

**Zizhi Tongjian as Data:
From Historical Chronicle to Monthly Computational Corpus**

September 2026
ISS Discussion Paper Series
F-203

Asei Ito^{*}
Jaehwan Lim[†]
Hongyong Zhang[‡]

^{*} Institute of Social Science, The University of Tokyo, Japan. asei@iss.u-tokyo.ac.jp

[†] School of International Politics, Economics, and Communication, Aoyama Gakuin University, Tokyo, Japan. t13025@aoyamagakuin.jp.

[‡] Research Group, The Research Institute of Economy, Trade, and Industry, Tokyo, Japan. zhang-hong-yong@rieti.go.jp.

Zizhi Tongjian as Data:

From Historical Chronicle to Monthly Computational Corpus

Version 1.0

Asei Ito

Institute of Social Science, The University of Tokyo, Tokyo, Japan

Jaehwan Lim

School of International Politics, Economics, and Communication, Aoyama Gakuin University, Tokyo, Japan

Hongyong Zhang

Research Group, The Research Institute of Economy, Trade, and Industry, Tokyo, Japan

Abstract

How can a continuous premodern Chinese chronicle be transformed into temporally structured, quantitatively usable data? Using the *Zizhi Tongjian* (資治通鑑, Comprehensive Mirror to Aid in Government), this paper develops a transparent and auditable workflow for reconstructing annual and monthly chronology and discovering recurrent lexical units through a corpus-internal procedure rather than a pretrained tokenizer. The main lexical-discovery analysis covers Tang China, 618–907 CE. It combines character n-grams of length two to four with corpus-internal measures of internal cohesion (minimum-split pointwise mutual information) and boundary freedom (left and right entropy), and then counts accepted units in a monthly temporal-lexical corpus. The procedure yields a final Tang lexicon of 6,528 recurrent multi-character units. In an LLM-classified validation sample, 97.6% of candidates in the top lexical-score decile are valid units; an independent known-vocabulary check recovers 46 of 61 items (75.4%). A separate blind human reliability check yields 84.0% exact human–LLM agreement for lexical classification and 93.3% for chronology cases. Monthly series also recover historically interpretable patterns, including the transition from Huihe (回紇) to Huigu (回鶻), two written forms of the Uyghur name, around 788 and month-specific peaks in Xuanwumen-related vocabulary around documented palace coups. A Song-period portability check using the *Xu Zizhi Tongjian Changbian* (續資治通鑑長編, Long Draft of the Continuation of the Comprehensive Mirror to Aid in Government) indicates that high-frequency Tang units remain attested while their relative salience changes substantially.

1. Introduction

The quantitative and digital study of Chinese history has developed substantially in recent years. China’s unusually rich documentary record has long supported quantitative inquiry, while digitization, database construction, and computational methods have widened the scale and range of feasible research (Chen, Campbell, and Ma 2026; Keller, Shiue, and Yan 2026; Bai, Jia, and Yang 2024). The China Biographical Database (CBDB) is a leading example: it extracts biographical information from heterogeneous primary and secondary sources, disambiguates and normalizes persons and relationships, and organizes persons, offices, places, kinship ties, social relations, and other entities in a relational structure for statistical, spatial, and network analysis (Tsui and Wang 2020; Chen and Wang 2022). Related work in historical political economy likewise demonstrates the value of constructing systematic evidence from Chinese historical sources. Wang and Yang (2025) combine epitaphs, government documents, and large-scale biographical information to study the long-run development of the imperial examination system. Keller, Shiue, and Yan (2026) use supervised machine learning to identify and geographically and temporally tag episodes of social unrest in the Veritable Records of the Qing Dynasty (清實錄). These projects represent substantial scholarly and technical achievements in converting historical records into quantitative evidence.

At the same time, historical data construction remains source-dependent. CBDB’s developers explicitly discuss source selection, automated and manual extraction, disambiguation, normalization, and the uneven representation of different activities in the surviving record; its holdings are rich in officeholding and kinship, for example, but sparse in economic activity (Chen and Wang 2022). Related annotation environments such as MARKUS also show how dictionaries, entity identification, and scholarly curation can make primary texts computationally tractable (De Weerd 2020). These approaches motivate a broader methodological lesson: source structure shapes dataset structure. A continuous narrative chronicle presents a different data-construction problem because the temporal and lexical units required for analysis may not be directly available as machine-readable observations. Digitization makes such a text searchable, but does not automatically make it analysis-ready.

The *Zizhi Tongjian* (資治通鑑, Comprehensive Mirror to Aid in Government; hereafter ZZTJ) is a useful case through which to examine this problem. Covering Chinese history from 403 BCE to 959 CE, it is organized annalistically. Its digital Classical Chinese text remains a continuous, unsegmented stream of characters. Dates appear within the narrative rather than as fields attached to passages, and the absence of whitespace means that recurrent lexical units are not directly marked. Quantitative use of the chronicle therefore requires the reconstruction of two related forms of structure: when a passage belongs in historical time, and which recurrent character sequences constitute units that can be counted through that time. These choices determine the observations, temporal resolution, and textual measures available for subsequent historical analysis.

For this specific source-construction task, we develop one transparent workflow. The main analysis focuses on Tang China (618–907 CE), while the *Xu Zizhi Tongjian Changbian* (續資治通鑑長編, Long Draft of the Continuation of the Comprehensive Mirror to Aid in Government; hereafter XZZTJ), covering 960–1100 CE in the corpus used here, provides a Song-period portability check. The analysis begins in 618 because the earlier portions of the ZZTJ are generally thinner and more uneven at monthly resolution, making them less suitable for a comparable long-run lexical panel; those earlier portions remain available for future source-specific analysis. The workflow reconstructs annual and monthly chronology from the continuous narrative, leaving genuinely ambiguous spans unassigned rather than interpolating them. It then identifies recurrent lexical units through source-internal character statistics, combining frequency, minimum-split pointwise mutual information (PMI), and left- and right-boundary entropy. Finally, the accepted units are counted back into the reconstructed calendar to produce a term-

by-year-by-month corpus, and the chronology and lexical output are evaluated through LLM-reviewed sampling audits, a separate author-conducted blind reliability check, known-vocabulary recovery, and historically documented cases.

This workflow is intended to complement existing work in digital and quantitative Chinese history as well as Classical Chinese NLP. The present study focuses on a continuous annalistic chronicle in which passage-level chronology and recurrent lexical units must be reconstructed together. Classical Chinese word-segmentation systems and pretrained language models offer important, complementary approaches to related tasks (Wang 2020; Tian et al. 2021; Wang et al. 2023). Our corpus-internal procedure is not proposed as a general-purpose segmentation algorithm or a replacement for those systems; it deliberately prioritizes inspectability for this application, so that each candidate can be traced to observed frequency, cohesion, and boundary statistics. In this respect, it also builds on the demonstrated utility of character-sequence methods in large digitized corpora and premodern textual analysis (Rappo 2018).

The paper’s contribution lies less in any single computational technique than in making the full process of historical data construction explicit and auditable. The case illustrates how source-specific historical knowledge and computational procedures can be combined in historical data construction. The units, temporal resolution, missing observations, ambiguities, and representations in a historical dataset reflect methodological choices as well as properties of the source. Making those choices explicit and auditable is therefore useful both for evaluating this corpus and for considering related workflows for other premodern chronicles.

The remainder of the paper proceeds as follows. Section 2 introduces the ZZTJ as a computational historical source. Section 3 describes chronology reconstruction and corpus coverage. Section 4 presents corpus-internal lexical discovery. Section 5 constructs the monthly temporal-lexical corpus. Section 6 illustrates historical applications. Section 7 reports validation and robustness. Section 8 concludes.

2. The *Zizhi Tongjian* as a Computational Historical Source

The ZZTJ, compiled under Sima Guang and completed in 1084, narrates Chinese history from 403 BCE to 959 CE in chronological (annalistic) form. Events are generally recorded under reign-years rather than grouped by topic or person, as in the standard dynastic histories. This organization makes the text especially useful as a computational source because it provides an underlying calendrical sequence, although that sequence must still be reconstructed from the narrative before temporal analysis. The XZZTJ extends a related annalistic form into the Song dynasty (960–1100 CE, within the scope used here) and is used here as an out-of-sample check on the portability of the lexicon discovered from the Tang corpus. The digital text underlying this project is drawn from the Kanseki Repository (漢籍リポジトリ), maintained by the Institute for Research in Humanities, Kyoto University. Content created by the repository is made available under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license.

Figure 1 illustrates the relationship between the page image of the ZZTJ and its machine-readable transcription in the Kanseki Repository. The passage records that, in the fifth month of 726, the Ministry of Revenue reported 7,069,565 registered households and a population of 41,419,712. In the seventh month, severe flooding across the regions north and south of the Yellow River drowned thousands of people. On the first day of the eighth month, Wei Prefecture reported that the Yellow River had overflowed its banks. In little more than sixty Chinese characters, the chronicle thus records a nationwide fiscal-demographic report, a large-scale flood with a mortality estimate, and a dated local report of a Yellow River overflow. The example illustrates the density of temporally structured historical

information contained in the source and the potential value of converting the chronicle into a machine-readable temporal corpus.

