A corpus-based study of lexical periodization in historical Chinese

Meng Ji
School of Law, Tohoku University, Japan

Abstract

The use of corpus material and methods represents a major methodological innovation in Chinese historical linguistics. The very exciting findings uncovered in this article may be seen as the first systematic large-scale investigation of the various morpho-syntactic patterns underpinning the evolution of Chinese lexis. In this article, we have made a ground-breaking investigation into the diverse lexical modes and patterns which have emerged and developed in each major period in Chinese history, in which the generation of corpus linguistic data and the subsequent computational statistical modelling have been essential.

Correspondence:
Meng Ji, School of Law, Tohoku University, Japan.
E-mail:
lolajimeng@yahoo.es;
lolajimeng09@gmail.com

1 Introduction

The practice of establishing a working timeline, also known as periodization in historical linguistics, is to provide a working guideline to the evolution of language across a wide time span. The division between any two main phases in the historical framework is supposed to be the point in time where major differences in the use of language seem to have occurred. Such variations may be reflected at various linguistic levels, e.g. phonological, grammatical, and lexical. While there has been some research on historical Chinese pursued mainly from a phonological or grammatical perspective (Karlgren, 1915–26; He, 1994), the broad area of Chinese historical lexis remains essentially underexplored. The purpose of this article is thus to provide the first corpus-based investigation on the distributional features of lexical constructs which have been instrumental in the multiplication and development of Chinese lexis in history, i.e. from 12th century BC until 1911, the year generally believed to the ending date of modern Chinese (Peyraube, 1996; Hu, 2005).

The use of corpus material and methods represents a major methodological innovation in Chinese historical linguistics. The very exciting findings uncovered in this article may be seen as the first systematic large-scale investigation of the various morpho-syntactic patterns underpinning the evolution of Chinese lexis. The complexities inherent in Chinese historical lexicography are well known (Yong *et al.*, 2008). The question of how Chinese lexis has developed across centuries from a predominantly monosyllabic language in remote ancient times to modern Chinese characterized by its enriched vocabulary has puzzled historical linguists for decades. Past studies pursued in this area hardly go beyond a philological recount based on observations in sketches. The lack of historical linguistic data gathered in large quantities has severely obstructed the formulation and advancement of theoretical hypotheses in the field. In this article, we have made a ground-breaking investigation into the diverse lexical modes and patterns emerged and developed in each major historical period in the Chinese history, in which the generation of corpus linguistic data and the subsequent computational statistical modelling have been essential.

For the purpose of this article, we have been using the Sheffield Corpus of Chinese, also known as SCC. The SCC is the first corpus of historical

Literary and Linguistic Computing, Vol. 25, No. 2, 2010.
doi:10.1093/llc/fqq002 Advance Access published on 12 April 2010

Chinese of its kind in terms of the wide time span it covers, the annotation depth it entails, and the great variety of texts it includes (Hu *et al*., 2007). At the first stage of expansion, SCC contains 432, 670 running words of Chinese historical texts divided into three sub-corpora reflecting the three major phases in the evolution of the language, i.e. archaic, medieval and modern Chinese. Within each sub-corpus, the corpus texts are further classified into early archaic (12th century to 206 BC) and late archaic (206 BC to AD 220); early medieval (AD 220–581), middle medieval (AD 581–860), and late medieval (AD 860–1368); early modern (AD 1368–1644) and late modern (AD 1644–1911).

As the corpus developers explained, such a chronological framework is mainly based on Peyraube (1988), where syntax serves as the main criterion for the periodization of historical texts. However, such practice has obviously simplified the internal complexities of the development of the Chinese language, where two key elements, i.e. syntax and lexis have exhibited very different patterns of development (Tai and Chan, 1998). The validity of the timeline used in the SCC has therefore become arguable where lexis is concerned. The suspicion is not to undermine the great efforts made by the developers of the SCC into categorizing the texts; instead, it posits the question of whether one could explore the SCC from an alternative perspective which is lexis-orientated rather than syntax-orientated.

The exploration and quantification of lexical patterns in the SCC is greatly facilitated by the dynamic annotation system established by its developers. Due to the particular nature of Chinese historical texts, the tagging system used in the SCC is largely empirical and experimental. That is, instead of applying a set of pre-defined general rules for the program to annotate the texts fully automatically—which has been achieved with a relatively high rate of success with subject-specific corpora in contemporary Chinese—the corpus builders of the SCC have to seek linguistic patterns constantly in the chronological texts, for the syntactic or lexico-grammatical rules governing the development of historical Chinese are still very much in the balance.

Our study is, therefore, to show that the construction of diachronic corpora, especially the development of annotation systems for historical linguistics offers new and promising perspectives on the investigation of issues such as lexical periodization which can be rather elusive to pin down without solid linguistic evidence extracted from language corpora. Moreover, apart from producing linguistic data for testing hypotheses based on one's intuitive judgements, novel linguistic categories as emerged from newly developed corpora have shown important theoretical implications: the computer-assisted documentation and processing of language material is increasingly shaping the way how we observe quantitative linguistic events and advance new hypotheses.

2 Research Question: Schematic Morpho-Syntactic Patterns in Chinese Historical Lexis

As may be noticed from its extensive list (see Appendix I), a distinctive feature of the mark-up scheme developed for the SCC is that in order to reflect the evolving nature of the language, individual items of annotation devised here cannot be readily classified into conventional tagging schemes, such as part-of-speech, syntactical, or lexical. On the contrary, it is largely morpho-syntactic exhibiting various recurrent patterns that the general lexicon of historical Chinese may well have followed in its evolution. By morpho-syntactic, within the context of Chinese historical linguistics, we refer to a particular language construct which usually derives from a monosyllabic word or morpheme as its core semantic component and extends into a larger lexical unit by agglutinating or using in conjunction with a functional morpheme. Within a morpho-syntactic unit, on the one hand, the monosyllabic word, or the semantic part of the construct is substitutive; or in other words, it may be changed for any other monosyllabic words or morphemes; on the other hand, the functional morpheme sustaining the morpho-syntactic unit remains intact as the semantic part varies. Typical examples are 相會,

相辭 (reciprocal_xiang_V); 自寬, 自縱 (reflexive_zi_V) and 所積, 所吟 (suo_V), etc.

