

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

Chenrong Zhao

Department of Chinese Language and Literature / Peking University  
crzhao@pku.edu.cn

## Abstract

This paper is concerned with annotating the syntax of ancient Chinese, which is a series of languages in the same development process. The major challenge is to ensure the annotations of languages at different stages are comparable. To this end, we propose a feature-based approach that integrates the deductive feature design from the Chomskyan school and the inductive feature design from traditional philological studies. We demonstrate the effectiveness of our approach by annotating a collection of representative sentences that cover various linguistic phenomena that are extensively discussed in the literature. As a result, we establish a corpus of 673 (for now) ancient Chinese sentences paired with syntactic analyses, covering from 700s B.C.E. to 1900s C.E. The corpus can be utilised as a guideline for future large-scale TreeBanking.

## 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. A fundamental methodology for diachronic linguistics, the comparative method (Meillet, 1925; Hoenigswald, 1950, 1965; Harris and Campbell, 1995) that identifies and explains form-meaning pairs (i.e., phonological and semantic correspondences) mainly in phonology and morphology among languages from different places or eras, encounters challenges in the field of syntax at the very beginning step of establishing corresponding sets. Various attempts have been made to identify relatively fixed comparable components within the evolving and generative (and therefore infinite) set of sentences (Winter, 1984; Rankin, 2017). One approach considers categories as fundamental, but lexical categorization and the functions of categories vary across languages or in different language periods. Another influential approach is the

Parameter Comparative Method (PCM; Guardiano et al., 2016; Longobardi, 2014, 2017; Crisma et al., 2020) that splits the parts of speech (POS; Lyons, 1968) i.e., word categories into syntactic features. It offers a more nuanced comparative framework for understanding syntactic functions across diverse languages based on streamlined parameters in phylogenetic comparisons, but it falls short on languages without morphological markers, such as Chinese.

A distinct perspective on features is needed to address the deficiency. In addition to formal features encoded by morphology (Chomsky, 1995; Adger and Svenonius, 2011), the categorial features (Chomsky, 1970) encompass information of syntactic position, which serves as a crucial foundation for the syntax of languages without formal markers. Another feature, individuation (Bisang, 1999, 2002; Imai and Mazuka, 2003) characterises syntactic functional components, which may be integrated with lexical formally in these languages. Moreover, the particularities of syntactic representation in specific languages are deeply considered. Drawing on Chinese as an example, we annotate features for structures and functional components that underwent significant changes during the language’s evolution over two millennia. Our proposed features can represent and effectively differentiate typical instances across different stages.

As we further develop this annotation approach into a large-scale endeavour, we could model language systems across different eras and extract patterns in functional features, revealing deeper rules of language development beyond traditional studies focused on individual structures.

## 2 Feature Design

Our proposed features aim to rectify the common oversight of languages lacking morphological markers. This in turn facilitates the comparative

analysis of syntactic evolution across different periods for such languages.

## 2.1 From Categories to Features

The cornerstone of language syntactic modelling lies in annotated databases, where syntactic information such as POS tags, categories, and syntactic functions are marked. Drawing on theoretical research and linguistic practice, we propose that features can serve as the foundation for such annotations, striking a balance between inductive and deductive approaches while also accounting for the influence of both syntax and lexicon on grammar rules.

Annotation of POS is challenging for languages like Chinese due to their flexibility. According to (Rijkhoff and van Lier, 2013), FLEXIBLE LANGUAGES have word classes covering functions associated with multiple traditional categories (verbs, nouns, and adjectives). “Traditional word classes”, also known as “semantic categories”, as suggested by (Rauh, 2010), vary in distribution across languages. In actual analysis, “word classes” are often distinguished semantically, while syntactic classes focus on functions and positions. It is tricky to establish a satisfactory category system following the principle of syntactic categories due to the “flexible” nature of languages like Chinese, where words can appear in various positions without markers.

In response to challenges in categories, linguistic theories have shifted towards lexicalized approaches, seen in frameworks like LFG (Kaplan et al., 1981; Bresnan et al., 2015) and HPSG (Pollard and Sag, 1988, 1994). The Borer–Chomsky Conjecture (BCC; Borer, 1984) pursued the possibility that the syntactic functions of vocabulary are carried by the lexical themselves in the form of features. The feature system and its values not only avoid the problem of classification while describing the syntactic function of the lexicon but also present a more systematic picture of syntactic evolution by means of the temporal change of feature values. However, the interpretable and uninterpretable feature structure in the minimalist program associated with lexical entries (Baker, 2008) barely suits isolating languages like Chinese; neither do the features or parameters have been highly developed in historical linguistics like PCM, for the widely accepted feature system is based on inflectional languages, while Chinese lacks formal inflexion. Therefore, we propose a feature system that mainly contains  $[\pm N]$ ,  $[\pm V]$  and  $[\pm IND(ividuation)]$  that concerns

whether words can be anchored to the real world when used grammatically (represented by the feature of individuation). When features are correlated with functional components (Borer, 1984; Fukui, 1988), this method enables comparisons not only across different stages of the same language but also across different languages. This section will provide details on the feature system.

