A Feature-Based Approach to Annotate the Syntax of Ancient Chinese

Chenrong Zhao[1]
[1]CLL, Peking University

1. Introduction

This paper proposes a feature-based method for annotation that makes the evolution of functional categories and structures in different language systems comparable. Comparing syntactic features across different periods is a challenging issue, and the two main opinions have deficiencies:

- POS Comparison: The definitions of categories change over time.
- Feature Comparison (of MP) is unsuitable for isolating languages like Chinese.

The feature system mainly contains [±N], [±V], and [±IND] (individuation, that grammatically anchors words to the real world). It can represent and effectively differentiate typical instances across different stages.

2. Categories → Features

No matter how finely detailed categorical classifications are, they cannot meet the syntactic annotation needs across periods, as demonstrated by ASACC [2]. Our method simplifies the process by using only three primary features and a few additional ones, instead of a complex categorical system.

Based on research into syntactic evolution, we selected a test suite that captures key changes in syntactic features.

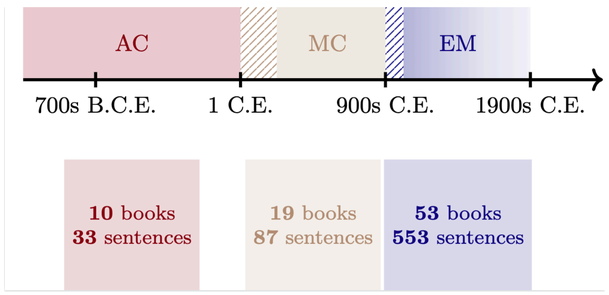


Figure 1: The illustration of the test suite with 673 sentences from 82 literary sources in total

Categorical features as more fundamental syntactic units:

- bypass the step of defining categories, allowing for a more detailed description of syntax across different periods.
- describe syntactic functions for flexible languages that lack clear categorical boundaries

Categories	Features			
	N	V	IND	VBLZ
CommonNoun	+	-	-	/
Copula	+	-	-	/
Pronoun, ProperNoun	+	-	+	/
NominalClassifier, Quantifier	+	-	+	/
Adjective	+	+	-	/
MotionVerb	-	+	-	/
Preposition	-	+	-	/
ModalVerb	-	+	+	/
VerbalClassifier, Disposal, Passive, Tense/Aspect, v(su$\check{o}$)	-	+	+	/
Sentence-finalParticle (SFP)	-	-	+	/
CoordinateMarker	-	-	-	/
Modifier-introducingParticle	+	±	-	/
VerbalisedActuation	+	+	-	+
VerbalisedConation	+	-	-	+
Nominaliser (zhě, zhī)	+	-	-	-

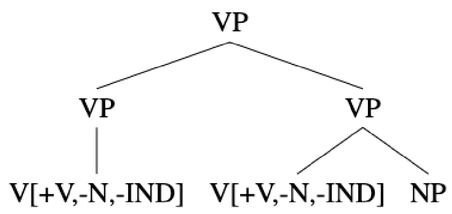
Table 1: Features of common categories in Chinese at different times

3. Annotation Examples

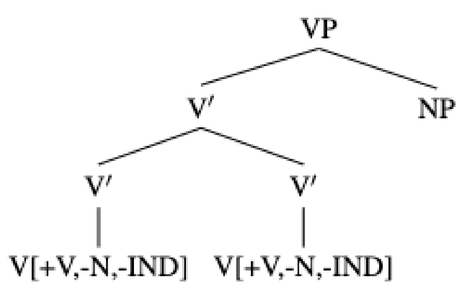
Syntax information is conveyed through both lexicon and structure, which can be represented by the three features. This approach aims to construct a large-scale diachronic treebank training set.

The development of new forms experiences three stages, take the development of complex predicates for an example:

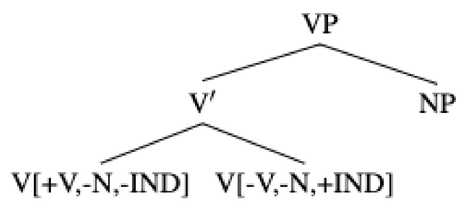
- Old Form:



- In-between:



- New Form:



(1) jī ér **shā** zhī.
hit coord kill 3pron
"Attact and kill him." OF

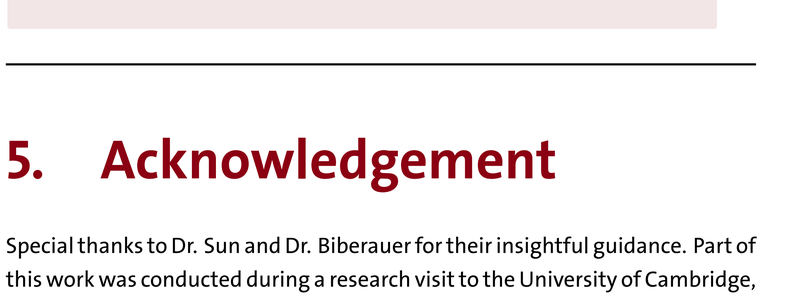
(2) dǎ **shā** qián jiā gē zi.
hit **kill** former household boy
"Beat the boy of former family to death." IB/NF

(3) é méi wù **shā** rén
pretty eyebrows impede largely people
"The pretty eyebrows (representing beauty) impeded her (entire life) to a large extent." NF

The second V in the NF is

- not an independent verb capable of taking patient arguments;
- attached to the first V, forming a compound-like construction that expresses actions with results.

Another example of the development of passive structure marked by "bèi":



4. Conclusions

- The test suite covers typical examples of Chinese syntactic evolution, which ensures effective annotation with limited data.
- The feature-based annotation aligns with the flexible nature of Chinese categories, minimizes bias and expresses syntactic as well as semantic information.
- The simplicity and comparability of our labels make them suitable for both Chinese and inflected languages.

5. Acknowledgement

Special thanks to Dr. Sun and Dr. Biberauer for their insightful guidance. Part of this work was conducted during a research visit to the University of Cambridge, which provided essential resources and support. We gratefully acknowledge the financial support provided by the China Scholarship Council (CSC).

References

[1] Hagit Borer. *Parametric Syntax: Case Studies in Semitic and Romance Languages*. Foris Publications, 1984.

[2] Academia Sinica; Research Group of Corpus Institute of Linguistics and Computational Linguistic Research Group. Academia Sinica Tagged Corpus. Academia Sinica Ancient Chinese Corpus, 1990. URL: `https://asac.iis.sinica.edu.tw/`.

[3] Carl Pollard and Ivan A Sag. *Information-based syntax and semantics: Vol. 1: fundamentals*. Center for the Study of Language and Information, 1988.

[4] Peiquan Wei. Donghan wei jin nanbeichao zai yufa shi shang de diwei [the position of the Eastern Han and Six Dynasties in the history of Chinese grammar]. *Chinese Studies*, 18:199–230, 12 2000.