The identification of functional morphemes is greatly facilitated by the SCC tagging system where Functional Morpheme is singled out as an individual category (See Appendix 1 from FMA to FML). Tagging items of the SCC which may be classified under the first type of morpho-syntactic constructs include EPC (generative_zhi_N), EPD (generative_zhi_suo_V), VBI (bei_V), VBJ (bu_V), VBO (jian_V), VBR (reflexive_xiang_V), VBS (reflexive_zi_V), VBU (stative_comparative), VBV (stative_superlative), VBW (suo_V), VBY (V_hua), VBZ (V_lai), VBBB (V_potential_bu_RVC), VBCC (V_potential_de_RVC), VBDD (V_qu), VBHH (V_yu), VBII (V_zhi), and VBJJ (yi_V). The proportions represented by these morpho-syntactic constructs for each of the main POS categories are summarized below: adjectives (0%), expressions (25%), noun (0%), onomatopoeia (0%), verb (44.4%).

Another unique and very productive type of morpho-syntactic constructs which emerged in the SCC annotation scheme is characterized by the repetition or alternation of constituent monosyllabic words within a fixed word span, ranging from two to four characters long. This linguistic phenomenon is known as *ge* (格) in Chinese, which may be translated as 'mould' or 'pattern'. Instead of being exploited by a specific type of words, this kind of morpho-syntactic construct is widely seen in a range of lexical and grammatical categories such as noun, adjective, onomatopoeia and verb. Sometimes, a four-character morpho-syntactic construct becomes so conventionalized that it enters gradually into the domain of fixed or idiomatic expressions, which gives the name to Chinese idioms, i.e. 四字格成语 (idioms composed of four characters; idioms in the four-character pattern). To some extent, it may be said that the second type of morpho-syntactic constructs in historical Chinese shows a higher level of structural abstractness compared to the first type, where the structural flexibility is still in part.

As different from words of distinctive semantic or syntactic properties, such morpho-syntactic constructs are highly dynamic. They lie at the heart of the evolution of Chinese, where the predominant position of monosyllabic words in pre-Qin Chinese (12th century to 206 BC) is gradually taken over by the proliferation of polysyllabic words at later times (Tai and Chan, 1998). That is, the Chinese language has entered into a historical process of lexical diversification in which the development of morpho-syntactic constructs may well have been essential.

Through their regular occurrence in the annotation scheme of the SCC, a number of underlying morpho-syntactic patterns seem to emerge, e.g. AA, AAB, ABA, ABB, AABB, ABAB, ABAC, ABCB, and ABBC (see Appendix 1). The crucial importance of lexical words which have been formed by means of this kind of schematic patterns is easy to capture by calculating the proportions that they represent for each category of the SCC marking up scheme: adjective (83.3%), adverb (33.3%), noun (38.9%), onomatopoeia (80%), and verb (25%).[1] Overall, the semi-fixed patterns account for more than 52.1% of the total of the annotation categories developed for the current version of the SCC (noting that the tagging system is still open to new lexical patterns as the corpus expands and more historical texts are added to it). This is a very revealing finding regarding the way in which the Chinese lexis may have evolved and multiplied over the centuries. To the best of our knowledge, very little research has been done in this area, neither from a computational linguistic nor from a theoretical linguistic perspective.

3 Research Methodology and Data

In the present article, we will have a close look at the distribution of these recurrent morpho-syntactic patterns in the SCC, in an effort to track down the lexical evolution of historical Chinese. Corpora and corpus analysis techniques used in this study have been widely used in English corpus linguistics and corpus statistics (Oakes, 1998; McEnery *et al.* 2006), though their validity and productivity for Chinese historical linguistics has been tested to a rather limited extent (Hu *et al.* 2007). Specifically, through the automatic retrieval function facilitated

by its in-built search engine, we shall first gather a large amount of textual data from the corpus and then subject the quantitative corpus data to statistical analysis to find out whether the seven major historical periods in the SCC are indeed different from each other in terms of the distribution of these semi-fixed morpho-syntactic patterns. If so, it is then necessary to find out whether it would be possible to set up a statistical model to re-group the seven historical periods into new diachronic clusters measured by the frequency of these linguistic variables as a very distinctive feature of the lexical evolution of the Chinese language. This entails the development of a new periodization framework for the SCC based on the quantitative empirical data gathered so far.

Table 1 provides a breakdown of the normalized frequency of occurrence of the forty-four types of morpho-syntactic constructs established in the SCC. We proceed with statistical analysis to mark out the systematic difference and correlation strength among the seven sub-corpora in terms of the distribution of the forty-four morpho-syntactic constructs. Before conducting any statistical analysis, one would need to examine the normality of the distribution of the data under investigation. This is a priori condition in any corpus-based statistical analysis, though its importance tends to be underestimated largely due to a lack of awareness of the potential complexities that data transformation may cause for the interpretation of the statistical outcome at a later stage.

Nevertheless, exploratory data analysis such as normality checking has to be run first to ensure the validity of the statistical analysis. This is also very important for non-parametric tests which are normally assumed to be 'distribution-free' (Zimmerman, 1998, pp. 55–68). This is especially necessary in the present corpus-based study. As mentioned above, the current version of the SCC is still at its first stage of expansion. The developing nature of the corpus is reflected in its less-balanced sampling structure (Hu *et al.* 2007), a fact which can only be remedied by the constant expansion of the corpus over a relatively long period of time. It is suspected that the corpus data extracted for the current version of the SCC might be susceptible to high skewness. To test this, we have performed the exploratory data analysis embedded in SPSS Version 15.0.

Table 2 offers an initial descriptive analysis of the data presented in Table 1. The skewness scores in the bottom row show that the seven sub-corpora included in the SCC are invariably positively skewed, which means that most of the values of the dependent variables are clustered to the left-hand side of the distribution curve. In order to improve the normality of the data distribution and makes it compliant with most statistical tests, it becomes essential to transform the data. There are three most commonly used data transformation techniques which are square root, logarithmic, and inverse transformation. The actual transformation process, however, is not that straightforward as it may appear to be. For instance, prior to computing the square root of the counts shown in Table 1, it is important to re-set the minimum value of the distribution to make sure of equidistant spacing between the transformed data.