## 2.2 Features: $[\pm N]$ , $[\pm V]$ and $[\pm IND]$

Flexible languages—take Chinese, especially ancient Chinese, as an example—have rather vague boundaries between nouns, verbs, and adjectives. Here is a typical example<sup>1</sup>:

- (1) ěr (尔) yù (欲) Wú wáng (吴王) wǒ (我)  
 2PRON want king of Wú 1PRON  
 hū (乎) ?  
 Q?  
 “Do you want to make me be (like) the king of Wu?”

The categorical features  $[\pm N]$  and  $[\pm V]$ , proposed by (Chomsky, 1970), delineating categories based on feature restrictions, address the problem. Despite the lack of morphological inflexion for  $\phi$  features in Chinese, the relatively strict word order (Sun and Givón, 1985; Sun, 1996; Rijkhoff and van Lier, 2013; Van Lier et al., 2013) of Chinese sentences, typically following the SVO sequence, allows for determinations of  $[\pm N]$  and  $[\pm V]$ . For instance, in Example (1), 吴王 (*Wú wáng*) precedes the pronoun 我 (*wǒ*) indicating it was used as a verb<sup>2</sup>.

The combination of  $[\pm N]$ ,  $[\pm V]$  and  $[\pm F]$  (functional) distinguishes thematic categories from functional categories<sup>3</sup> (Chomsky, 1970; Grimshaw, 2000). However, in ancient Chinese, where functional categories were less developed, many func-

<sup>1</sup>In this sentence, 吴王 (*Wú wáng*) is a proper noun (refers to the certain king of 吴 (*Wú*) who was assassinated), yet could still take an object and express causative meaning. The shift of semantics (if there is one) without any morphological representation is periphrastic for containing a complex argument structure of “make (into)”.

<sup>2</sup>In our annotation, 吴王 (*Wú wáng*) is tagged as  $[+N]$ ,  $[+V]$ ,  $[-IND]$ ,  $[+VBLZ]$ , where  $[+N]$  and  $[+V]$  aligns with the adjectives in Chomsky (1970), while  $[+VBLZ]$  represents “verbalisation”, indicating the noun here means to make somebody be 吴王 (*Wú wáng*).  $[-VBLZ]$  is used to express the opposite of verbalisation – nominalisation. These usages are common in ancient Chinese.

<sup>3</sup>Every thematic category would exhibit the  $[-F]$  feature, while every functional category would exhibit the  $[+F]$  feature. Thematic categories include verbs, nouns, adjectives and prepositions. Functional categories include inflexions, determiners, degree adverbs, and complementizers.

tional features, like  $\phi$  features and TAM (tense, aspect, and mood), were carried by lexical items. The process of grammaticalization involves the emergence of these functional features as independent from lexical items, resulting in the formation of functional categories.

The mixed state of functional and thematic categories can be described from the point of view of INDIVIDUATION. Specifically, when a thematic category enters a sentence, it has to be individuated in some manner. Individuation (Imai and Mazuka, 2003) serves to anchor objects or events, indicating grammatical information regarding whether an object is identifiable and a motion is anchored<sup>4</sup>. While nouns are generally individuated by the determiner (or classifiers in certain languages) (Chierchia, 1998; Davidse, 2004; Zhang, 2013), the individuation of verbs includes all the components that wrap around the outside of the verb to make it legitimately used, like TAM, little *v*, etc.. We can draw the mapping relations of categories and the three features  $[\pm N]$ ,  $[\pm V]$ ,  $[\pm IND]$ , with an additional  $[\pm VBLZ]$  in table 1. The feature  $[\pm VBLZ]$  represents “verbalised actuation” that refers to the action of making the object to be a particular role or status, and “verbalised conation” refers to the action of regarding the object as treating someone.  $[-VBLZ]$  mainly denotes the nominalizers in ancient Chinese.

Categories	Features			
	N	V	IND	VBLZ
CommonNoun	+	-	-	/
Copula	+	-	-	/
Pronoun, ProperNoun	+	-	+	/
NominalClassifier, Quantifier	+	-	+	/
Adjective	+	+	-	/
MotionVerb	-	+	-	/
Preposition	-	+	-	/
ModalVerb	-	+	+	/
VerbalClassifier, Disposal, Passive, Tense/Aspect, <i>v</i> (所 <i>suǒ</i> )	-	+	+	/
Sentence-finalParticle (SFP)	-	-	+	/
CoordinateMarker	-	-	-	/
Modifier-introducingParticle	+	±	-	/
VerbalisedActuation	+	+	-	+
VerbalisedConation	+	-	-	+
Nominaliser (者 <i>zhě</i> , 之 <i>zhī</i> )	+	-	-	-

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

<sup>4</sup>The feature is effective, especially in languages lacking inflexion, but it could also be used on morphological languages, for it is based on cognitive theory (c.f. Langacker (1991))

### 2.3 $[\pm IND]$ : Case Study on Nominal Features

The  $[\pm IND]$  feature will be further illustrated through changes in the annotation of quantifiers within the domain of nouns.