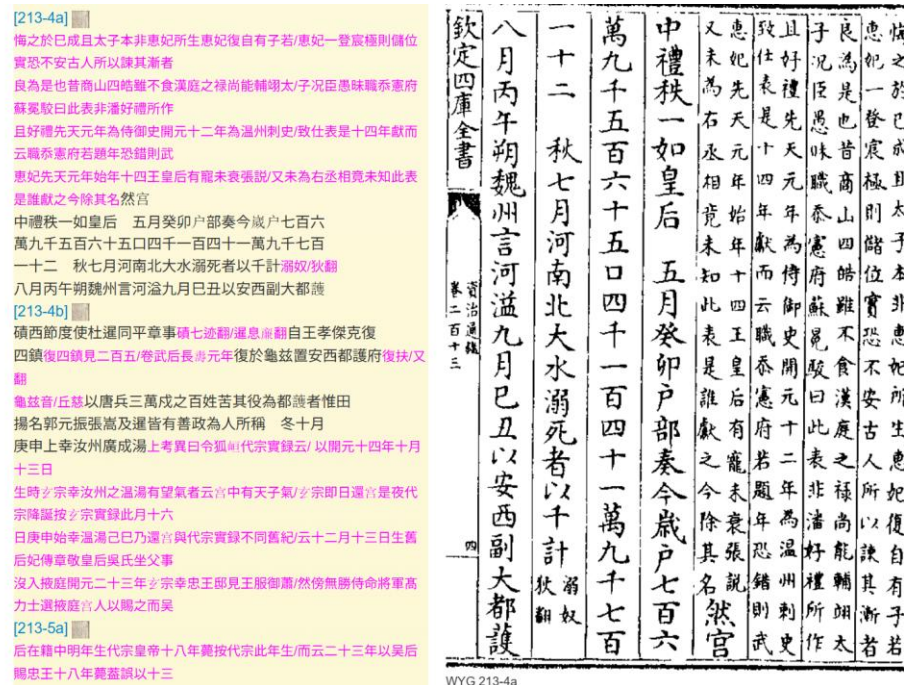


Figure 1. Historical page image and machine-readable transcription of the *Zizhi Tongjian*, 726 CE

Source: *Zizhi Tongjian*, juan 213, Kanseki Repository, KR2b0007. Electronic transcription from the Kanseki Repository, with corresponding page image from the *Wenyuange Siku Quanshu* edition. Kanseki Repository content created by the repository is licensed under CC BY-SA 4.0 (<https://www.kanripo.org/text/KR2b0007/213#>).

A feature of the source is that chronological coverage is uneven across months. As the coverage statistics presented below will show, not every year contains separately identifiable sections for all twelve months. This does not imply that months without an identifiable boundary necessarily represent missing text. For example, in 734 CE (*Kaiyuan* 22), the chronicle explicitly marks the first through eighth months and then the twelfth month, while providing an extensive continuous narrative for the year. The record includes, among other events, an earthquake in Qinzhou that reportedly killed more than 4,000 people, a court debate over private coinage, changes in senior appointments, military activity involving the Khitan, and detailed reforms to grain transport and storage. The following passage shows the chronicle's continuous record for 734 CE, with the identifiable month markers highlighted in bold and grey.

開元二十二年**春正月**己巳上發西京己丑至東都張九齡自韶州入見求終喪不許**二月**壬寅秦州地震震壞公私屋殆盡吏民壓死者四千餘人命左丞相蕭嵩賑恤方士張果自言有神仙術誑人云堯時為侍中於今數千歲多往來恒山中則天以來屢徵不至相州刺史韋濟薦之上遣中書舍人徐嶠齎璽書迎之庚寅至東都肩輿入宮恩禮甚厚張九齡請不禁鑄錢**三月**庚辰敕百官議之裴耀卿等皆曰一啟此門恐小人棄農逐利而濫惡更甚祕書監崔沔曰若稅銅折役則官冶可成計估度庸則私鑄無利

易而可久簡而難誣且夫錢之為物貴以通貨利不在多何待私鑄然後足用也右監門錄事參軍劉秩曰夫人富則不可以賞勸貧則不可以威禁若許其私鑄貧者必不能為之臣恐貧者益貧而役於富富者益富而逞其欲漢文帝時吳王濞富埒天子鑄錢所致也上乃止秩子玄之子也夏四月壬辰以朔方節度使信安王禕兼關內道採訪處置使增領涇原等十二州吏部侍郎李林甫柔佞多狡數深結宦官及妃嬪家伺候上動靜無不知之由是每奏對常稱旨上悅之時武惠妃寵幸傾後宮生壽王清諸子莫得為比太子浸疎薄林甫乃因宦官言於惠妃願盡力保護壽王惠妃德之陰為內助由是擢黃門侍郎五月戊子以裴耀卿為侍中張九齡為中書令林甫為禮部尚書同中書門下三品上種麥於苑中帥太子以下親往芟之謂曰此所以薦宗廟故不敢不親且欲使汝曹知稼穡艱難耳又偏以賜侍臣曰比遣人視田中稼多不得實故自種以觀之六月壬辰幽州節度使張守珪大破契丹遣使獻捷薛王業疾病上憂之容髮為變七月己巳薨贈諡惠宣太子上以裴耀卿為江淮河南轉運使於河口置輸場八月壬寅於輸場東置河陰倉西置柏崖倉三門東置集津倉西置鹽倉鑿漕渠十八里以避三門之險先是舟運江淮之米至東都含嘉倉僦車陸運三百里至陝率兩斛用十錢耀卿令江淮舟運悉輸河陰倉更用河舟運至含嘉倉及太原倉自太原倉入渭輸關中凡三歲運米七百萬斛省僦車錢三十萬緡或說耀卿獻所省錢耀卿曰此公家贏縮之利耳奈何以之市寵乎悉奏以為市糴錢張果固請歸恒山制以為銀青光祿大夫號通玄先生厚賜而遣之後卒好異者奏以為尸解上由是頗信神仙冬十二月戊子朔日有食之乙巳幽州節度使張守珪斬契丹王屈烈及可突干傳首時可突干連年為邊患趙含章薛楚玉皆不能討守珪到官屢擊破之可突干困迫遣使詐降守珪使管記王悔就撫之悔至其牙帳察契丹上下殊無降意但稍徙營帳近西北密遣人引突厥謀殺悔以叛悔知之牙官李過折與可突干分典兵馬爭權不叶悔說過折使圖之過折夜勒兵斬屈烈及可突干盡誅其黨帥餘眾來降守珪出師紫蒙川大閱以鎮撫之梟屈烈可突干首于天津之南突厥毗伽可汗為其大臣梅錄啜所毒未死討誅梅錄啜及其族黨既卒子伊然可汗立尋卒弟登利可汗立庚戌來告喪禁京城旬者置病坊以廩之 (1036 characters; Source: Sima Guang, Zizhi Tongjian, juan 214, Kaiyuan 22 [734 CE], Kanseki Repository, KR2b0007)

Thus, the absence of a separately identifiable month does not necessarily indicate missing source text; it may instead reflect the temporal organization of the chronicle itself.

Two data-quality issues specific to this source are handled as follows. First, some characters are represented in the source digitization by literal gajii placeholder codes (e.g., &KRnnnn;). These placeholders are detected and removed before the calculation of frequencies, cohesion, or entropy, rather than being treated as ordinary characters. Second, the corpus's orthography reflects its textual transmission and digitization as well as the language of the periods it describes. For example, the standard form 玄武門 (*Xuanwumen*, “Gate of the Black Tortoise”), examined further in Section 6, does not occur in the corpus. The Kanripo texts instead consistently use 玄武門, with the variant character 玄 in place of 玄. The same substitution occurs in other words containing 玄. This study therefore treats 玄 as an edition-level orthographic variant and normalizes it to 玄 only in the English discussion and figure labels. The substitution is compatible with the avoidance of 玄 associated with the personal name of the Kangxi Emperor, Xuanye 玄燁, but the present computational evidence alone does not establish when or at which stage of transmission the substitution was introduced. It should therefore not be interpreted as evidence of Tang- or Song-period spelling. This example illustrates why source-specific orthography matters for lexical discovery: a dictionary or tokenizer based exclusively on normalized spellings would not identify the attested form without orthographic normalization or adaptation.

Across the twelve-cell analytical calendar, the reconstructed corpus contains 5,170,451 characters: 1,013,273 from the ZZTJ and 4,157,178 from the XZZTJ. Intercalary months are not treated as thirteenth analytical cells. Instead, their text is grouped with the corresponding numbered month in the

twelve-cell analytical panel, while the intercalary flag, text, and character count are retained separately. The corpus contains 72,682 characters identified as belonging to intercalary months (1,939 in the ZZTJ and 70,743 in the XZZTJ), allowing the full chronological sequence to be reconstructed when required. This scale permits corpus-statistical lexical discovery while remaining suitable for direct validation of sampled candidates.

3. Reconstructing Annual and Monthly Chronology

Neither source carries machine-readable date fields. Years are marked within the narrative by reign-era names and year counts, while months are indicated by phrases such as “春正月” (spring, first month) or “閏 X 月” (intercalary month X). These chronological markers are embedded in continuous prose alongside retrospective references, cross-year citations, and, in some cases, multiple narrative threads. Recovering calendar structure is therefore a problem of historical text reconstruction rather than simple reformatting.

The reconstruction proceeds in two stages. First, the continuous text is segmented into chronicle-year units, using reign-era names, year counts, and associated calendrical evidence to identify annual boundaries and connect the relevant textual spans. Second, each reconstructed year is segmented into calendar months. This stage combines explicit month-boundary phrases with calendrical cross-checks and rules for handling retrospective references, cited material, and secondary narrative threads. A contextual audit revealed a source-specific feature of the XZZTJ: one reconstructed year may span several continuation volumes, each introduced by a compiler heading containing the emperor, reign era, reign-year number, and month. An earlier rule could misclassify headings beyond the beginning of the annual record as embedded year citations. The revised parser recognizes these continuation-volume headings when they agree with the current annual record after conservative normalization of attested graphic variants. No month identified through the parser audit is hard-coded; the four separately documented Tang annual-boundary repairs are summarized in Appendix Table C1. When the evidence does not support confident assignment, the procedure leaves the cell unassigned rather than interpolating. Cells within the documented XZZTJ lacuna from the fourth month of 1067 through the third month of 1070 are labeled `source_lacuna` rather than treated as parser failures.