This has much to do with the specific mathematical test selected for the purpose of data transformation (for detailed explanations cf. Osborne, 2002). The optimized minimum score of a distribution is anchored at 1.0, where the data transformation will be most effective. The addition of a constant will not alter the essential properties of a distribution curve, namely its standard deviation or variance and skewness. That is, after the transformation, all the data points remain in the same relative order, allowing the researcher to make predictions or to interpret the statistical results in terms of increasing scores. If we go back to Table 1, we will see that the minimum value for each of the seven sub-corpora is zero. To optimize the transformation process, we increase the minimum value from zero to one (Fig. 1).

P–P plots show that the datasets for the seven sub-corpora, which have undergone the natural log transformation, now exhibit greatly improved distribution curves which resemble that of a normal distribution. This has been very important preparation for the use of statistical analyses which are contingent on the availability of a normal

Table 1 Distribution of schematic morpho-syntactic patterns in the SCC[a]

POS	Tag	Pattern	Early Archaic	Late Archaic	Early Medieval	Middle Medieval	Late Medieval	Early Modern	Late Modern
Expression	EPC	genitive_zhi_N	312	435	104	179	165	85	114
Expression	EPD	genitive_zhi_suo_V	19	5	0	2	1	1	1
Verb	VBI	Bei_V	0	2	0	0	0	1	0
Verb	VBJ	Bu_V	262	232	96	136	14	113	129
Verb	VBO	Jian_V	1	5	4	3	0	5	3
Verb	VBR	reciprocal_xiang_V	11	21	24	45	18	31	16
Verb	VBS	reflexive_zi_V	14	25	17	15	24	15	17
Verb	VBU	stative_comparative	1	0	0	1	2	1	0
Verb	VBV	stative_superlative	3	2	0	0	4	2	3
Verb	VBW	suo_V	18	32	48	20	16	12	18
Verb	VBY	V_hua	1	2	1	1	2	3	3
Verb	VBZ	V_lai	1	0	3	6	15	38	48
Verb	VBBB	V_potential_bu_RVC	0	0	0	0	12	15	11
Verb	VBCC	V_potential_de_RVC	0	0	0	0	4	3	2
Verb	VBDD	V_qu	0	0	1	3	10	19	20
Verb	VBHH	V_yu	76	55	13	16	17	7	7
Verb	VBII	V_zhi	165	146	63	46	53	21	32
Verb	VBJJ	yi_V	2	4	0	4	2	11	15
Adjective	AJB	AA	21	9	3	23	24	18	22
Adjective	AJC	AAB	3	0	0	0	0	0	0
Adjective	AJD	AABB	3	2	0	4	4	8	2
Adjective	AJE	ABAB	0	0	0	0	0	0	1
Adjective	AJF	ABB	1	0	0	2	13	4	2
Adverb	AVB	AA	0	0	0	1	4	2	4
Noun	NNB	AA	1	2	1	7	3	7	14
Noun	NNC	AAB	0	0	0	0	0	0	0
Noun	NND	AABB	0	0	0	0	0	1	1
Noun	NNE	ABAB	0	0	0	0	0	0	1
Noun	NNF	ABAC	1	1	1	0	0	3	2
Noun	NNG	ABB	0	0	0	1	3	2	3
Noun	NNH	ABAC	0	0	0	0	1	0	0
Onoma	ONA	AA	1	0	0	2	0	2	3
Onoma	ONB	AAA	0	0	0	0	2	0	0
Onoma	ONC	AABB	0	0	0	1	0	0	0
Onoma	OND	ABBC	0	0	0	0	0	0	0
Verb	VBB	AA	2	0	0	0	1	5	3
Verb	VBC	AAB	0	0	0	0	0	0	0
Verb	VBD	AABB	0	0	0	1	0	2	1
Verb	VBE	ABAB	2	0	0	0	0	1	2
Verb	VBF	ABAC	18	33	1	6	15	21	2
Verb	VBG	ABB	0	0	0	0	3	1	0
Verb	VBH	ABCB	0	3	0	1	1	0	0
Verb	VBX	V_bu_V	0	0	0	0	1	0	0
Verb	VBGG	V_yi_V	0	0	0	0	1	2	2

[a]Due to the different sizes of sub-corpora, raw frequencies have been converted into frequencies per 10K words.

distribution of the data. This article aims to test the applicability of the current diachronic framework of the SCC in the study of the lexical evolution in historical Chinese. To examine the similarities among the variables and to arrange them accordingly into new sub-groups regarding the underlying patterns in the use of semi-fixed morpho-syntactic constructs, we perform the Hierarchical Cluster

Table 2 Exploratory data analysis on the dataset

	Early Archaic	Late Archaic	Early Medieval	Middle Medieval	Late Medieval	Early Modern	Late Modern
N							
Valid	44	44	44	44	44	44	44
Missing	0	0	0	0	0	0	0
Mean	153.93	86.89	36.77	49.07	84.39	114.84	52.39
Standard deviation	467.550	284.320	100.316	138.512	210.185	238.934	118.956
Variance	218603.274	80837.964	10063.203	19185.600	44177.545	57089.532	14150.429
Skewness	3.694	4.497	3.199	4.060	3.972	4.585	3.672

Analysis (SPSS 15.0), one of the most straightforward clustering techniques.

Figure 2 shows the dendrogram generated by the Hierarchical Cluster Analysis. The items listed on the left stand for the transformed datasets for the seven historical periods. The statistical method used is agglomerative clustering where the measure selected is squared Euclidean distance under the interval data category. Under such conditions, variables with most similarities are first identified and merged into a cluster (Cluster 1). Cluster 1 then moves on to a higher level seeking for its closest 'friend' which can be either an individual variable or a cluster formed in a similar manner. In this way, the dendrogram gradually grows from left to the right, where the differences among all the variables are finally enclosed within a single cluster. The distance between two clusters or between a cluster and a single variable indicates how different the two constituent parts are from each other. This is illustrated on the tree graph by the length of the branch measuring the relative position of any two connecting points: the longer the branch, the more different the two conjoined clusters.