In Mandarin Chinese, there are primarily two major categories of noun forms. One type requires a classifier when modified by a numeral, while the other functions directly as an argument without requiring a classifier. As shown in example (7) and (3).

- (2) sān (三) (běn (本) shū (书)  
three CL books  
“three books”
- (3) bān (搬) zhuō zi (桌子)  
move table  
“move the table (at the corner)/moving tables (is a simple job)”

In the nominal domain, the most significant change is the emergence of classifiers (Wei, 2000). Therefore, to depict the diachronic evolution of noun representations in Chinese, it is necessary to account for the function of classifiers. As is shown above, classifiers are obligatory when nouns are quantified by numbers, therefore they are numeral classifiers as Allan (1977) proposed and followed by others (Croft, 1994; Craig, 1994; Grinevald, 2000; Aikhenvald, 2000). Classifiers bear the features  $[\pm N]$ ,  $[-V]$ ,  $[\pm IND]$ , where the  $[\pm N]$  and  $[-V]$  are projected from N. In sentence (7), the noun takes the features  $[\pm N]$ ,  $[-V]$ ,  $[-IND]$ , while the combination of the  $[-IND]$  feature and the  $[\pm IND]$  feature of the classifier 本 (*běn*) results in the entire expression being marked as  $[\pm IND]$ , meaning that the noun 书 (*shū*) is instantiated of the concept of book, and is individuated from other books.

Following the principles of generative linguistics, determining the syntactic position of classifiers is necessary. Under the influence of the DP (Determiner Phrase) hypothesis (Abney, 1987), a prevailing view among scholars suggests a developmental tendency of classifiers in Chinese to adopt the functional role of the determiner (D), as evidenced by specific instances of classifier usage in certain dialects (Cheng and Sybesma, 1999, 2005; C-TJ et al., 2009; Gebhardt, 2011; Li, 2013). For example, the classifier 只 (*tsəʔ*) in sentence (4) (Li and Bisang, 2012) does not appear with numbers and indicates definiteness. There are also numerous opponents to this viewpoint, with the majority arguing that such usage is constrained by other factors.

For instance, [Wu and Bodomo \(2009\)](#) suggests that definiteness is not an inherent attribute of Chinese classifiers. The definite reading in examples like (4) is provided by the context.

- (4) tsəʔ (只) kiu (狗) ɕan kan (像看) san (生)  
 CL dog look-like have  
 mao biŋ (毛病) die (喋)  
 sickness PFV  
 “This dog looks ill.” Fuyang (Wu dialect)

Due to the relatively weak syntactic constraints on nominal expressions in Chinese, we argue that feature selection should be guided by communicative needs. The concept of communicative needs reminds us to prioritise aspects in Chinese speakers’ cognition that are more readily conceptualised, thus exhibiting greater universality and systematicity. From the functional perspective, individuation, the pragmatic function of numeral classifiers, is defined as “to establish a sensory perception as an individual by actualizing the inherent properties which constitute its conceptual unity” ([Bisang, 1999, 2002](#)), contrasting with “identification”. The concept of “identification” here, which differs from the function of DP (determiner phrase), does not explicitly treat an object as an individual. For instance, it’s conceivable to associate a sensory perception with the concept of, for example, an “apple” without explicitly delineating its inherent boundaries ([Bisang, 2002](#)).

The function of individuation, expressed through classifiers, emerged only after Middle Chinese. Before the appearance of individual classifiers, numerals directly modified nouns in Archaic Chinese, similar to English. According to [Huang \(1964\)](#); [Li \(2000\)](#) and others, it is commonly accepted that the evolution of classifiers progressed through four stages: from “noun + numeral” or “numeral + noun”<sup>5</sup>, to “noun + numeral + noun”, further evolving into “noun + numeral + classifier”, and ultimately forming structures like “numeral + classifier + noun”. The forms in the second and third stages serve as transitional forms, leading to the emergence of classifiers in Modern Chinese, where classifiers originated from nouns occupying the same position. As for the structure of “numeral + classifier + noun” phrases in Mandarin Chinese, some scholars propose that it stems from the reposition-

<sup>5</sup>In Archaic Chinese, both forms existed: the former was primarily used for counting, while the latter tended to convey predication ([Cheng, 2015](#)). These meanings were inherently inclined, as [Yao \(2008\)](#) pointed out.

ing of “numeral + classifier” from “noun + numeral + classifier” structures ([Wang, 1957](#)), as shown in (5), while others argue that after the formation of classifiers, “numeral + classifier + noun” structures directly replaced “numeral + noun” ([Yang, 1993](#); [Zhang, 2010](#)).