Figure 2 illustrates this transformation using 734 CE (*Kaiyuan* 22) as an example. The source presents the year as a continuous historical narrative rather than as a set of monthly observations. The reconstruction procedure identifies explicit month boundaries and assigns the intervening text to the corresponding calendar month, transforming the annual narrative into a structured year–month–text dataset. In this case, nine months can be separately reconstructed: months 1–8 and 12. Months 9–11 remain unassigned rather than being inferred or filled by interpolation. Importantly, these unassigned cells should not be interpreted as evidence that the source contains no historical information pertaining to those months; they indicate only that separate monthly units cannot be confidently reconstructed under the stated procedure. The transformation thus preserves both the text associated with identifiable months and the uneven temporal structure of the source, producing the basic observational unit from which the subsequent monthly lexical corpus is constructed.

Table 1. Corpus coverage and density after chronology reconstruction

Measure	ZZTJ	ZZTJ, Tang	XZZTJ
Year range	618–959	618–907	960–1100
Represented years	342	290	136
Possible month cells	4,104	3,480	1,632
Reconstructed month cells	3,314	2,724	1,585
Coverage	80.8%	78.3%	97.1%
Mean reconstructed months/year	9.69	9.39	11.65
Median reconstructed months/year	10	9	12
Total characters	1,013,273	762,516	4,157,178
Mean characters/year	2,962.8	2,629.4	30,567.5
Median characters/year	2,169.5	1,814.0	16,802.0

Note: “Possible month cells” refers to the twelve-cell analytical calendar (months 1–12). Intercalary-month text is included in the corresponding numbered month for this analytical panel. Intercalary status, text, and character counts are additionally retained in separate fields (ZZTJ: 1,939 characters; XZZTJ: 70,743 characters), allowing intercalary months to be recovered as separate chronological units when required. “Represented years” counts years present in the attached monthly corpus; gaps within a source’s stated range are not imputed. Annual characters are the sum of `n_char_month` across the twelve-cell analytical panel for each represented year; means and medians are calculated across represented years.

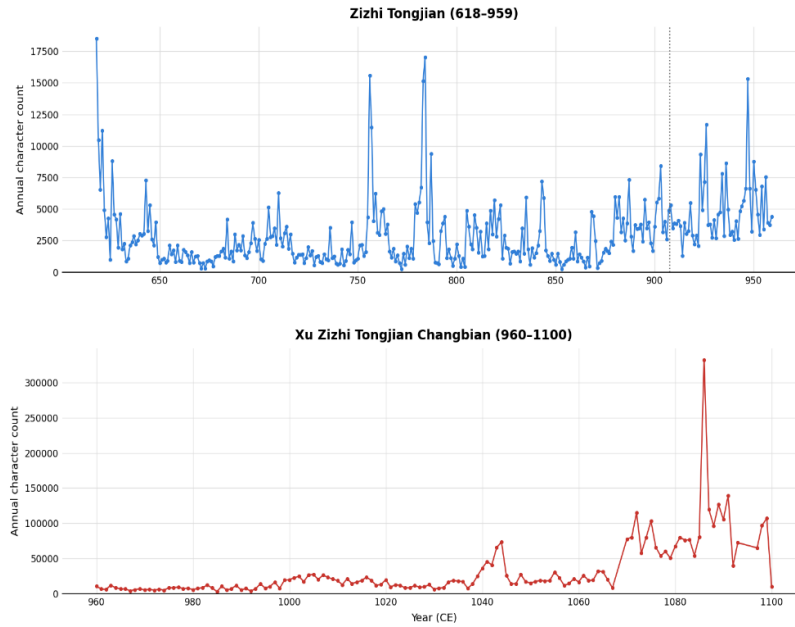


Figure 3. Annual character counts in the ZZTJ and XZZTJ

Note: Annual counts sum reconstructed characters assigned to the twelve-cell analytical panel; intercalary-month text is included in the corresponding numbered month. Intercalary status, text, and character counts remain separately recoverable. The dotted line marks 907, the end of the Tang period. ZZTJ years 908–959 are retained for coverage assessment but excluded from Tang lexical discovery.

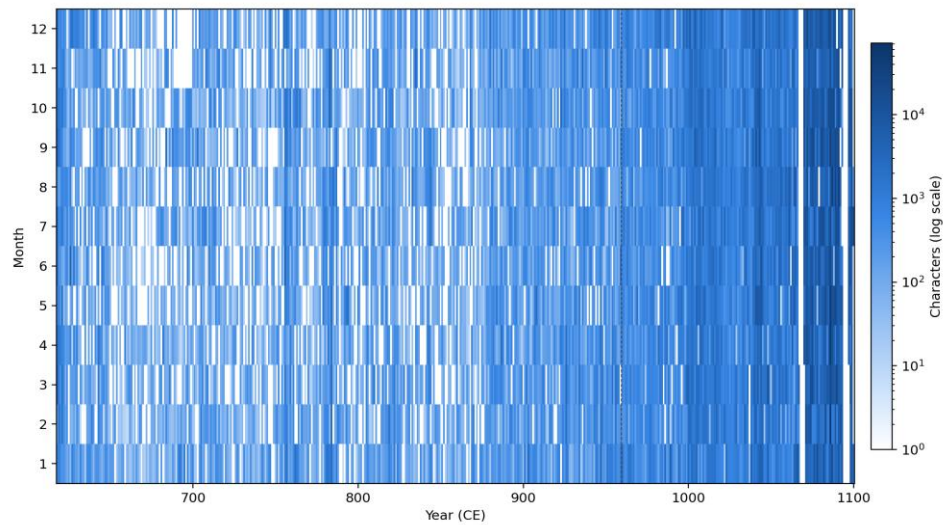


Figure 4. Monthly coverage of the chronicles, by character count

Note: Each cell is one chronicle-year-month; color records assigned character count on a logarithmic scale. White cells have zero assigned characters. Full-height white bands at 1068–1069 and 1094–1096 are years absent from the supplied XZZTJ dataset. The transmitted XZZTJ lacuna from 1067/4 through 1070/3 is labeled `source_lacuna` in the data. ZZTJ year 853 remains unusually sparse (three

reconstructed months; 241 characters). Other white cells may reflect source silence or unresolved boundaries. The dashed line marks the source boundary at 959/960.

The chronology reconstruction and its iterative validation are reported in Section 7.1 and Appendix Tables A1–A2.

4. Tokenizer-Free Lexical Discovery in Classical Chinese

Modern Chinese tokenizers are not specifically designed for the vocabulary and register of the ZZTJ, while Classical Chinese segmentation systems and pretrained models offer complementary approaches that introduce their own modeling choices and training resources. For the source-construction task pursued here, we adopt a corpus-internal approach based on character co-occurrence statistics, prioritizing the inspectability of its acceptance criteria. Lexical discovery is restricted to the Tang subcorpus (618–907 CE, 762,516 reconstructed characters in the twelve-cell analytical panel), so that it operates on a single historical span rather than pooling across dynasties.

Statistical approaches to unsupervised Chinese word discovery have long used two complementary types of distributional evidence: association among characters within a candidate string and diversity in the contexts surrounding it. Character association, including mutual-information-based measures, provides evidence of internal cohesion, while accessor variety and branching or boundary entropy measure the extent to which a string occurs independently across different external contexts (Feng et al. 2004; Jin and Tanaka-Ishii 2006). Building on this logic, we treat internal cohesion and boundary freedom as complementary criteria for identifying recurrent lexical units in the ZZTJ corpus.

Candidate units are all contiguous character n -grams, for $n = 2, 3, 4$, that actually occur in the reconstructed monthly text. Each candidate is scored on two independent criteria. Internal cohesion is measured by pointwise mutual information (PMI), taken as the minimum PMI over every possible binary split of the candidate – a conjunctive requirement that every internal boundary, not just one favorable one, shows above-chance character co-occurrence. Boundary freedom is measured by left- and right-context character entropy, taking the minimum of the two sides; a recurrent lexical unit is expected to occur across varied external contexts while remaining internally cohesive. The two criteria are combined into a single `lexical_score`, computed as the minimum of each measure’s percentile rank within its own n -gram length, so that a candidate receives a high combined score only when both cohesion and boundary freedom rank highly.

Four thresholds turn scores into a lexicon: a frequency floor (occurring at least 10 times), a PMI floor (minimum-split PMI above 3), an entropy floor (minimum boundary entropy above 1.0), and a final pass that removes shorter fragments largely accounted for by accepted longer units. Table 2a reports the rerun on the revised monthly data: of 1,527,124 raw candidate types ($n = 2$ –4 pooled), 14,844 clear the frequency floor, 8,686 also clear the PMI floor, and 6,555 also clear the entropy floor, of which 6,528 survive fragment removal into the final Tang lexicon (4,466 bigrams, 1,644 trigrams, and 418 four-character units). Threshold sensitivity is reported in Appendix Table A3.

Table 2a. Tang lexicon assembly funnel

Stage	Types remaining
Raw candidate types ($n = 2$ –4, pooled)	1,527,124
+ frequency ≥ 10	14,844

Stage	Types remaining
+ minimum-split PMI > 3	8,686
+ minimum boundary entropy > 1.0	6,555
+ redundant-fragment removal (final lexicon)	6,528

Table 2b provides concrete examples of how the two statistics differentiate candidate strings. Candidates occurring at least ten times were ranked separately, within each n-gram length, by minimum-split PMI and minimum boundary entropy. High and low correspond to the top and bottom deciles of the within-length distribution, while middle refers to the 47.5th–52.5th percentiles. Within each stratum, the five most frequent candidates are reported, with the raw score shown in parentheses. The examples illustrate that the two measures capture different properties of recurrent character sequences.

Minimum-split PMI captures internal cohesion. High-PMI candidates such as 節度, 刺史, 陛下, 侍郎, and 吐蕃 consist of character combinations that co-occur much more strongly than would be expected from the marginal frequencies of their component characters. By contrast, low-PMI strings such as 之以, 之為, 之人, and 使之 are frequent combinations formed largely from common function characters and show little evidence of behaving as internally cohesive units. Middle-ranked candidates occupy an intermediate zone in which some combinations may be meaningful in particular contexts but are less strongly bound internally.