In Fig. 2, the initial level of clusters forms between Early Archaic and Late Archaic (1–2), Early Medieval and Middle Medieval (3–4). As can be seen, within each cluster, the distance between the two variables is the shortest in the whole dendrogram, suggesting that among others, the similarities between Early Archaic and Late Archaic, and those between Early Medieval and Middle Medieval are the most prominent. The second cluster occurs when the first two initial clusters are joined together with a relatively long distance in between. The first distributional pattern emerged in the dendrogram as divergent from the chronological framework deployed in the SCC is that there seems to be a greater similarity between Late Medieval and Early Modern than between Middle Medieval and Late Medieval. Therefore, a dividing point as suggested by the Hierarchical Clustering Analysis should be drawn between Middle Medieval and Late Medieval, rather than between Middle Medieval and Late Medieval. As a result, the third cluster is identified between Late Medieval and Early Modern. Lastly, the fourth cluster converges with the remaining variable Late Modern, where the dissimilarity between the two, as visualized by the tree branch is rather small.

The new periodization framework uncovered above brings us to the next question of how the seven sub-corpora have been re-clustered. In other words, what could be the main dimensions of factors that have influenced the statistical modelling of the underlying patterns in the lexical development of historical Chinese? To answer this question, we will have to go back to Table 1, which summarizes the distribution of the forty-four morpho-syntactic constructs in the SCC. As an initial classification, all the dependent variables have been categorized as (1) semi-fixed morpho-syntactic constructs and (2) schematic morpho-syntactic patterns (or *ge*). This criterion used here is largely structural based on the researcher's observation of the distinctive linguistic features of the lexical units under investigation. This, however, may well not be the criterion deployed by the statistical test, i.e. Hierarchical Clustering Analysis in the actual text classification. To probe into the internal structure of the observable variables, we will use the exploratory factor analysis, i.e. principal component analysis (PCA).

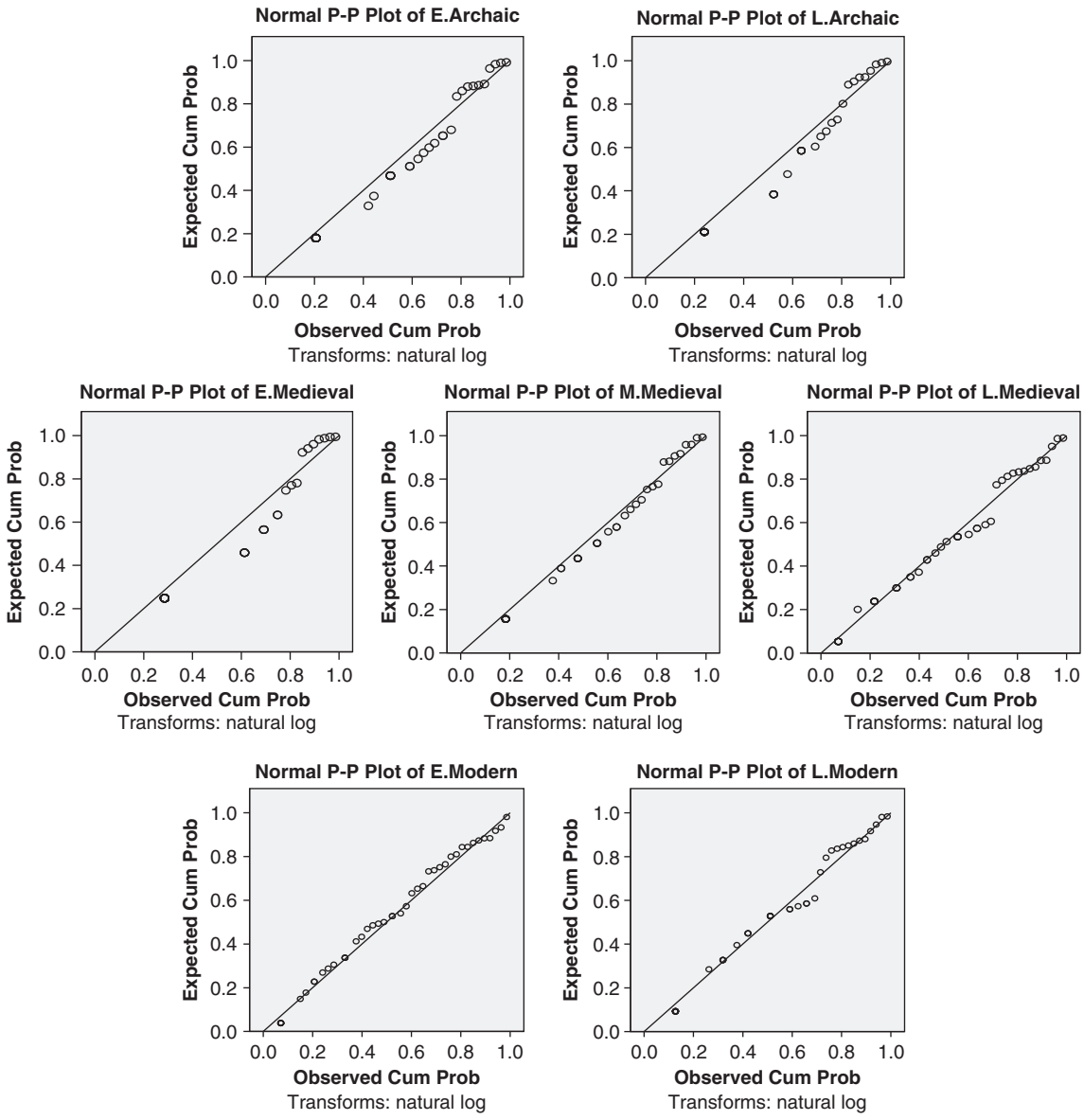

Fig. 1 P–P plots of log transformed seven sub-corpora of the SCC

PCA is a classical statistical analysis used for the purpose of data reduction and predictive model construction. Its main function is to reveal the internal structure of the data that may best explain the behaviour of the independent variables. It classifies the original observable variables into a reduced number of statistical dimensions, normally two or three (in the case of high dimensional data, the principal components abstracted may increase accordingly), to maximally account for the variations in the original dataset. The principal components or factors are constructed in such a way that they are minimally correlated with each other. The first factor of dependent variables is always the most