- (5) [*NumP* numeral classifier]<sub>i</sub> noun t<sub>i</sub>

To describe the noun forms in different periods of Chinese, it is crucial to address the question: did individuation exist in Archaic Chinese before the emergence of individual classifiers? If so, how was this function realised grammatically? If in Mandarin Chinese, common nouns behave akin to mass nouns in English, such as “water”, requiring individuation through classifiers to serve as arguments, then common nouns in Archaic Chinese exhibited characteristics of count nouns. With the emergence and development of classifiers, bare nouns and numeral-noun structures tended to convey generic reference ([He et al., 2011](#); [Krifka et al., 1995](#)) and individuated reference, respectively. In other words, before the emergence of classifiers, Chinese nouns lacked gender, number, and case markings, with generic and individuated references both conveyed by bare nouns without formal distinction. We argue that in Archaic Chinese, the grammaticalization level of the noun domain was low, and the individuation function was not yet fully isolated from count nouns. The distinction between the nominal domain and the verbal domain was also not clear. At this stage, it is challenging to assign specific word-class labels with external distinctiveness and internal consistency, let alone functional categories. This is why we propose the use of feature marking. In the feature system we proposed, nouns in numeral-noun and noun-numeral structures in Archaic Chinese possess the feature [+IND], inherited from count nouns. Before the emergence of individual classifiers, countable nouns, proper nouns, demonstrative pronouns, and personal pronouns all bore the [+IND] feature. However, with the development of quantifiers, this feature shifted to functional elements like classifiers and pronouns, while lexical elements such as common nouns and proper nouns became [-IND]. For instance, in Mandarin Chinese, classifiers are now obligatory before proper nouns, as illustrated in example (8).

- (6) sān[+N, -V, +NUM] (三)  
 three



- rén[+N, -V, +IND] (人)  
person  
“three persons”      Archaic Chinese
- (7) sān[+N, -V, +NUM] (三)  
three  
gè[+N, -V, +IND] (个)  
CL  
rén[+N, -V, -IND] (人)  
person  
“three persons”      Mandarin Chinese
- (8) yī (一) gè (个) yuè liang (月亮)  
one    CL    moon  
“one moon/the moon”

These features are deductively constructed to represent lexical and functional categories while inductively drawing from linguistic insights, thus covering a wider range of interconnected phenomena and a flexible, appropriate representation of syntactic information.

### 3 Test Suite

The three-feature annotation can be applied to a representative test suite that covers typical changing structures in Chinese based on Wei (2000), ensuring the feasibility of the annotation method with minimal annotation required.

Utilising prior research on classical Chinese syntax, we have created a test suite of 673 sentences spanning three major historical periods, annotated with syntactic structures and features from 82 literary sources. The research-oriented dataset spans from the East-Zhou dynasty (which began in the 700s B.C.E.) to the Qing dynasty (which ended in the 1900s C.E.), encompassing influential documents and Chinese classics widely referenced by scholars and native speakers. According to the periodization of Chinese historical syntax by Pan (1982), these materials can be categorised into three stages, as shown in Table 2.

Stage	#Book	#Sent
Archaic Chinese (700s B.C.E.–1 C.E.)	10	33
Middle Chinese (200s C.E.–900s C.E.)	19	87
Early Mandarin (1000s C.E.–1900s C.E.)	53	553
Total	82	673

Table 2: Distribution of test suite sentences.

These linguistic phenomena signify significant grammatical shifts in ancient Chinese over time, extensively explored in traditional research. Wei (2000) evaluates the frequency of certain typical phenomena at different times to determine the exact period when the reanalysis and analogy happened. The phenomena examined by Wei (2000) cover changes related to nominal and verbal domains. Nominal expressions include the development of suffixes, plural markers, 3rd personal pronouns, classifiers and quantifiers that modify N. Verbal expressions include the changes of actualisation, perfective marker, passive structure, disposal marker, etc.. The test suite encompasses crucial processes of reanalysis wherein these items (including lexical and functional) and structures undergo transformation.

In the nominal domain, aside from the grammaticalization of pronouns and classifiers, noun affixes in word formation (子 (*zi*), 儿 (*er*), 头 (*you*), etc.) and morphology (们 (*men*)) formed. Specifically, the test suite covers pronouns that have been largely simplified since Middle Chinese compared with those in Archaic Chinese (Wei, [1990] 2004), as well as the third-person pronouns that formed in Middle Chinese (Wang, 1945; Wei, [1990] 2004). The emergence of individual classifiers and plural affix 们 (*men*) are included in the test suite as well.

The aforementioned functional categories originate from the development of nouns, while other functional categories stem from the evolution of verbs. For instance, 了 (*le*) (perfective marker) evolves from verbs denoting completion, 着 (*zhe*) (durative marker) evolves from verbs indicating attachment, and 过 (*guo*) (past tense marker) derives from verbs conveying experiential meanings. Another typical change, complex predicates that express action results or depict the degree of a state, originates from the development of coordinated verbs. The special sentence structures, such as 把 (*bǎ*), 被 (*bèi*), and 比 (*bǐ*), all function as action verbs capable of taking objects in Archaic Chinese.

Furthermore, the grammaticalization of the particles 的/地/得 (*de*) indicating modification relationships, prepositions, and conjunctions has intensified.

These selected representative instances of structures have reflected the significant changes across different periods in the Chinese language system. This ensures the accuracy and professionalism of

manual annotations and enables the extraction of units that effectively characterise Chinese syntactic structures. In the following sections, we will present our data annotated by the features.