Minimum boundary entropy captures contextual freedom rather than internal association. Candidates with high boundary entropy, such as 節度使, 州刺史, 陛下, and 太子, occur with a relatively diverse set of characters on both their left and right boundaries, which is consistent with their functioning as reusable lexical units across different contexts. Low-entropy strings such as 度使, 州刺, 平章, 章事, and 同平, by contrast, tend to occur as recurrent internal fragments of longer expressions and have highly restricted boundary environments. Their low entropy therefore helps distinguish productive lexical units from substrings that recur mainly because they are embedded within a larger fixed expression.

The two statistics are therefore complementary rather than interchangeable. PMI asks whether the characters inside a candidate belong together; boundary entropy asks whether the candidate can occur as a relatively independent unit across varied external contexts. A candidate that scores highly on only one dimension may still be a fragment or context-specific sequence. For this reason, the final lexical score is deliberately conjunctive: a candidate receives a high combined score only when it performs well on both internal cohesion and boundary freedom. The procedure should not be interpreted as identifying “words” in an absolute linguistic sense. Rather, it provides a transparent way to rank recurrent multi-character sequences by the extent to which they behave like reusable lexical units in this corpus.

For this implementation, lexical discovery does not rely on a seed dictionary or pretrained segmentation model. Human judgment instead enters through the choice of n-gram lengths, frequency floor, PMI and entropy thresholds, and fragment-removal rules. Their implications are evaluated through the LLM-classified sample, a separate blind human reliability check, known-vocabulary recovery, and threshold-sensitivity checks reported in Section 7. Appendix B provides an intuitive illustration of how internal cohesion and boundary freedom distinguish reusable lexical units from recurrent fragments.

Table 2b. Illustrative candidates across PMI and boundary-entropy strata

Measure	Stratum	Example 1	Example 2	Example 3	Example 4	Example 5
Minimum-split PMI	High	節度 (7.74)	刺史 (8.16)	陛下 (7.72)	侍郎 (8.48)	吐蕃 (9.57)
Minimum-split PMI	Middle	其將 (2.81)	使者 (2.81)	也上 (2.94)	人皆 (2.81)	詔以 (2.95)
Minimum-split PMI	Low	之以 (0.09)	為節度 (2.21)	之為 (-0.99)	之人 (-0.17)	使之 (-0.44)
Minimum boundary entropy	High	節度使 (5.61)	州刺史 (7.62)	以為 (7.36)	陛下 (6.75)	太子 (6.96)
Minimum boundary entropy	Middle	尚書 (3.07)	大將 (3.07)	遣其 (3.05)	部侍郎 (2.64)	門下 (3.07)
Minimum boundary entropy	Low	度使 (0.01)	州刺 (0.03)	平章 (0.13)	章事 (0.02)	同平 (0.03)

5. Building a Monthly Temporal-Lexical Corpus

The resulting Tang lexicon is then counted back against the reconstructed monthly text: for every lexicon entry and every chronicle-year x month cell, the pipeline records how many times that entry occurs, and normalizes the count per 10,000 characters of that month’s surviving text. This normalization matters because, as Figures 3–4 show, the amount of reconstructed text varies substantially across years and months; a raw occurrence count would largely track how much text is represented in a given cell rather than only changes in the term’s relative salience. The result is a 178,443-row long-format panel (term x year x month x frequency x per-10,000-character rate), which is the paper’s principal data product. All visualizations in Section 6 use this common representation.

6. Visualizing Historical Change

Three figures illustrate what the resulting lexicon and monthly chronology together support, each presenting a distinct use of the corpus.

Figure 5 plots annual frequencies per 10,000 characters for the eight most frequent entries in the final Tang lexicon. The terms display markedly different temporal profiles. 節度使 (jiedushi, military governor) rises strongly during the later Tang, while 將軍 (jiangjun, general) declines; 太子 (taizi, crown prince) is characterized by episodic spikes; and 吐蕃 (Tubo, the Tibetan Empire) is concentrated in periods of intense Tang-Tibetan interaction. These contrasting profiles show temporal variation after normalization for corpus volume and provide historically interpretable examples of the resulting annual series.

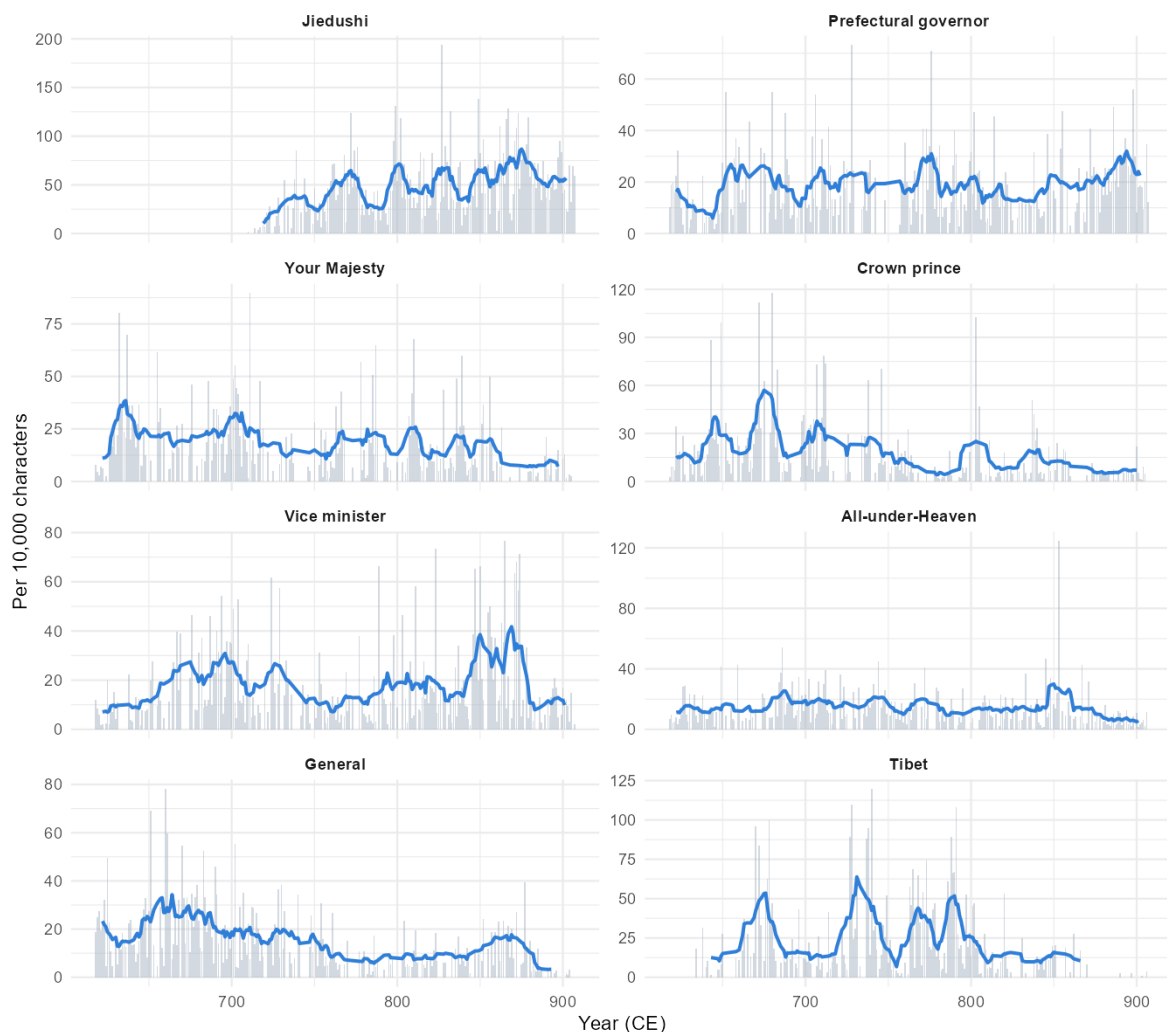


Figure 5. Annual frequencies of the eight most frequent Tang lexical units

Note: Bars show annual values and the line is a centered ten-year moving average. Frequencies are normalized per 10,000 reconstructed characters. The eight terms are selected mechanically as the most frequent entries in the final Tang lexicon. English labels correspond to the following original Chinese terms: Jiedushi (節度使, military governor), Prefectural governor (州刺史), Your Majesty (陛下), Crown prince (太子), Vice minister (侍郎), All-under-Heaven (天下), General (將軍), and Tibet (吐蕃, Tubo).

Figure 6 examines two Chinese transliterations of the same Inner Asian polity's name (the Uyghurs), whose distribution is historically patterned: the *Xin Tangshu* (新唐書), juan 217A, dates the court's requested change from 回紇 to 回鶻 to 788 CE, whereas the *Jiu Tangshu* (舊唐書), juan 195,

places the episode in 809 CE.¹ In this corpus, 回紇’s last attested year and 回鶻’s first attested year both fall in 788 – a pronounced lexical shift concentrated at the earlier, disputed date, though not an absolute cutover. A handful of 回紇 mentions persist afterward and may reflect retrospective reference, quotation, or orthographic persistence. The transition aligns closely with the 788 dating reported in the *Xin Tangshu*, although the purpose here is not to adjudicate between competing historical accounts.