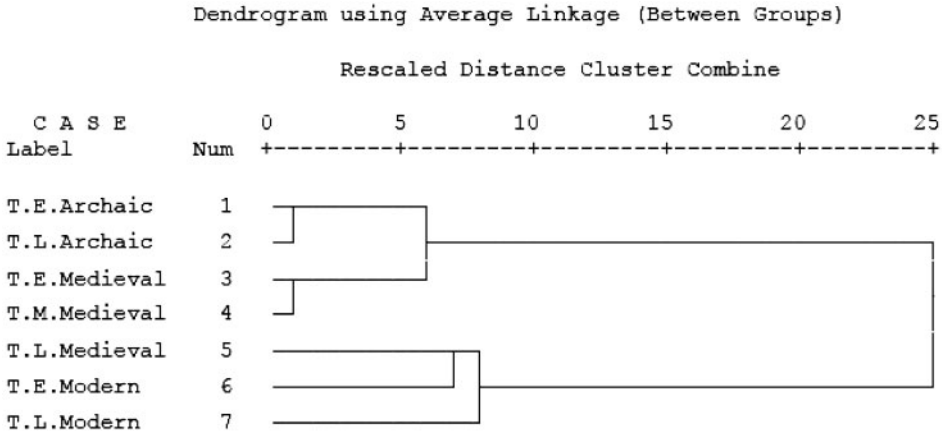

Fig. 2 Hierarchical clustering analysis

Table 3 PCA of the distribution of morpho-syntactic constructs in the SCC

Component	Initial Eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings		
	Total	Percentage of variance	Cumulative percentage	Total	Percentage of variance	Cumulative percentage	Total	Percentage of variance	Cumulative percentage
1	22.204	50.463	50.463	22.204	50.463	50.463	20.417	46.403	46.403
2	8.625	19.601	70.064	8.625	19.601	70.064	9.309	21.156	67.559
3	6.740	15.317	85.381	6.740	15.317	85.381	6.844	15.554	83.113
4	3.585	8.147	93.528	3.585	8.147	93.528	2.868	**6.517**	89.630
5	1.517	3.447	96.975	1.517	3.447	96.975	2.664	**6.054**	95.685
6	1.331	3.025	100.000	1.331	3.025	100.000	1.899	**4.315**	100.000

stable one and explains the largest amount of variance in the dataset; the second factor is constructed to account as much as possible for the remaining variations. The efficiency of a factor in a PCA model is expressed through the statistical indicator, Variance Accounted For (VAF), usually reported as a percentage. As different from parametric tests, e.g. cluster analysis, normality in data distribution is not considered a critical assumption in factor analysis.

Table 3 shows the result of the PCA of the original observable variables presented in Table 1. The PCA has extracted six factors from the forty-four dependent variables, the cumulative VAF of which account for the total variance in the dataset. However, if we have a close look at the cross-tabulation, it is easy to detect that not all the principal components are worth retaining in the statistical model. Compared to the first three main dimensions of variables, the efficiency of the last three dimensions is rather limited. For instance, the nested VAF of the last three principal components is just slightly higher than that of the third principal component in rank. Without the last three, the PCA model already explains the great majority of variance in the observed variables. To obtain the most parsimonious model, we suggest limiting the highest dimension of the PCA to three, rather than six. To look into the internal structure of the dataset as modelled by the PCA, we have produced a breakdown of the loadings of each observed variable on the three principal components (Table 4).

Table 4 Component loadings

Rotated principal component matrix							
Variable	**1**	**2**	**3**		**1**	**2**	**3**
VBJJ	0.984	−0.104	−0.091	NNC	0.740	−0.113	−0.145
VBZ	0.984	0.063	−0.162	VBS	0.736	0.533	0.197
NNF	0.979	−0.079	0.076	VBE	0.730	−0.237	0.581
VBB	0.978	−0.033	0.126	VBF	0.708	0.128	0.393
VBGG	0.976	0.113	−0.163	ONB	−0.128	0.987	−0.042
VBDD	0.971	0.153	−0.158	NNH	−0.128	0.987	−0.042
ONA	0.962	−0.101	−0.066	VBX	−0.128	0.987	−0.042
NND	0.957	0.236	−0.125	OND	−0.128	0.987	−0.042
NNB	0.941	−0.064	−0.240	VBG	0.314	0.936	-0.050
VBBB	0.930	0.308	−0.112	AJF	0.422	0.898	−0.052
VBC	0.925	−0.194	0.308	VBU	0.496	0.710	0.249
VBY	0.923	0.281	−0.069	AVB	0.581	0.707	−0.250
AJD	0.896	0.124	0.102	AJE	0.069	0.598	−0.242
VBO	0.870	−0.232	−0.089	VBHH	−0.174	−0.097	0.978
VBCC	0.816	0.554	−0.119	EPD	−0.055	−0.138	0.977
VBR	0.806	0.030	−0.128	AJC	−0.014	−0.018	0.970
NNG	0.792	0.531	−0.238	VBII	−0.205	−0.083	0.964
AJB	0.777	0.394	0.372	VBJ	0.411	0.068	0.894
VBD	0.757	0.582	−0.117	EPC	−0.516	0.006	0.805
VBV	0.755	0.471	0.420				

Table 4 shows the distribution of the observed variables on the statistically constructed dimensions. As can be seen from the table, the stability of the first principal component is reflected in the fact that there are numerous observed variables heavily loaded on this dimension, which together account for nearly half of the total variance in the original dataset (see Table 3). The first dimension includes items listed in the first column as well as the first types of morpho-syntactic patterns highlighted in the second column, i.e. from VBJJ (yi_V, e.g. 一訪, 一望) to VBF (ABAC, e.g. 包長包短). To a lesser extent, the second dimension is substantiated by another nine variables from ONB (AAA, e.g. 騰騰騰, 撒撒撒) to AJE (Non-predicate_ABAB, e.g. 筍條筍條) in the second column. This accounts for as many as one third of the variables attributed to Dimension 1. The reliability of this scale is confirmed by the fact that all nine variables are weighted heavily and clearly on that dimension, without any confusing element with equally high loadings on any other dimensions. Dimension 3 also exhibits a proven level of robustness with all of the last six variables from VBHH (V_yu, e.g. 起於) to EPC (genitive_zhi_N, e.g. 之理, 之屬) loading unequivocally high on that component, ranging from 0.805 to 0.978. While the PCA has successfully constructed a robust statistical model, it is interesting to notice that the initial classificatory line drawn between the two main types of morpho-syntactic constructs has not been taken into account in the statistical modelling; on the contrary, individual lexical units of distinctive structural properties now have all been mixed up and re-clustered into new dimensions of factors: 1, 2, and 3.

