0.00	0.00	0.00	2861
V	0.33	0.86	0.48	4965
VO	0.00	0.00	0.00	3305
VS	0.00	0.00	0.00	3
Accuracy			0.29	
Macro Avg	0.05	0.17	0.07	14623
Weighted Avg	0.11	0.29	0.16	14623

Table 1: Baseline frequency model classification report on the UD Treebank corpus.

Label	**Precision**	**Recall**	**F1-score**	**Support**
N/A	0.00	0.00	0.00	0
OV	0.00	0.00	0.00	34
SOV	0.00	0.00	0.00	32
SV	0.43	0.35	0.39	3423
SVO	0.39	0.29	0.33	2861
V	0.45	0.47	0.46	4965
VO	0.40	0.35	0.38	3305
VS	0.00	0.00	0.00	3
Accuracy			0.38	
Macro Avg	0.21	0.18	0.19	14623
Weighted Avg	0.42	0.38	0.40	14623

Table 2: spaCy Model Classification Report on UD Treebank Corpus

4 Discussion

4.1 Limitations of the UD-GSD Corpus

Although the UD treebanks are designed with robust and consistent annotation standards, the UD-GSD corpus presents notable limitations for this study. In particular, word order patterns such as SOV, OVS, and OV are extremely sparsely represented. This class imbalance makes it difficult to meaningfully evaluate model performance on rare but significant constructions. As a result, overall accuracy is inflated by performance on frequent patterns (e.g., V, SV, SVO), while failures on minority classes remain hidden.

This limitation confines the conclusions that can be drawn about the ability of models to generalize to all of Mandarin word orders. Future research should therefore expand beyond UD-GSD, incorporating other Chinese UD treebanks or additional annotated corpora that contain more balanced distributions of constructions such as BA and BEI. Doing so would provide stronger evidence of how computational models handle non-canonical word orders and whether they can detect gradient variation across registers and contexts.

4.2 spaCy and Clausal Word-Order Patterns

The spaCy model did slightly outperform the baseline, but its performance on non-canonical word orders remained weak. One contributing factor is the mismatch between UD's annotation and spaCy's internal tagging conventions. For example, spaCy occasionally failed to return valid subject, verb, object dependencies under UD standards, which is reflected in the `N/A` label in Table 2.

Even when the subject, verb, and object relations were clear, spaCy often failed to recover them consistently. Even the most common word order pattern "V" was only correctly identified in roughly half of the cases. This is surprising given that spaCy's parsing pipeline incorporates transformer-based models that typically yield strong results in other domains. The under performance here suggests that pretrained parsers, while powerful in general-purpose applications, are not necessarily reliable tools for specific syntactic analysis in Mandarin without further problem-specific adaptations.

Future work could address this gap by testing other parsers explicitly trained on UD annotation (e.g., Stanza) which may better align with the labels/tags used in this study. Additionally, lightweight heuristic adaptation could improve recognition of passive BA/BEI and other relevant constructions. Such improvements would be essential for using computational tools to support corpus-based studies of gradient word order variation in Mandarin.

4.3 Implications for Mandarin and NLP

The baseline and spaCy model results show that automatic parsing is strongly biased towards the standard SVO pattern. This bias risks reinforcing oversimplified views of Mandarin syntax if researchers blindly rely on pretrained tools such as spaCy. Constructions that deviate from SVO (or SV, VO, or V), such as BA/BEI constructions, remain under-recognized despite their appearance in natural language use.

These findings show the importance of aligning computational tools with linguistic theory. If the goal is to study gradient word order variation, corpora must include richer distributions of non-canonical structures, and parsers must be adapted to reliably capture them. Otherwise, what appears to be a linguistic "fact" may instead be an artifact of parser bias or corpus imbalance.

Beyond linguistics, there are practical implications for other NLP applications in the pipeline such as machine translation, information extraction, and dialogue systems. Future work should therefore focus on task-specific retraining and the use of evaluation metrics that explicitly account for class imbalance. These steps are essential for ensuring that computational approaches can meaningfully contribute to syntactic corpus-based analyses of Mandarin.

4.4 A Future Experiment

While this study focused exclusively on the UD-GSD treebank, future work could investigate variation in word order patterns across different registers of Mandarin. Registers such as news, blogs, social media posts, or subtitles may exhibit more frequent use of uncommon structures like BA and BEI constructions. By comparing distributions of SVO, SOV, OV, and OVS patterns across registers, researchers could assess whether non-canonical patterns are register-specific or more broadly distributed.

To undertake such an experiment, additional UD treebanks and annotated corpora from diverse domains should be incorporated. However, as the present findings suggest, spaCy may not provide sufficiently accurate or UD compatible labels for this task. Instead, future experiments could employ dependency parsers that are directly trained on UD annotations, such as Stanza, which would provide more consistent treatment of Mandarin's syntactic relations. Using such tools would enable both a more accurate characterization of gradient word order and a better foundation for downstream corpus annotation or syntactic research.

5 Conclusion

This paper examined the ability of different supervised models to classify clausal word order patterns in Mandarin. A simple baseline frequency model was compared with the pretrained spaCy parser. While spaCy achieved slightly higher accuracy, its improvements were only marginal. This limitation stems partly from inconsistencies between UD annotation standards and spaCy's internal labels, but also from the broader challenge of applying pretrained parsers to fine-grained syntactic analysis. These findings suggest that third-party tools such as spaCy should be used with caution in corpus-based studies of Mandarin word order, and that more task-specific models or UD-aligned parsers may be necessary for reliable results. Future work could explore UD-trained parsers such as Stanza and evaluate word order variation across additional registers.

Acknowledgments

LING 413 Project 1. Could not have been possible without the support of Professor Brian Lin and Professor Jonathan Dunn.