#### 4 Related Works

The previously annotated corpora included partial syntactic information that mainly relied on POS tags, which are not comparable for diachronic variations. Moreover, there is a lack of detailed syntactic relations among components across different periods.

Due to the extensive timespan of Chinese, it’s challenging to employ unified annotation rules, yet inconsistent rules hinder comparisons across different periods of the language system. For example, the biggest annotated corpus, *Academia Sinica tagged corpus* adopts the second strategy to pursue annotation accuracy, while the POS tags designed for early languages may not be suitable for languages that have changed in later periods. Moreover, the flexibility of word classes in Chinese (Rijkhoff and van Lier, 2013)—where nouns, verbs, and adjectives lack clear boundaries—poses challenges in establishing a theoretical basis for categorization when tagging POS. In contrast, features are more suitable than POS tags for comparing syntactic systems across languages, which had been successfully put into practice by the PCM. However, the PCM primarily focuses on verifying and quantifying phylogenetic relations among languages, and the features proposed by the PCM are not precise enough to describe the historical evolution in languages outside the Indo-European language family.

The scarcity of functional markers in Chinese, particularly evident in ancient times compared to languages with rich morphological markings, further increases the difficulty of syntactic analysis and annotation. For inflected languages, lexical categories and functional information can be validated through morphological markings, making it easier to correspond to diachronic changes in morphological markers. For instance, syntactic information in Middle Portuguese (Rocio et al., 2003)<sup>6</sup> is mapped based on Modern Portuguese.

While addressing the above issues with the three-feature annotation, we also annotate syntactic structures, supplementing syntactic information beyond

<sup>6</sup>MPPT (Mediaeval Portuguese Partial Treebank) uses the tagging resources of modern Portuguese as part of the training materials for automatic tagging of mediaeval Portuguese.

lexicons. In languages like English with mature annotation standards such as Taylor et al. (2003), historical data annotated with syntactic treebanks have seen significant development, for instance, the Penn-Helsinki Parsed Corpus of Middle English (PPCME; Taylor and Kroch, 1994). However, there is still a lack of comprehensive development in non-inflectional languages like Chinese.

An annotated corpus that facilitates comparisons of syntactic changes across different periods and reflects the development process of functional categories is required for depicting diachronic syntax.

#### 5 Annotation examples

Syntax information is conveyed through both lexicon and structure. In this regard, we are employing the three-feature standard to annotate vocabulary while referencing the phrase structure grammar (Gazdar, 1985) to annotate syntactic trees. This approach represents our endeavour to further construct a large-scale diachronic treebank training set. We will illustrate our annotation method through the development of resultative predicates and passive constructions, demonstrating that the data collected in the test set encompasses typical instances of structural changes in Chinese across various periods.

One of the most noticeable structural changes in VP is the development of complex predicates. The structure contains a verb and a postverbal constituent that modifies the verb, expressing the result, manner, or degree. The structure of VP changed from figure 1 and 2 (Archaic Chinese) to figure 3 (Middle Chinese<sup>7</sup>).

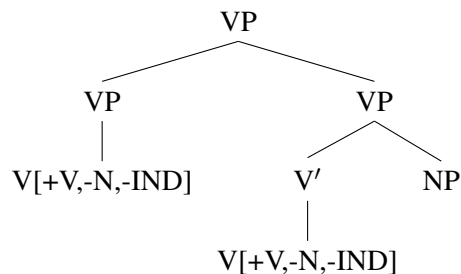


Figure 1: Coordination of VPs

<sup>7</sup>The period depicted in the illustration signifies the theoretical emergence of resultative complements, a topic that has sparked debate among scholars regarding its specific historical emergence. In this context, we refer to the perspectives of scholars such as (Wang, 1957; Ota, [1958] 1987; Mei, 1991; Wei, 2000).

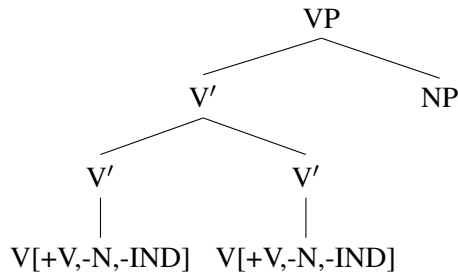


Figure 2: Coordination of V's

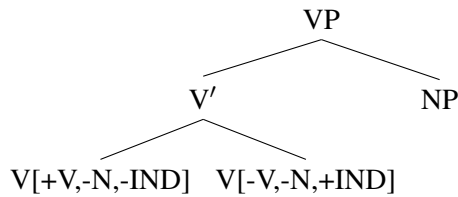


Figure 3: Complex predicate structure (Coordination of Vs)

The patterns demonstrate the development of complex predicate structures. During this process, the ability of the second verb to take a complement gradually diminishes until it merges with the first verb in the third stage. For instance, the *shā* in (9) and (10) is an action verb indicating killing, while that in (11) is a degree adverb representing the degree of the predicate. This process involves the merging of two VPs, i.e. two predicates. When two verbs are adjacent, the second verb gradually loses its function as a verb, therefore, the feature becomes [-V]. Adjacent to the first verb in the linear sequence, the second verb undergoes structural reanalysis, merging into a single VP with the first verb. Consequently, the original second verb expresses either the degree of the first verb or the resulting status, indicating a perfective predicate with a positive value for the [IND] feature.