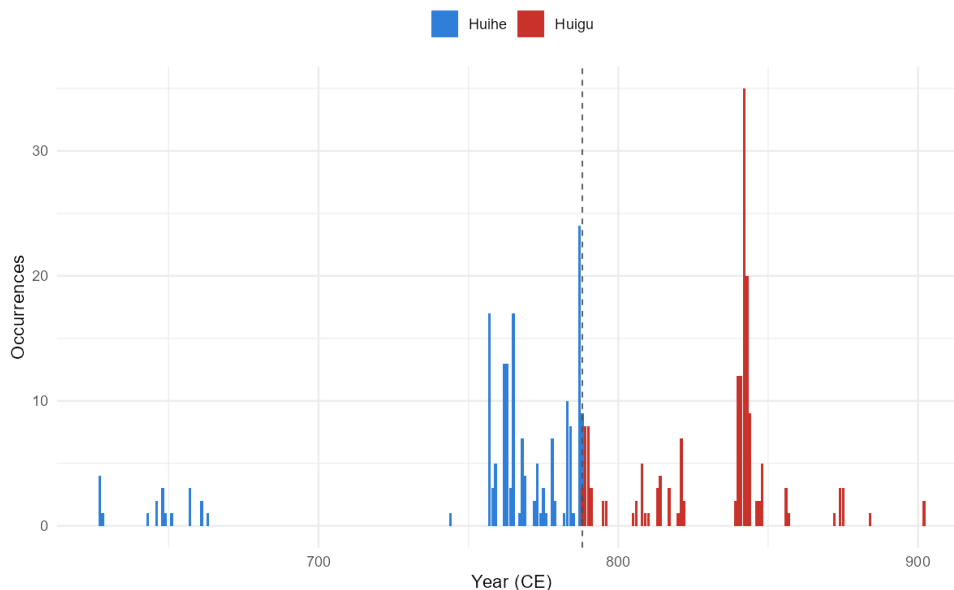


Figure 6. Annual occurrences of Huihe (回紇) and Huigu (回鶻), 618–907

Note: Bars show annual occurrences in the ZZTJ Tang corpus. The dashed line marks 788 CE, the dating in the *New Book of Tang* of the Uyghurs’ request to change the written form from 回紇 to 回鶻; the *Old Book of Tang* dates the request to 809 CE.

Figure 7 examines 玄武門 (*Xuanwumen*, conventionally written 玄武門; “Gate of the Black Tortoise”). The standard spelling 玄武門 does not occur in the corpus: the Kanripo edition consistently uses the variant character 𪚩 in place of 玄, as discussed in Section 2. This study normalizes 𪚩玄武門 as 玄武門 in the English discussion and figure labels for readability. *Xuanwumen* does not designate a single location. The name follows a conventional directional scheme in which the northern gate of a palace complex was named after *Xuanwu* 玄武, the Black Tortoise associated with the north; consequently, several unrelated Sui- and Tang-period palaces had gates bearing this name. All occurrences in the analyzed corpus fall between 618 and 710. Every month containing two or more mentions corresponds to one of the four annotated events listed in Table A7: the Jiangdu Mutiny (618/3), including the assassination of Emperor Yang of Sui at the Jiangdu palace’s *Xuanwumen*; the

¹ *Xin Tangshu* 新唐書, juan 217A, “Huihu shang” 回鶻上; *Jiu Tangshu* 舊唐書, juan 195, “Huihe” 回紇. Electronic texts available via 維基文庫 (Wikisource) and 中國哲學書電子化計劃 (Chinese Text Project): *Xin Tangshu*, juan 217A (<https://zh.wikisource.org/zh-hant/%E6%96%B0%E5%94%90%E6%9B%B8/%E5%8D%B7217%E4%B8%8A>); *Jiu Tangshu*, juan 195 (<https://ctext.org/wiki.pl?if=gb&chapter=295984>).

Xuanwumen Incident (626/6); the Shenlong Revolution (705/1); and the Tanglong Coup (710/6). The absence of occurrences after 710 should be interpreted as a corpus pattern, not as evidence that succession conflict ceased. The series therefore serves as an event-alignment illustration rather than as a fully independent historical validation.

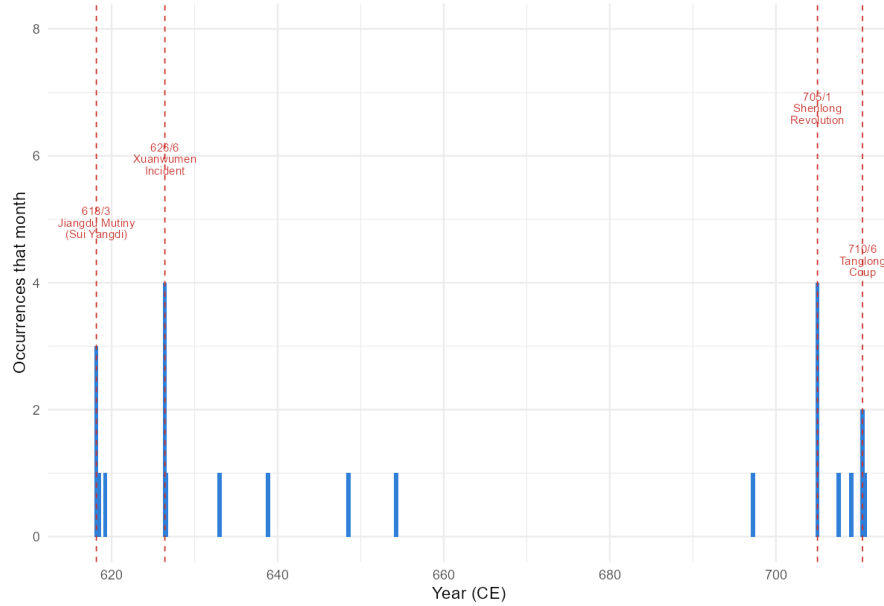


Figure 7. Monthly mentions of Xuanwumen, 618–710

Note: All corpus occurrences fall in 618–710. Dashed lines identify four months with at least two mentions: the Jiangdu Mutiny (618/3), Xuanwumen Incident (626/6), Shenlong Revolution (705/1), and Tanglong Coup (710/6).

The exercise illustrates the temporal resolution of the corpus but is not presented as a general event-detection method: no threshold, model, or systematic search was developed for that purpose. A broader event-detection contribution would require separate design and evaluation.

7. Validation and Robustness

7.1 Chronology reconstruction and iterative audit

Chronology validation was iterative rather than a one-time accuracy check. Using an LLM (ChatGPT, OpenAI GPT-5.6 Sol), the first fixed-seed audit reviewed 80 detected boundaries and 20 unassigned months using expanded annual context. It found 79 correct and one ambiguous detected boundary; among 19 evaluable unassigned months, 16 were correct absences and three were missed boundaries, while one additional case lay in a documented source lacuna. The three errors shared a common XZZTJ continuation-volume-heading mechanism. This diagnosis motivated the general source-specific rule described in Section 3, which recovered those cases without hard-coding them and restored 207 previously empty months. The original 100 cases then served only as a development-set regression test, which the revised parser passed in all 100 cases. Two independent post-revision LLM audits followed. A fixed-seed holdout found 50 of 50 detected boundaries correct and no missed boundary among 45 evaluable unassigned months; five further cells were source lacunae. A third, non-overlapping fixed-

seed sample found all 80 detected boundaries and all 20 unassigned-month checks correct. Across the two independent post-revision LLM samples, detected-boundary precision was therefore 130/130, and the missed-boundary rate was 0/65 among evaluable unassigned cells, with five source-lacuna cases reported separately. All classifications in this iterative audit sequence were made by an LLM, not by the author. Because the detected and unassigned strata were deliberately sampled rather than population-weighted, these figures are audit rates for the reviewed cases, not a single corpus-wide accuracy estimate. Appendix Table A1 summarizes the LLM audit trail.

As a separate reliability check, one author evaluated 30 additional chronology cases while blinded to the LLM classifications and parser-validation outcomes. The sample comprised 15 detected boundaries and 15 unassigned months, drawn from the post-revision independent audit pools after excluding the 100 development cases. Exact human–LLM agreement was 28/30 (93.3%): 15/15 for detected boundaries and 13/15 for unassigned months. The two disagreements concerned months within the documented XZZTJ source lacuna. The human reviewer classified them as correct absences, whereas the LLM classified them as source lacunae; neither disagreement concerned the presence of a missed month boundary. Human judgments were saved before the LLM results were merged. Appendix Table A2 summarizes this blind reliability check.

7.2 Lexical-unit precision

A 209-item sample was drawn from the scored candidate pool, stratified by `lexical_score` into five strata (bottom quartile through top decile), and each item was classified by an LLM as valid, borderline, or a fragment. No author judgment entered these 209 classifications. Figure 8 shows the result: validity rises monotonically from 2.4% in the bottom quartile to 97.6% in the top decile. This monotonic gradient provides direct evidence that `lexical_score` discriminates between LLM-judged valid units and fragments in this sample.

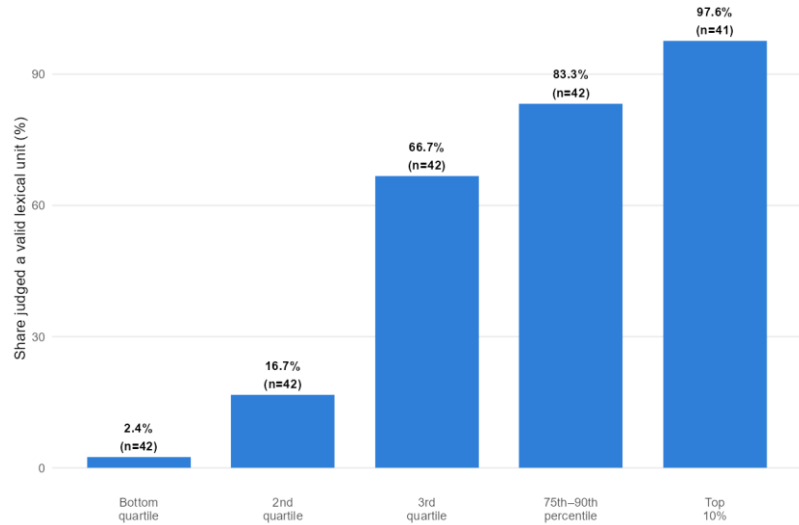


Figure 8. LLM-validation precision by `lexical_score` stratum (Anthropic, Claude Sonnet 5)

Note: The score-stratified sample contains 209 judged candidates. Labels report the valid share and sample size in each lexical-score stratum.