4 Findings

What is the conceptual significance of these new dimensions of variables within the context of this corpus-based study of Chinese historical lexicography? Has the statistical data reduction been a random process leading up to an algorithm assigning original observed variables to new dimensions without real sense? The picture drawn by the PCA seems to have made the original question more

complex than we expected—what are the main factors of linguistic variables that have distinguished the seven historical periods in the SCC? The construction of conceptual scales is another major function of the PCA. It has a wide and long-standing application in medical and biological sciences where the principal components abstracted from the statistical modelling serve as the defining features in the nomination of new genetic diseases (Reich *et al.,* 2008). In literary and linguistic research, past studies have used the PCA in the classification of historical manuscripts (Ousaka and Yamazaki, 2002) and the detection of stylistic variation and its various linguistic representations (McKenna and Antonia, 2001), while discussion on the construction and interpretation of conceptual scales in the PCA as revealing theoretical instruments has been rather light.

The interpretation of conceptual scales, however, entails a critical analytical process in which the theoretical insights brought about by statistical analysis need to be interpreted in terms of the common features shared by the original variables newly classified and attributed to each dimension of variables. The interpretive process of the latent classificatory criteria used in the statistical modelling can be rather difficult and requires a discerning eye to spot the defining feature of a principal component from the diverse linguistic traits shown by a normally large number of observed variables sustaining that component. However, the identification or decipherment of the hidden classificatory scheme suggested by the statistical analysis holds the key to a deeper understanding of the internal structure of the subject under investigation, and more importantly, potential theoretical values for further research.

At first sight the component score matrix shown in Table 4 seems very difficult to interpret with regards to any possible defining features of each principal component: individual items classified under each dimension exhibit very different structural properties and are distributed across a wide range of part-of-speech categories. They do not seem to share any semantic similarities and are found in a number of text genres assuming different pragmatic functions. They do not show a uniform pattern in terms of their phonological composition in syllables and the stylistic connotations of their use in historical texts also vary considerably. It seems fairly elusive to pin down the latent structure of the statistical model by relying solely on the linguistic features of the observed variables. Another important factor that we have not considered is the diachronic status of each variable, i.e. the historical period in which a morpho-syntactic construct is first coined and then fully explored in successive periods. By using the online search engine of the SCC, we track down the chronological distribution of the variables categorized under each dimension and striking patterns begin to emerge to our surprise.

Table 5 summarizes the chronological distribution of the variables sustaining each principal component. Despite the considerable diversity in their structural and linguistic features, revealing patterns regarding the internal coherency and divergence in the data distribution gradually come to light. In the third column counting from the left, we have the main historical period in which a certain lexical construct occurs most prominently. It is followed by another two 'minor' historical periods in which the lexical construct is found with a relatively significant frequency of occurrence. As can be seen, the main historical period of distribution has naturally clustered the numerous variables into three larger groups which correspond exactly to the three principal components abstracted by the PCA. This suggests that the main historical period of distribution may well be the latent classificatory criteria used in the statistical modelling.

5 Conclusion

This article offers an original corpus-based investigation of the evolution of Chinese lexis by focusing on the distributional patterns of various morpho-syntactic patterns in the Sheffield Corpus of Chinese, or SCC for short. Through the annotation and extraction of quantitative linguistic data from the SCC, a large amount of textual information was uncovered, with a view to establishing an empirically solid framework of periodization for Chinese historical linguistics. The statistical analysis of the corpus data constructed a linguistic model which classified the original dataset into three main dimensions of variables. The qualitative

Table 5 Chronological distribution of morpho-syntactic constructs in the SCC

Dimension	Morpho-syntactic patterns	Main diachronic distribution (%)					
1	VBJJ	Ming	51.8	Qing	24.8	n/a	
1	VBZ	Ming	57.7	Qing	25.4	Song and Yuan	11.6
1	NNF	Ming	50.0	Qing	18.8	Pre-Qin	12.5
1	VBB	Ming	60.0	Qing	16.0	Pre-Qin	14.0
1	VBGG	Ming	57.1	Qing	28.6	Song and Yuan	14.3
1	VBDD	Ming	58.4	Qing	21.2	Song and Yuan	15.7
1	ONA	Ming	43.3	Qing	23.3	Sui and Tang	13.3
1	NND	Ming	60.0	Qing	20.0	Song and Yuan	20.0
1	NNB	Ming	41.7	Qing	27.8	Sui and Tang	12.2
1	VBBB	Ming	60.1	Qing	15.5	Song and Yuan	23.5
1	VBC	Ming	50.0	Qing	25.0	Pre-Qin	25.0
1	VBY	Ming	39.5	Qing	14.0	Song and Yuan	18.6
1	AJD	Ming	57.8	Song and Yuan	13.3	Pre-Qin	11.1
1	VBO	Ming	45.4	W and E Han	14.2	Qing	11.3
1	VBBC	Ming	50.0	Song and Yuan	35.0	Qing	15.0
1	VBR	Ming	39.0	Sui and Tang	17.6	Song and Yuan	11.4
1	NNG	Ming	39.1	Song and Yuan	30.4	Qing	21.7
1	AJB	Ming	29.8	Song and Yuan	20.1	Pre-Qin	19.1
1	VBD	Ming	50.0	Song and Yuan	35.7	n/a	
1	VBV	Ming	32.7	Song and Yuan	24.5	Pre-Qin	20.4
1	NNC	Ming	66.7	Sui and Tang	33.3	n/a	
1	VBS	Ming	26.0	Song and Yuan	20.4	Pre-Qin	13.7
1	VBE	Ming	37.5	Pre-Qin	37.5	Qing	25.0
1	VBF	Ming	42.2	Pre-Qin	19.3	W and E Han	18.7
2	ONB	Song & Yuan	100	n/a		n/a	
2	NNH	Song & Yuan	100	n/a		n/a	
2	VBX	Song & Yuan	100	n/a		n/a	
2	OND	Song & Yuan	100	n/a		n/a	
2	VBG	Song & Yuan	66.7	Ming	33.3	n/a	
2	AJF	Song & Yuan	51.8	Ming	32.1	Qing	7.1
2	VBU	Song & Yuan	31.6	Ming	31.6	Pre-Qin	15.8
2	AVB	Song & Yuan	34.2	Ming	26.3	Qing	23.7
2	AJE	Song & Yuan	50	Qing	50	n/a	
3	VBHH	Pre-Qin	49.2	W and E Han	18.6	Song and Yuan	10.1
3	EPD	Pre-Qin	69.7	W and E Han	9.1	Ming	9.1
3	AJC	Pre-Qin	80	Song and Yuan	10	Ming	10
3	VBII	Pre-Qin	40.3	W and E Han	18.5	Song and Yuan	11.6
3	VBJ	Pre-Qin	28.2	Ming	22.0	Song and Yuan	13.7
3	EPC	Pre-Qin	29.1	W and E Han	21.1	Ming	14.4