- (9) *jī* (击) *ér* (而) *shā* (杀) *zhī* (之).  
hit COORD kill 3PRON  
“Attack and kill him.” Archaic Chinese

- (10) *dǎ* (打) *shā* (杀) *qián* (前) *jiā* (家)  
hit kill former household  
*gē zi* (哥子).  
boy  
“Beat the boy of former family to death.”  
Middle Chinese

- (11) *é méi* (娥眉) *wù* (误) *shā* (杀) *rén* (人)  
pretty eyebrows impede largely people  
“The pretty eyebrows (representing beauty)  
impeded her (entire life) to a large extent.”

Generally, in the process of linguistic evolution, new linguistic forms must coexist with old forms for a certain period until they are widely accepted by the linguistic community, thus replacing the older forms. The study of the syntactic evolution of ancient Chinese is particularly concerned with this transition from the old to the new. Take the example of passive sentences and their structural analyses, *bèi* in sentence (12) is an action verb expressing to receive. In contrast, *bèi* has lost its lexical meaning and introduces the agent and indicates the passive voice in typical passive constructions like example 4. The feature [+IND] is attributed to *bèi* due to the passive voice, which also shows that the action is completed. The usage of *bèi* presented in example (14) further demonstrates its grammaticalization. Some scholars argue that this usage implies a passive subject (Jiang, 1994). In this construction, the predicate can take an object, distinguishing it from the typical passive usage described in example 4. Specifically, the subject of the expression *nà rén* (“that person”, mentioned in preceding texts) undergoes an event involving the action by the agent and is directly affected by this associated event. The structures are illustrated in 4 and 5.

- (12) *yòu* (幼) *bèi* (被) *cí* (慈) *mǔ* (母)  
child receive beloved mother  
*sān* (三) *qiān* (迁) *zhī* (之) *jiào* (教)  
three move NMLZ education  
“When I was a child, (I) received the education from (my) beloved mother by moving three times.” Archaic Chinese

- (13) *lǎo* (老) *sēng* (僧) *bèi* (被) *rǔ* (汝) *qí* (骑)  
old monk PASS 2PRON ride  
“(I) Old monk, was ride by you.” Middle Chinese

- (14) *bèi* (被) *Wūsōng* (武松) *bù* (不)  
PREP PN NEG  
*guǎn* (管) *tā* (他)  
take-care 3PRON  
“Wūsōng does not take care of him.”  
被武松不管他。 Early Mandarin

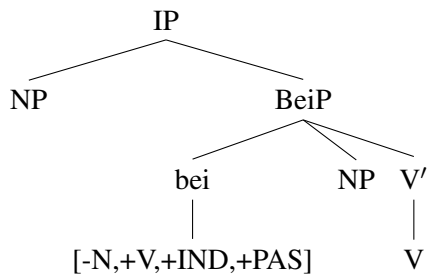


Figure 4: Passive constructions: “bèi” as a passive marker

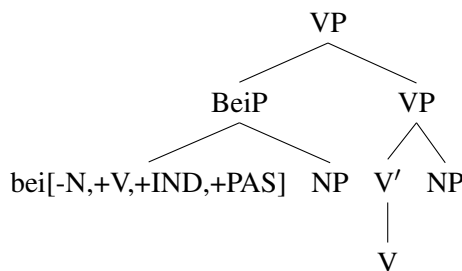


Figure 5: Passive constructions: “bèi” as a preposition

## 6 Conclusion

The feature annotation system is based on intuitions from a philological study of ancient Chinese syntax evolution, as well as features from both formal and functional grammar. The materials annotated are typical examples representing the process of Chinese syntactic evolution. This approach ensures annotation effectiveness and feasibility when data volume is limited. Our proposed three-feature system aligns well with the flexible characteristics of Chinese parts of speech, minimising researcher bias while expressing all known syntactic and semantic information. The simplicity and cross-temporal, cross-linguistic comparability of our feature labels make them suitable for languages like Chinese lacking morphological markers and adaptable for inflected languages as well.

## References

Steven P Abney. 1987. *The English noun phrase in its sentential aspect*. Ph.D. thesis, Massachusetts Institute of Technology.

David Adger and Peter Svenonius. 2011. Features in minimalist syntax. *The Oxford handbook of linguistic minimalism*, 1:27–51.

Alexandra Y Aikhenvald. 2000. *Classifiers: A typology of noun categorization devices*. OUP Oxford.

Keith Allan. 1977. Classifiers. *Language*, 53(2):285–311.

Mark C Baker. 2008. The macroparameter in a microparametric world. In Theresa Biberauer, editor, *The limits of syntactic variation*. Amsterdam: John Benjamins Pub.