The blind human reliability check additionally sampled 50 lexical candidates, with 10 drawn from each of the same five lexical-score strata. The author saw only the candidate and three to five corpus examples; PMI, entropy, lexical_score, stratum, and all prior LLM judgments were withheld. Exact agreement across the three categories was 42/50 (84.0%; Cohen's $\kappa = 0.694$). When Valid was contrasted with non-Valid (Borderline and Fragment combined), agreement was 43/50 (86.0%; $\kappa = 0.715$). The lower and upper ends of the score distribution each showed 10/10 exact agreement; disagreements were concentrated in intermediate cases involving partial personal names, productive phrases, and the boundary between ordinary compounds and reusable lexical units. Appendix Table A2 reports the combined human reliability results.

7.3 Known-vocabulary recovery

A complementary, independent check asks the opposite question: of vocabulary independently known to belong to Tang institutional and diplomatic life, how much does the pipeline actually recover? A small ($n = 61$), hand-compiled, citation-backed list was traced through the same four-stage funnel as the discovered lexicon itself. This is a known-vocabulary recovery rate rather than recall against a complete gold-standard lexicon of Tang institutional vocabulary, which does not exist in machine-readable form. The list was built from standard reference works – Hucker (1985) for official titles, Twitchett (1970) for fiscal administration, and Twitchett, ed. (1979) for foreign polities and ethnonyms – to provide stable and citable reference points. Reference entries were then expressed in the graph forms actually attested in the Tang ZZTJ subcorpus. In particular, 倭國 (Woguo, the polity later and elsewhere designated 日本, Japan) replaces 日本 because the relevant Tang passages in this corpus use 倭國; and 兩税法 (the Two-Tax Law) retains the corpus form 稅 rather than the normalized traditional graph 稅. The former occurs four times, while 兩税法 occurs three times: twice in the account for the first month of 780, including the description of Yang Yan's reform, and once in the seventh month of 787. Both are therefore attested, although neither reaches the baseline frequency floor of ten. With these source-form corrections, all 61 reference terms are attested in the Tang monthly corpus and 46 enter the final lexicon, giving a recovery rate of 46/61 (75.4%). The category rates in Table 3 use all listed terms in each category as their denominators. Recovery is high for central-government titles, military titles, and foreign-polity names (79.2–100%), and lower for fiscal-administration terms (40.0%) and examination-system terms (66.7%). The weaker recovery in some categories may reflect both the chronicle's narrative register and the pipeline's frequency, length, and boundary choices. Court-political and military terms often appear as recurrent multi-character strings, whereas some fiscal terminology occurs mainly within longer official titles (e.g. 度支 within 戶部度支 or related titles). The correction also illustrates a broader methodological point: a reference item may appear absent when the lookup list silently normalizes a historical or edition-specific spelling. The complete 61-item reference list and each item's passage through the discovery funnel are reported in Appendix Table A8.

Table 3. Known-vocabulary recovery, by category

Category	n terms	Attested	In final lexicon	Overall recovery
Central government titles	13	13	12	92.3%
Regional administration	8	8	6	75.0%
Military titles	3	3	3	100%
Examination system	3	3	2	66.7%
Fiscal administration	10	10	4	40.0%
Foreign / ethnonyms	24	24	19	79.2%

7.4 Threshold sensitivity

Threshold sensitivity was evaluated with a one-at-a-time grid in which each parameter was varied while the remaining settings were held at the baseline (frequency ≥ 10 , minimum-split PMI > 3 , minimum boundary entropy > 1.0 , and an 80% containment threshold for redundant-fragment removal). Appendix Table A3 reports the resulting lexicon size and recovery of the 61-item known-vocabulary list. Lowering the frequency floor to five expands the lexicon from 6,528 to 14,978 units and raises known-vocabulary recovery from 46/61 (75.4%) to 50/61 (82.0%), whereas raising it to twenty reduces the lexicon to 2,843 units and recovery to 39/61 (63.9%). By contrast, reasonable changes to the PMI and entropy floors or to the fragment-removal threshold change lexicon size but leave known-vocabulary recovery within 75.4–77.0%. The substantive recovery result is therefore stable to the cohesion, boundary-freedom, and fragment-removal choices, but—as expected for a frequency-based discovery procedure—more sensitive to the minimum-frequency requirement.

7.5 Cross-period portability and limitations

An out-of-sample portability check examines whether the discovered terms also appear beyond the Tang discovery corpus. The top Tang-discovered terms are almost all attested in the Song-period XZZTJ subcorpus, but at substantially different relative frequencies (節度使 falls from 26.1 to 3.5 per 10,000 characters; 陛下 rises from 13.5 to 20.6). Their attestation supports the cross-period portability of these lexical forms, while the frequency differences indicate period-specific salience. Four independently documented Tang reign-era names (貞觀, 天寶, 元和, and 開元) were checked against the discovered lexicon's own attested date ranges. Three match their documented start years exactly; 開元 appears earlier because the same character string also has the ordinary-language sense "to inaugurate." The attested end years extend beyond the official end of each era because the chronicle can mention reign-era names retrospectively. Table A4 therefore evaluates start-year alignment only and does not treat the last attested year as an estimate of an era's end.

8. Conclusion

This paper develops and validates a transparent and auditable workflow for converting a continuous, digitally unstructured premodern Chinese chronicle into a monthly-resolution temporal-lexical corpus. The contribution lies less in any single computational technique than in combining chronology reconstruction, corpus-internal lexical discovery, monthly measurement, and explicit validation. The resulting representation does not eliminate ambiguities in the source; rather, it makes their treatment

explicit, leaving unresolved temporal assignments unfilled and distinguishing reconstruction problems from documented source lacunae.

The value of computational historical analysis need not rest on discovering facts previously unknown to historians. Recovering well-established historical patterns can itself provide an important validation of a newly constructed corpus. The broader opportunity is to translate qualitative historical propositions into explicit textual expectations and assess their timing, magnitude, persistence, and sequence consistently across long periods. Computational methods thus complement rather than replace historical interpretation: they extend the capacity to search, compare, and measure at scale, while source criticism remains necessary to determine what those patterns mean (Pichler and Reiter 2022).

Such data construction also makes historical sources available for social-scientific analysis. Recent studies of Chinese history illustrate how textual and other historical records can be transformed into geographically and temporally structured evidence for questions about state capacity and long-run institutional development (Keller, Shiue, and Yan 2026; Bai and Jia 2023). Monthly reconstruction creates similar possibilities here. Textual measures can potentially be linked to rulers, political transitions, external shocks, warfare, and other externally dated events, allowing questions about timing and response that annual or unstructured narrative data cannot address as directly.

These possibilities require a clear distinction between textual measurement and the historical phenomena those measures are intended to represent. A rise in the frequency of an institutional title is direct evidence of changing textual salience, but only indirect evidence of institutional change; the latter requires historical interpretation and, where possible, independent evidence. The corpus developed here is therefore best understood as research infrastructure for subsequent historical and social scientific inquiries. The broader progression is from chronicle as text, to chronicle as data, and ultimately to chronicle as evidence: the first transition requires computational reconstruction and explicit representation choices, while the second requires historically informed questions, validation, and research design.

Funding

This work was supported by the Japan Society for the Promotion of Science KAKENHI grant numbers JP24K03158.

References

- Bai, Ying, and Ruixue Jia. 2023. “The Economic Consequences of Political Hierarchy: Evidence from Regime Changes in China, 1000–2000 CE.” *Review of Economics and Statistics* 105 (3): 626–645.
- Bai, Ying, Ruixue Jia, and Jiaojiao Yang. 2024. “Knowledge Suppression and Resilience under Censorship: Three-century *Book Publications in China*.” No. w33258. National Bureau of Economic Research.
- Chen, Song, and Hongsu Wang. 2022. “China Biographical Database (CBDB): A Relational Database for Prosopographical Research of Pre-Modern China.” *Journal of Open Humanities Data* 8: 4.
- Chen, Zhiwu, Cameron Campbell, and Debin Ma. 2026. “Editorial Introduction: Quantitative History of China: State Capacity, Institutions and Development.” In *Quantitative History of China: State Capacity, Institutions and Development*, edited by Zhiwu Chen, Cameron Campbell, and Debin Ma, 1–15. Singapore: Springer.

- De Weerdt, Hilde. 2020. "Creating, Linking, and Analyzing Chinese and Korean Datasets: Digital Text Annotation in MARKUS and COMPARATIVUS." *Journal of Chinese History* 4 (2): 519–527.
- Feng, Haodi, Kang Chen, Xiaotie Deng, and Weimin Zheng. 2004. "Accessor Variety Criteria for Chinese Word Extraction." *Computational Linguistics* 30 (1): 75–93.
- Harvard University, Academia Sinica, and Peking University. 2018. *China Biographical Database (CBDB)*. <https://cbdb.hsites.harvard.edu/>
- Hucker, Charles O. 1985. *A Dictionary of Official Titles in Imperial China*. Stanford, CA: Stanford University Press.
- Jin, Zhihui, and Kumiko Tanaka-Ishii. 2006. "Unsupervised Segmentation of Chinese Text by Use of Branching Entropy." In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 428–435. Sydney: Association for Computational Linguistics.
- Keller, Wolfgang, Carol H. Shiue, and Sen Yan. 2026. "Mining Chinese Historical Sources at Scale: A Machine Learning Approach to Qing State Capacity." *Historical Methods* 59 (2): 75–99.
- Pichler, Axel, and Nils Reiter. 2022. "From Concepts to Texts and Back: Operationalization as a Core Activity of Digital Humanities." *Journal of Cultural Analytics* 7 (4).
- Rappo, Gaëtan. 2018. "Data Mining in the Works of the Shingon Monk Monkan (1278–1357): Using Digital Methods to Assess the Contested Authorship of Three Religious Texts from Medieval Japan." *Journal of the Japanese Association for Digital Humanities* 3 (1): 114–149.
- Tian, Huishuang, Kexin Yang, Dayiheng Liu, and Jiancheng Lv. 2021. "AnchiBERT: A Pre-Trained Model for Ancient Chinese Language Understanding and Generation." In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.
- Tsui, Lik Hang, and Hongsu Wang. 2020. "Harvesting Big Biographical Data for Chinese History: The China Biographical Database (CBDB)." *Journal of Chinese History* 4 (2): 505–511.
- Twitchett, Denis. 1970. *Financial Administration under the T'ang Dynasty*. Cambridge: Cambridge University Press.
- Twitchett, Denis, ed. 1979. *The Cambridge History of China. Vol. 3, Sui and T'ang China, 589–906*. Cambridge: Cambridge University Press.
- Wang, Dongbo, Chang Liu, Zhixiao Zhao, Si Shen, Liu Liu, Bin Li, Haotian Hu, Mengcheng Wu, Litao Lin, Xue Zhao, and Xiyu Wang. 2023. "GujiBERT and GujiGPT: Construction of Intelligent Information Processing Foundation Language Models for Ancient Texts." arXiv:2307.05354.
- Wang, Erik H., and Clair Z. Yang. 2025. *The Political Economy of China's Imperial Examination System*. Elements in Political Economy. Cambridge: Cambridge University Press.
- Wang, Yu-Chun. 2020. "Word Segmentation for Classical Chinese Buddhist Literature." *Journal of the Japanese Association for Digital Humanities* 5 (2): 154–172.