interpretation of the three conceptual scales sustaining the statistical model pointed to an alternative scheme of periodization for historical Chinese which was largely lexis based (see Table 6). The newly proposed periodization framework is a very exciting and monumental finding in Chinese historical linguistics. Past studies of Chinese historical lexicography have rarely reached this micro-structural level of investigation.

It is the first time that we have succeeded in mapping out the diachronic distribution of *a number* of very productive linguistic constructs in the evolution of Chinese lexis, which have not even been noticed in the past. The systematicity of our research and the linguistic evidence that we have gathered so far has been ground breaking and has greatly promoted the theoretical advancement of the field. However, at the same time, it should be

Table 6 PCA model for lexical periodization of historical Chinese

Dimension	Historical periods(in which most variables on the component occur)	Conceptual scale
1	**Ming [Early Modern];** Qing [late modern]	**Modern**
2	**Song and Yuan** [late Medieval]; Ming [early modern]	**Late Medieval**
3	**Pre-Qin** [early Archaic]; Western and Eastern Han [late Archaic]	**Archaic**

pointed out that due to the developing nature of the SCC, which is still quite limited in size to reflect the extremely complex process of the evolution of Chinese lexis, more empirical evidence will be needed to verify the validity and wider applicability of the initial investigation that we have made here. For example, since the current version of the SCC has gathered different text genres for each sub-corpus representing each sub-period, one may want to find out apart from the temporal factor, if text genre might also have a role to play in the distribution of the various morpho-syntactic patterns attributed to each historical phase as suggested in Table 5; or how dynamic this distributional map could be as more and more lexical-grammatical tags are constantly added to the current version of the annotation scheme. Such questions which are to be addressed at a later stage as the SCC expands imply that the theoretical model proposed in this article will necessarily be modified and enhanced in our future research through the testing of the model on historical texts to be incorporated in the SCC.

References

Hu, X., Williamson, N., McLaughlin, J. *et al.* (2005). Sheffield Corpus of Chinese for diachronic linguistic study. *Literary and Linguistic Computing*, **20**(3): 28193.

Hu, X., McLaughlin, J., and Williamson, N (2007). Syntactic positions of prepositional phrases in the history of Chinese: using the developing Sheffield Corpus of Chinese for diachronic linguistic study. *Literary and Linguistic Computing*, **22**(4): 419–34.

McEnery, A. M., Piao, S., Wilson, A. *et al.* (2006). A large semantic lexicon for corpus annotation. In *The Proceedings of the Corpus Linguistics Conference 2005.* Centre for Corpus Research, Birmingham University, available at http://www.corpus.bham.ac.uk/pclc/ (last accessed March 2010).

McKenna, C. and Antonia, A. (2001). The statistical analysis of style: reflections on form, meaning and ideology in the 'Nausicaa' episode of Ulysses. *Literary and Linguistic Computing*, **1**(4): 353–73.

Oakes, M. (1998). *Statistics for Corpus Linguistics.* . Edinburgh: Edinburgh University Press.

Osborne, J. (2002). Notes on the use of data transformation. *Practical Assessment, Research and Evaluation.* **8**(6). http://pareonline.net/getvn.asp?v=8&n=6.

Ousaka, Y. and Yamazaki, M. (2002). Genealogical classification of saddharmapundarika manuscripts based on many-variable analysis. *Literary and Linguistic Computing*, **17**(2): 193–206.

Peyraube, A. (1996). Recent issues in Chinese historical syntax. In Huang, J. C. T. and Li, A. Y. H. (eds), *New Horizons in Chinese Linguistics, Studies in Natural Language and Linguistics Theory 35.* London and Boston: Kluwer, pp. 161–214.

Reich, D., Price, A., and Patterson, N. (2008). Principal component analysis of genetic data. *Nature*, **40**: 491–2.

Tai, H. Y. and Chan, K. M. (1998). Some reflections on the periodization of the Chinese language. In Peyraube, A. and Sun, C. (eds), *Studies in Chinese Historical Syntax and Morphology: Linguistic Essays in Honour of Meit Tsu-lin.* Paris: Ecole des Hautes Etudes en Sciences Sociales, pp. 223–9.

Yong, H. and Peng, J. (2008). *Chinese Lexicography: A History from 1046 BC to AD 1911.* New York: Oxford University Press.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education*, **67**: 55–68.