Walter Bisang. 1999. Classifiers in east and southeast asian languages: Counting and beyond. *TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS*, 118:113–186.

Walter Bisang. 2002. Classification and the evolution of grammatical structures: a universal perspective. *STUF-Language Typology and Universals*, 55(3):289–308.

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

Joan Bresnan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler. 2015. *Lexical-Functional Syntax*, 2 edition. Blackwell Textbooks in Linguistics. Wiley-Blackwell, Chichester, England.

Huang C-TJ, Li A, and Li Y. 2009. *The Syntax of Chinese*. Cambridge University Press.

Lisa Lai-Shen Cheng and Rint Sybesma. 1999. Bare and not-so-bare nouns and the structure of np. *Linguistic inquiry*, 30(4):509–542.

Lisa Lai-Shen Cheng and Rint Sybesma. 2005. Classifiers in four varieties of chinese. *Handbook of comparative syntax*, pages 259–292.

Wenwen Cheng. 2015. Cong chutu wenxian kan “shu + liang + ming” jiegou de lishi xingcheng guocheng [the forming of the structure “numeral + quantifier + noun” in excavated documents] 从出土文献看“数+量+名”结构的历时形成过程. *Guhanyu Yanjiu [Research in Ancient Chinese Language] 古汉语研究*, (4):14–21.

Gennaro Chierchia. 1998. Reference to kinds across language. *Natural language semantics*, 6(4):339–405.

Noam Chomsky. 1970. Remarks on nominalism. In R. Jacobs and P. S. Rosenbaum, editors, *Readings in English Transformational Grammar*.

Noam Chomsky. 1995. *The Minimalist Program*. Cambridge University Press.

Colette G Craig. 1994. Classifier languages. *The encyclopedia of language and linguistics*, 2:565–569.

Paola Crisma, Cristina Guardiano, Giuseppe Longobardi, et al. 2020. Syntactic diversity and language learnability. *Studi e saggi linguistici*, 58(2):99–130.

William Croft. 1994. Semantic universals in classifier systems. *Word*, 45(2):145–171.



- Kristin Davidse. 2004. The interaction of quantification and identification in English determiners. Language, culture and mind, pages 507–533.
- Naoki Fukui. 1988. Deriving the differences between english and japanese: A case study in parametric syntax. English Linguistics, 5:249–270.
- Gerald Gazdar. 1985. Generalized phrase structure grammar. Harvard University Press.
- Lewis Gebhardt. 2011. Classifiers are functional. Linguistic Inquiry, 42(1):125–130.
- Jane Grimshaw. 2000. Locality and extended projection. AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4, pages 115–134.
- Colette Grinevald. 2000. A morphosyntactic typology of classifiers. Systems of nominal classification.
- Cristina Guardiano, Dimitris Michelioudakis, Andrea Ceolin, Monica Alexandrina Irimia, Giuseppe Longobardi, Nina Radkevich, Giuseppina Silvestri, Ioanna Sitaridou, et al. 2016. South by southeast. a syntactic approach to greek and romance microvariation. L'Italia dialettale, 77:95–166.
- Alice C Harris and Lyle Campbell. 1995. Historical syntax in cross-linguistic perspective, volume 74. Cambridge University Press.
- C He, Y Jiang, et al. 2011. Nominative xp-de and semantics of-de. (Contemporary linguistics), 13(1):49–62.
- Henry M Hoenigswald. 1950. The principal step in comparative grammar. Language, pages 357–364.
- Henry M Hoenigswald. 1965. Language change and linguistic reconstruction. University of Chicago Press, Chicago.
- Zaijun Huang. 1964. Cong jiawen, jinwen liangci de yingyong kaocha hanyu liangci de qiyuan yu fazhan [a study of the origin and development of classifiers in chinese, beginning from their use in the oracle bone inscriptions and metal inscriptions] 从甲骨文、金文量词的应用考察汉语量词的起源与发展. Zhongguo Yuwen, 133(6).
- Mutsumi Imai and Reiko Mazuka. 2003. Re-evaluating linguistic relativity: Language-specific categories and the role of universal ontological knowledge in the construal of individuation. Language in mind: Advances in the study of language and thought, pages 429–464.
- Shaoyu Jiang. 1994. Jindai Hanyu Yanjiu Gaikuang [A Survey Study on Early Mandarin] 近代汉语研究概况. Beijing: Peking University Press.
- Ronald M Kaplan, Joan Bresnan, et al. 1981. Lexical-functional grammar: A formal system for grammatical representation. Massachusetts Institute Of Technology, Center For Cognitive Science.
- M. Krifka, F. Pelletier, G. Carlson, and A. ter meulen. 1995. The Generic Book [C]. Chicago: The University of Chicago Press, 1-124.
- R. Langacker. 1991. Foundations of Cognitive Grammar. Vol. 2. Descriptive Application, volume 2. Stanford University Press, Stanford.
- Xuping Li. 2013. Numeral Classifiers in Chinese. Mouton de Gruyter, Berlin.
- Xuping Li and Walter Bisang. 2012. Classifiers in sinitic languages: From individuation to definiteness-marking. Lingua, 122(4):335–355.
- Yuming Li. 2000. Kaobei xing liangci ji qi zai han-zangyuxi liangci fazhan zhong de diwei [The copying type quantifiers and their position in the development of quantifiers in the Sino-Tibetan language family] 拷贝型量词及其在汉藏语系量词发展中的地位. Zhongguo Yuwen 中国语文, (1):27–34.
- Giuseppe Longobardi. 2014. Theory and experiment in parametric minimalism. Language description informed by theory. Amsterdam: John Benjamins, pages 217–262.
- Giuseppe Longobardi. 2017. Principles, Parameters, and Schemata: A Constructivist UG. Linguistic Analysis, 41(3 & 4):517–558.
- John Lyons. 1968. Introduction to theoretical linguistics, volume 510. Cambridge university press.
- Tsu-lin Mei. 1991. The development of predicate-complement structure based on *V-shā*, *V-sǐ* in the han dynasty. Yuyanxue Luncong [Essays on Linguistics], 16:169–171.
- Antoine Meillet. 1925. La méthode comparative en linguistique historique. Aschoug, Paris.
- Tatsuo Ota. [1958] 1987. A Historical Grammar of Modern Chinese. Peking University Press.
- Yunzhong Pan. 1982. Hanyu Yufa Shi Gaiyao [Outline of Historical grammar of Chinese] 汉语语法史概要. Zhongzhou Chinese Classics Publishing House.
- Carl Pollard and Ivan A Sag. 1988. Information-based syntax and semantics: Vol. 1: fundamentals. Center for the Study of Language and Information.
- Carl Pollard and Ivan A Sag. 1994. Head-driven phrase structure grammar. University of Chicago Press.
- Robert L. Rankin. 2017. The Comparative Method, chapter 1. John Wiley Sons, Ltd.
- Gisa Rauh. 2010. Syntactic categories: Their identification and description in linguistic theories. OUP Oxford.
- Jan Rijkhoff and Eva van Lier. 2013. Flexible word classes: Typological studies of underspecified parts of speech. OUP Oxford.