Appendix A. Supplementary Tables and Figures

Table A1. Iterative LLM audit of monthly chronology reconstruction

Audit stage	Sample design	Detected boundaries	Unassigned-month checks	Outcome / use
Version 1: Initial contextual audit	80 + 20	79 correct; 1 ambiguous	16 correct; 3 missed; 1 source lacuna	Diagnosed continuation-volume-heading failure
Version 1: Development regression	Same 100 cases after repair	80 / 80 passed	20 / 20 passed, including lacuna label	Regression only; not independent validation
Version 2: Independent holdout	50 + 50	50 / 50 correct	45 / 45 correct; 5 source lacunae	No error among 95 evaluable cases
Version 3: Independent replication	80 + 20; no overlap with prior 200	80 / 80 correct	20 / 20 correct	No error among 100 evaluable cases
Post-revision independent total	Two independent samples	130 / 130 correct	65 / 65 correct; 5 source lacunae	Detected precision 100%; missed-boundary rate 0% in reviewed evaluable cases

Note: ChatGPT (OpenAI GPT-5.6 Sol) performed all classifications reported in this table. The first sample was used to diagnose and repair the parser and is therefore not independent final validation. Post-revision rates pool the two later, non-overlapping LLM audits by audit type. Source-lacuna cases are excluded from the missed-boundary denominator. The sampling strata were deliberately fixed and are not population-weighted. The separate author-conducted blind check appears in Table A2.

Table A2. Blind human reliability check against LLM classifications

Component	Human sample	Exact agreement	Cohen's κ	Interpretation
Lexical: three categories	50	42/50 (84.0%)	0.694	Valid / Borderline / Fragment
Lexical: binary	50	43/50 (86.0%)	0.715	Valid vs. Borderline + Fragment
Chronology: detected	15	15/15 (100.0%)	—	All detected boundaries agreed
Chronology: unassigned	15	13/15 (86.7%)	—	Two source-lacuna coding disagreements; no

				missed boundary
Chronology: overall	30	28/30 (93.3%)	—	Detected and unassigned questions reported separately in text

Note: The author completed and saved all 80 judgments before LLM classifications were merged. Lexical cases were sampled evenly across five lexical-score strata. Chronology cases were drawn from the independent post-revision pools, excluding the 100 development cases. Cohen’s κ is not reported for the chronology subsets because the human reviewer used a single response category within each subset; agreement rates are more informative in this setting.

Table A3. One-at-a-time sensitivity of lexicon size and known-vocabulary recovery

Specification	Final lexicon	Known terms recovered	Recovery rate
Baseline	6,528	46/61	75.4%
Frequency floor = 5	14,978	50/61	82.0%
Frequency floor = 20	2,843	39/61	63.9%
PMI floor = 2.5	7,337	47/61	77.0%
PMI floor = 3.5	5,725	46/61	75.4%
Entropy floor = 0.5	6,777	46/61	75.4%
Entropy floor = 1.5	6,135	46/61	75.4%
Containment threshold = 70%	6,413	46/61	75.4%
Containment threshold = 90%	6,555	47/61	77.0%
No fragment filter	6,555	47/61	77.0%

Note: Each alternative changes one parameter while holding the others at the baseline. The containment threshold is the share of a shorter candidate’s occurrences accounted for by an accepted longer candidate at either edge; candidates meeting or exceeding that share are removed as redundant fragments.

Table A4. Reign-era names: attested date ranges vs. documented reign dates

Reign era	Documented	Attested	Note
貞觀 (Zhenguan)	627–649	627–848	Start year matches; later retrospective mentions

Reign era	Documented	Attested	Note
開元 (Kaiyuan)	713–741	621–833	Earlier homograph attestations; later retrospective mentions
天寶 (Tianbao)	742–756	742–794	Start year matches; later retrospective mentions
元和 (Yuanhe)	806–820	806–873	Start year matches; later retrospective mentions

Table A5. Top 15 Tang-discovered terms: Tang vs. Song (XZZTJ) frequency

Term	Tang freq	Song freq	Tang per10k	Song per10k
節度使	1993	1439	26.14	3.46
州刺史	1099	1011	14.41	2.43
陛下	1031	8555	13.52	20.58
太子	871	836	11.42	2.01
侍郎	822	1592	10.78	3.83
天下	804	4868	10.54	11.71
將軍	780	685	10.23	1.65
吐蕃	744	48	9.76	0.12
全忠	710	4	9.31	0.01
不可	700	4983	9.18	11.99
不能	670	2737	8.79	6.58
同平章事	617	221	8.09	0.53
中書	613	3185	8.04	7.66
尚書	605	2057	7.93	4.95
上曰	574	2827	7.53	6.80

Table A6. 回紇 and 回鶻: attested year ranges

ngram	First attested year	Last attested year	Total occurrences
回紇	627	788	174

ngram	First attested year	Last attested year	Total occurrences
回鶻	788	902	169

Table A7. Xuanwumen (玄武門; see Section 2): highest-count months

Year	Month	Occurrences	Documented event
626	6	4	Xuanwumen Incident (Chang'an)
705	1	4	Shenlong Revolution (Chang'an)
618	3	3	Jiangdu Mutiny: assassination of Emperor Yang of Sui (Jiangdu)
710	6	2	Tanglong Coup (Chang'an)

Table A8. Reference vocabulary used in the known-vocabulary recovery check

Category	Term	Tang frequency	Attested	In final lexicon	First exclusion stage
Central government titles	中書令	150	Yes	Yes	Included
Central government titles	侍中	159	Yes	Yes	Included
Central government titles	尚書令	25	Yes	Yes	Included
Central government titles	尚書左丞	29	Yes	Yes	Included
Central government titles	尚書右丞	14	Yes	Yes	Included
Central government titles	御史大夫	110	Yes	Yes	Included
Central government titles	御史中丞	83	Yes	Yes	Included
Central government titles	給事中	101	Yes	Yes	Included
Central government titles	諫議大夫	62	Yes	Yes	Included
Central government titles	翰林學士	108	Yes	Yes	Included
Central government titles	太子太傅	4	Yes	No	Frequency floor
Central government titles	太子太保	13	Yes	Yes	Included
Central government titles	太子太師	18	Yes	Yes	Included
Regional administration	節度使	1993	Yes	Yes	Included
Regional administration	觀察使	208	Yes	Yes	Included
Regional administration	都督	347	Yes	Yes	Included
Regional administration	刺史	1370	Yes	No	Fragment filter
Regional administration	都護	10	Yes	Yes	Included
Regional administration	都護府	2	Yes	No	Frequency floor
Regional administration	防禦使	73	Yes	Yes	Included
Regional administration	團練使	35	Yes	Yes	Included
Military titles	大將軍	315	Yes	Yes	Included
Military titles	將軍	780	Yes	Yes	Included
Military titles	兵馬使	222	Yes	Yes	Included
Examination system	進士	50	Yes	Yes	Included
Examination system	明經	15	Yes	Yes	Included
Examination system	制舉	2	Yes	No	Frequency floor

Category	Term	Tang frequency	Attested	In final lexicon	First exclusion stage
Fiscal administration	度支	172	Yes	Yes	Included
Fiscal administration	鹽鐵	92	Yes	Yes	Included
Fiscal administration	轉運使	55	Yes	Yes	Included
Fiscal administration	兩税法	3	Yes	No	Frequency floor
Fiscal administration	租庸調	7	Yes	No	Frequency floor
Fiscal administration	常平倉	2	Yes	No	Frequency floor
Fiscal administration	均田	1	Yes	No	Frequency floor
Fiscal administration	戶部	3	Yes	No	Frequency floor
Fiscal administration	屯田	20	Yes	Yes	Included
Fiscal administration	和糴	9	Yes	No	Frequency floor
Foreign polities / ethnonyms	吐蕃	744	Yes	Yes	Included
Foreign / ethnonyms	突厥	573	Yes	Yes	Included
Foreign / ethnonyms	契丹	140	Yes	Yes	Included
Foreign / ethnonyms	靺鞨	35	Yes	Yes	Included
Foreign / ethnonyms	南詔	126	Yes	Yes	Included
Foreign / ethnonyms	新羅	69	Yes	Yes	Included
Foreign / ethnonyms	渤海	7	Yes	No	Frequency floor
Foreign / ethnonyms	回紇	174	Yes	Yes	Included
Foreign / ethnonyms	回鶻	169	Yes	Yes	Included
Foreign / ethnonyms	百濟	49	Yes	Yes	Included
Foreign / ethnonyms	高麗	39	Yes	Yes	Included
Foreign / ethnonyms	倭國	4	Yes	No	Frequency floor
Foreign / ethnonyms	大食	21	Yes	No	PMI floor
Foreign / ethnonyms	波斯	15	Yes	Yes	Included
Foreign / ethnonyms	天竺	11	Yes	Yes	Included
Foreign / ethnonyms	吐火羅	8	Yes	No	Frequency floor
Foreign / ethnonyms	突騎施	45	Yes	Yes	Included
Foreign / ethnonyms	沙陀	61	Yes	Yes	Included
Foreign / ethnonyms	黠戛斯	24	Yes	Yes	Included
Foreign / ethnonyms	室韋	21	Yes	Yes	Included
Foreign / ethnonyms	龜茲	23	Yes	Yes	Included
Foreign / ethnonyms	疏勒	1	Yes	No	Frequency floor
Foreign / ethnonyms	于闐	22	Yes	Yes	Included
Foreign / ethnonyms	焉耆	36	Yes	Yes	Included

Note: The list was compiled independently of the corpus-statistical discovery results. Central-government, regional-administration, military, and examination terms were selected with reference to Hucker (1985); fiscal and economic terms with reference to Twitchett (1970) and Hucker (1985); and foreign-polity names and ethnonyms with reference to Twitchett, ed. (1979). Entries use the graph forms found in the Tang ZZTJ subcorpus: 兩税法 retains the corpus graph 税 rather than normalized 稅. “First exclusion stage” identifies the first stage at which an item failed to enter the final lexicon; Included

indicates that it survived all stages. Tang frequency is the raw occurrence count in the twelve-cell analytical panel of the ZZTJ Tang subcorpus.