Note

1 Two types of verbs, i.e. VBX [V_bu_V] and VBGG [V_yi_V] have been categorized under the schematic pattern type of ABA.

Appendix I Annotation scheme of the Sheffield Corpus of Chinese

Tag label	Word class	Category
AJA	Adjective	non_predicate (e.g. 溜清, 噴香)
AJB	Adjective	non_predicate_aa (e.g. 溥溥, 蕩蕩)
AJC	adjective	non_predicate_aab (e.g. 黯黯然)
AJD	adjective	non_predicate_aabb (e.g. 熟熟馴馴)
AJE	adjective	non_predicate_abab (e.g. 筍條筍條)
AJF	adjective	non_predicate_abb (e.g. 酸蔭蔭)
AVA	adverb	general (e.g. 約莫, 直截)
AVB	adverb	AA (e.g. 常常, 往往)
AVC	adverb	negative (e.g. 未, 休)
CJA	conjunction	coordinating (e.g. 和, 但是)
CJB	conjunction	subordinating (e.g. 假若, 因爲)
CLA	classifier	(e.g. 粒, 幅)
EPA	expression	direction (e.g. 庵北, 其西)
EPB	expression	formulaic (e.g. 端的, 不期)
EPC	expression	genitive_zhi_N (e.g. 之理, 之屬)
EPD	expression	genitive_zhi_suo_V (e.g. 之所恃)
EPE	expression	location (e.g. 廳外, 崖下)
EPF	expression	nominal (e.g. 吊孝的)
EPG	expression	order (e.g. 吸前)
EPH	expression	time (e.g. 嘉祐中, 慶歷中)
FMA	functional_morpheme	adverbial (e.g. 地)
FMB	functional_morpheme	aspect_durative (e.g. 着)
FMC	functional_morpheme	aspect_experiential (e.g. 過)
FMD	functional_morpheme	aspect_perfective (e.g. 了)
FME	functional_morpheme	causative (e.g. 使)
FMF	functional_morpheme	complement (e.g. 得)
FMG	functional_morpheme	emphatic (e.g. 所)
FMH	functional_morpheme	general (e.g. 聿)
FMI	functional_morpheme	genitive (e.g. 之)
FMJ	functional_morpheme	objective (e.g. 把)
FMK	functional_morpheme	passive (e.g. 見, 被)
FML	functional_morpheme	plural (e.g. 們)
FMM	functional_morpheme	relative (e.g. 的)
IDA	idiom	(e.g. 斐然成章)
ITA	interjection	(e.g. 嗚呼, 哎)
LCA	localizer	(e.g. 上, 后)
NNA	noun	common (e.g. 劍客, 糧食)
NNB	noun	AA (e.g. 根根, 人人)
NNC	noun	AAB (e.g. 三三行, 萬萬慈)
NND	noun	AABB (e.g. 般般件件)
NNE	noun	ABAB (e.g. 一對一對)
NNF	noun	ABAC (e.g. 僮男僮女)
NNG	noun	ABB (e.g. 一層層, 汗珠珠)
NNH	noun	ABCB (e.g. 千世萬世)
NNI	noun	honorific (e.g. 貴庚, 仙鄉)
NNJ	noun	proper (e.g. 黃巾)
NNK	noun	proper_dynasty_name (e.g. 春秋戰國)
NNL	noun	proper_person_name (e.g. 蘧伯玉)
NNM	noun	proper_place_name (e.g. 黃山)
NNN	noun	proper_title (e.g. 孫子兵法)
NNO	noun	proper_year_name (e.g. 天章)
NMA	numeral	cardinal (e.g. 十八, 千)

(continued)

Appendix I Continued

Tag label	Word class	Category
NMB	numeral	indefinite (e.g. 數十, 幾百)
NMC	numeral	ordinal (e.g. 第一, 第八)
ONA	onomatopoeia	AA (e.g. 哇哇, 嘻嘻)
ONB	onomatopoeia	AAA (e.g. 騰騰騰, 撒撒撒)
ONC	onomatopoeia	AABB (e.g. 隱隱轟轟)
OND	onomatopoeia	ABBC (e.g. 撲通通冬, 吉丁丁瑞)
ONE	onomatopoeia	general (e.g. 耶穌咿啞)
PNA	pronoun	demonstrative (e.g. 這, 其)
PNB	pronoun	honorific (e.g. 寡人, 在下)
PNC	pronoun	personal (e.g. 我們, 俺)
PND	pronoun	possessive (e.g. 我的, 厥)
PNE	pronoun	reciprocal (e.g. 彼此)
PNF	pronoun	reflexive (e.g. 自己)
PPA	preposition	(e.g. 據, 至於)
PRA	particle	tag (e.g. 吧, 乎)
PTA	punctuation	general_separating_mark ('◦', '!', '?', ',', ':')
PTB	punctuation	left_bracket (e.g. 『, 〈, or 「)
PTC	punctuation	right_bracket (e.g. 』, 〉, or 」)
PTD	punctuation	secondary_separating_mark (e.g. '·', ' ')
QWA	question_word	general (e.g. 爲何, 甚麼)
QWB	question_word	tag (e.g. 麼)
UND	unidentified	(e.g. □)
VBA	verb	general (e.g. 剮, 頂)
VBB	verb	AA (e.g. 演演, 走走)
VBC	verb	AAB (e.g. 散散心)
VBD	verb	AABB (e.g. 哭哭啼啼)
VBE	verb	ABAB (e.g. 接待接待)
VBF	verb	ABAC (e.g. 包長包短)
VBG	verb	ABB (e.g. 哭啼啼)
VBH	verb	ABCB (e.g. 手之舞之)
VBI	verb	bei_V (e.g. 被戮)
VBJ	verb	bu_V (e.g. 不宜)
VBK	verb	copular_shi (e.g. 是)
VBL	verb	copular_shi _negative (e.g. 不是)
VBM	verb	existential_you (e.g. 有)
VBN	verb	existential_you_negative (e.g. 未有)
VBO	verb	jian_V (e.g. 見信, 見教)
VBP	verb	modal_auxiliary (e.g. 必, 該)
VBQ	verb	modal_auxiliary_negative (e.g. 不必, 不該)
VBR	verb	reciprocal_xiang_V (e.g. 相會, 相辭)
VBS	verb	reflexive_zi_V (e.g. 自寬, 自縊)
VBT	verb	stative (e.g. 惆悵, 廣厚)
VBU	verb	stative_comparative (e.g. 更深)
VBV	verb	stative_superlative (e.g. 最早)
VBW	verb	suo_V (e.g. 所積, 所吟)
VBX	verb	V_bu_V (e.g. 念不念, 定不定)
VBY	verb	V_hua (e.g. 羽化)
VBZ	verb	V_lai (e.g. 宣來, 討來)
VBAA	verb	V_N (e.g. 守法, 聽話)
VBBB	verb	V_potential_bu_RVC* (e.g. 睡不穩)
VBCC	verb	V_potential_de_RVC (e.g. 躲得過)
VBDD	verb	V_qu (e.g. 消去, 拿去)

(continued)

Appendix I Continued

Tag label	Word class	Category
VBEE	verb	V_RVC (e.g. 學成, 生出)
VBFF	verb	V_V (e.g. 敘說, 思慮)
VBGG	verb	V_yi_V (e.g. 晝一晝, 嘗一嘗)
VBHH	verb	V_yu (e.g. 起於)
VBII	verb	V_zhi (e.g. 刑之)
VBJJ	verb	yi_V (e.g. 一訪, 一望)