- Vitor Rocio, Mário Amado Alves, J Gabriel Lopes, Maria Francisca Xavier, and Graça Vicente. 2003. Automated creation of a Medieval Portuguese partial treebank. In Treebanks, pages 211–227. Springer.
- Chao-Fen Sun and Talmy Givón. 1985. On the so-called SOV word order in Mandarin Chinese: A quantified text study and its implications. Language, pages 329–351.
- Chaofen Sun. 1996. Word-order change and grammaticalization in the history of Chinese. Stanford University Press.
- Ann Taylor and Anthony S Kroch. 1994. The pennhelsinki parsed corpus of middle english. MS. University of Pennsylvania, page 30.
- Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The penn treebank: an overview. Treebanks: Building and using parsed corpora, pages 5–22.
- Eva Van Lier, Jan Rijkhoff, et al. 2013. Flexible word classes in linguistic typology and grammatical theory. Flexible word classes. Typological studies of underspecified parts of speech, pages 1–30.
- Li Wang. 1945. Zhongguo Yufa Lilun [Theory of Grammar in Chinese] 中国语法理论. Zhonghua Book Company.
- Li Wang. 1957. Hanyushi Gao [A Draft History of the Chinese Language] 汉语史稿. Number v. 1 in . China Science Publishing Media Ltd.
- Peiquan Wei. [1990] 2004. Han Wei Liuchao Chengdaici Yanjiu [Studies on Pronouns in the Eastern Han and Six Dynasties] 汉魏六朝称代词研究. Language and Linguistics. Institute of Linguistics, Academic Sinica.
- Peiquan Wei. 2000. 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] 东汉魏晋南北朝在语法史上的地位. Chinese Studies 汉学研究, 18:199–230.
- Werner Winter. 1984. Reconstructional comparative linguistics and the reconstruction of the syntax of undocumented stages in the development of languages and language families. Mouton Publishers.
- Yicheng Wu and Adams Bodo. 2009. Classifiers ≠ determiners. Linguistic Inquiry, 40(3):487–503.
- Jianguo Yang. 1993. Jindai Hanyu Yinlun [An introduction to Early Mandarin] 近代汉语引论. Huangshan Publishing House.
- Zhenwu Yao. 2008. “Hanyu ‘shu + liang + ming’ geshi de lai yuan” du hou [After reading ‘the origin of the ‘number + quantifier + noun’ format in Chinese’] 《汉语“数+量+名”格式的来源》读后. Zhongguo Yuwen 中国语文, (3):247–253.
- Cheng Zhang. 2010. Fazhan chuqi de hanyu mingliangci tedian—handai liangci yanjiu [the characteristics of measure words in early Chinese development: A study of measure words in the Han dynasty] 发展初期的汉语名量词特点—汉代量词研究. Journal of Chinese Language History 汉语史学报, (1):120–130.
- Niina Ning Zhang. 2013. Classifier Structures in Mandarin Chinese, volume 263. Walter de Gruyter.