Appendix B. Illustrating Internal Cohesion and Boundary Freedom

The lexical-discovery procedure described in Section 4 combines two complementary forms of distributional evidence: internal cohesion and boundary freedom. This appendix illustrates the intuition behind these measures using examples from the corpus. The purpose is not to provide an alternative segmentation of Classical Chinese, but to clarify how recurrent character sequences are distinguished from frequent fragments of longer expressions.

B.1. Internal cohesion

Internal cohesion asks whether the characters within a candidate occur together more strongly than would be expected from their individual frequencies. For a two-character candidate, this is measured directly by pointwise mutual information (PMI). For candidates of three or four characters, however, there is more than one possible internal boundary. The procedure therefore calculates PMI across every possible binary split and retains the minimum value.

Consider the three-character candidate 節度使 (*jiedushi*, military governor). It has two possible binary splits:

節 | 度使
節度 | 使

A high minimum-split PMI requires both divisions to exhibit strong association. The candidate cannot therefore receive a high cohesion score simply because one pair of adjacent characters is particularly strongly associated. For a four-character candidate, the same principle is applied to all three possible binary splits. This conservative rule treats the weakest internal boundary as the relevant evidence for the cohesion of the candidate as a whole.

Internal cohesion alone, however, cannot determine whether a recurrent string is an independent lexical unit. A substring of a highly frequent longer expression may itself occur frequently and display strong internal association. The sequence 度使, for example, occurs repeatedly because it forms the final two characters of 節度使. A second criterion is therefore needed to distinguish reusable units from recurrent internal fragments.

B.2. Boundary freedom

Boundary freedom asks whether a candidate occurs across varied external contexts. The intuition follows statistical approaches to unsupervised Chinese word extraction and segmentation that use contextual diversity around a character sequence as evidence of its relative independence (Feng et al. 2004; Jin and Tanaka-Ishii 2006).

Corpus-derived KWIC examples illustrate this contrast. Occurrences of 節度使 appear across varied left and right contexts (dates in parentheses indicate the reconstructed chronicle year and month):

置朔方 | 節度使 | 領單于都護府夏鹽等六州 (721/12)

是歲以幽州 | 節度使 | 兼河北採訪處置使 (732/12)

以安祿山為 | 節度使 | 是時天下聲教所被之州 (742/1)

今請除一親王充 | 節度使 | 且令萬榮知留後 (793/12)

These are KWIC excerpts from the Tang-period ZZTJ monthly corpus rather than schematic examples. They show that both the characters preceding the candidate and the material following it vary across occurrences, producing relatively high entropy at both boundaries.

By contrast, the recurrent substring 度使 appears in the same passages as:

置朔方節 | 度使 | 領單于都護府夏鹽等六州 (721/12)

是歲以幽州節 | 度使 | 兼河北採訪處置使 (732/12)

以安祿山為節 | 度使 | 是時天下聲教所被之州 (742/1)

今請除一親王充節 | 度使 | 且令萬榮知留後 (793/12)

Of the 1,995 occurrences of 度使 in the Tang corpus, 1,993 (99.9%) are immediately preceded by 節; the remaining two occur within 支度使. The substring therefore has almost no left-boundary freedom, although its right contexts vary. This helps explain why boundary entropy distinguishes the complete official title 節度使 from the recurrent fragment 度使. As reported in Table 2b, the minimum boundary-entropy score is 5.61 for 節度使 but only 0.01 for 度使. The same contrast appears between 州刺史 (7.62) and its recurrent fragment 州刺 (0.03).

Boundary entropy therefore provides information that frequency and internal association alone cannot supply. A recurrent sequence may be internally cohesive while still being externally constrained as part of a longer expression. Conversely, a sequence may occur in varied contexts without exhibiting strong internal cohesion. Neither property by itself is sufficient for the purpose of lexical discovery used here.

B.3. Combining the two signals

The procedure therefore treats internal cohesion and boundary freedom as complementary rather than competing criteria. PMI asks whether the components within a candidate belong together; boundary entropy asks whether the resulting sequence behaves as a relatively independent unit across different surrounding contexts.

The final lexical_score implements this conjunctive logic. Minimum-split PMI and minimum boundary entropy are first converted to percentile ranks within each n-gram length, and the lower of the two percentile ranks is retained. A candidate therefore receives a high lexical score only when it ranks highly on both internal cohesion and boundary freedom. This deliberately conservative construction prevents an exceptionally high value on one dimension from compensating for weak evidence on the other.

The procedure should consequently be interpreted as identifying recurrent character sequences that behave like reusable lexical units within this corpus, rather than as determining word boundaries in an absolute linguistic sense. This distinction is especially important for an unsegmented premodern corpus,

where the objective is to construct transparent and countable textual units for historical analysis rather than to impose a complete modern linguistic segmentation.

Appendix C. Annual and Monthly Chronology Reconstruction

This appendix records implementation choices that materially affect the temporal corpus but are too detailed for Section 3. It summarizes the rules rather than reproducing the code.

1. Source ingestion and cleaning. Files were ordered by embedded JUAN metadata rather than filename sequence; front matter and noncanonical files were retained in QA logs but excluded from the narrative corpus. Kanripo metadata, page markers, paragraph marks, whitespace, structural headings, and recursively nested parenthetical annotations were removed from the analytical text. An unstripped version was retained for provenance. A minimal 撰 sentinel was preserved at volume openings because it identifies continuation-volume headings needed by the XZZTJ parser.

2. Annual boundaries. Candidate headers combine a Chinese year ordinal with nearby 春 or 正月 and a preceding reign-era name. Era names are normalized only for matching a small set of attested graphic variants; the analytical text retains source spellings. Candidates are resolved as an ordered chronological sequence using exact or fuzzy era matching, continuation of reign-year counts, and guards against backward or implausibly large jumps. Quoted and retrospective dates are not treated as top-level boundaries.

3. Restricted gap filling and annual repairs. If a year remains missing, a second pass searches only between already accepted adjacent boundaries and admits a broader set of month expressions or 改元 evidence. Four Tang years still required documented structural repairs (Table C1); these are exact-anchor or volume-boundary splits, not imputed text. Thus, the revised monthly parser contains no audit-case-specific recovery rule, while the preceding annual stage includes four disclosed exceptions.

Table C1. Documented Tang annual-boundary repairs

Year	Problem	Evidence and treatment
663	Embedded or future-date interference	Split at 三年春正月; end before 麟德元年春正月.
665	Annual material merged with an adjacent year	Split at 二年春正月; end before 乾封元年春正月.
670	Duplicated or mislabelled 咸亨 heading across volumes	Split at the juan 201/202 boundary; assign the first block to 咸亨元年.
696	Multiple era changes and nonstandard opening	Split at 萬歲通天元年 and verify the internal sequence through 臘月, 改元, and 春一月.

Note: Adjacent years were redistributed only as required to conserve the continuous text; execution stops if a required anchor is absent.

4. Month-boundary classification. The parser detects month names with optional season and leap-month prefixes, then rejects matches marked as retrospective, embedded in an explicit different reign-year citation, or inconsistent with non-decreasing narrative time. A following 朔 or ganzhi day within a

short window is retained as supporting evidence rather than a universal requirement. This prevents ranges, duration expressions, quoted precedents, and calendrical discussions from being treated automatically as current narrative dates.

Table C2. Individually documented month-marker exclusions

Source-year	False marker	Reason for exclusion
ZZTJ 731	二八月	Recurring biannual ritual schedule preceding 臣光曰, not narrative month 8.
XZZTJ 1006	三月...改朔	Conditional ritual-calendar rule, not an attested date.
XZZTJ 1007	三年二月...四月	Tang ritual precedent quoted within a Song memorial.
XZZTJ 1038	閏十二月...正月朔	Future-year eclipse prediction.
XZZTJ 1057	正月朔	Eclipse citation keyed to another calendrical year.

Note: These exclusions are applied before thread construction. They remain case-specific because broader rules would also discard genuine dated entries.

5. Secondary narrative threads and intercalary months. When a year contains a later internally ordered sequence of months, a new thread may open only at a calendrically supported marker; each thread must then remain non-decreasing. Noncontiguous segments assigned to the same month are concatenated, while *n_segments* and *n_threads* preserve their structure. Intercalary months are grouped with the corresponding numbered month in the twelve-cell panel, but their flags, text, and character counts are stored separately, allowing later reconstruction of the full chronological sequence, including intercalary months.

6. Assignment, missingness, and invariants. Text preceding the first confirmed month marker is assigned to that first reconstructed month; this preserves annual-opening material but may attach undated seasonal prose to the first explicit month. Years with no defensible month marker are not defaulted to month 1. Empty cells distinguish unresolved or unnarrated months from the documented XZZTJ source lacuna. Character totals must be conserved across annual and monthly segmentation, and all adjacent Tang year boundaries are exported as tail/head pairs for QA.