

# Towards Context-aware Normalization of Variant Characters in Classical Chinese Using Parallel Editions and BERT

Florian Kessler

Friedrich-Alexander-University Erlangen-Nuremberg

florian.kessler@fau.de

## Abstract

For the automatic processing of Classical Chinese texts it is highly desirable to normalize *variant characters*, i.e. characters with different visual forms that are being used to represent the same morpheme, into a single form. However, there are some variant characters that are used interchangeably by some writers but deliberately employed to distinguish between different meanings by others. Hence, in order to avoid losing information in the normalization processes by conflating meaningful distinctions between variants, an intelligent normalization system that takes context into account is needed. Towards the goal of developing such a system, in this study, we describe how a dataset with usage samples of variant characters can be extracted from a corpus of paired editions of multiple texts. Using the dataset, we conduct two experiments, testing whether models can be trained with contextual word embeddings to predict variant characters. The results of the experiments show that while this is often possible for single texts, most conventions learned do not transfer well between documents.

## 1 Introduction

A lack of orthographic norms is a common feature of ancient writing systems. In the case of Classical Chinese, the written language of ancient China, this manifests prominently in a high number of *variant characters* (*yitizi* 異體字), that is in a broad sense, characters that are graphically distinct from each other but are used to write the same morpheme. For many downstream tasks such as full-text search, identification of parallel passages or the analysis of vocabulary, normalization of variant characters is desirable, as often, they are completely interchangeable and merely reflect arbitrary choices of copyists or woodblock carvers. However, there is also a class of *quasi-variant* characters that are only interchangeable in some

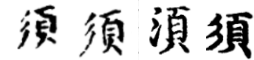


Figure 1: Images of four characters taken from the *Sibu congkan* editions of the *Zhaimin yaoshu* (first and second from the left) and the *Baishi changqing ji* (third and fourth from the left), all representing the same word “*xu* to need”, transcribed as 「*xu* 須」 (first and third from the left) and 「*xu* 須」 (second and fourth from the left) in the digital editions.

contexts, with one variant often being preferred for one of multiple words<sup>1</sup> that can be written with the characters, more strongly associated with a particular word sense, or only found in certain compounds. For example, the two homophonous and etymologically related words “*li* to experience, to undergo” and “*li* calender” should, according to most dictionaries, be written with the two characters 「*li* 歷」<sup>2</sup> and 「*li* 曆」 respectively. While the usage of these two characters in some editions of Classical Chinese texts agrees with this distinction, in others, we find either character used to write both words, or other variant forms such as 「*li* 厯」 replacing them. Hence, a simplistic approach to normalization based on lists of variant characters must either risk conflating variants that were intentionally kept apart such as 「*li* 歷」 and 「*li* 曆」, potentially impacting the understanding of the text, or ignore such cases, which could e.g. mean missing a parallel passage in two texts just because one scribe decided to use 「*li* 厯」 for both “to experience” and “calender”.

<sup>1</sup>Since Classical Chinese is a largely monosyllabic language, most morphemes are also words, so in the following, we will be mostly concerned with words rather than morphemes, although this is of course a simplification.

<sup>2</sup>In order to distinguish between characters and the words they represent, we use English quotation marks “” for our glosses for the latter, and Chinese quotation marks 「」 for the former. To improve readability for readers unfamiliar with Chinese, Pinyin transliterations for both are supplied, although it should be noted that characters can represent multiple words with different pronunciations.

The complex history of many characters further complicates matters, with specialised variants appearing and disappearing and characters being borrowed to write additional words over time (for a detailed overview, see Qiu et al., 2000, Chapters 10-12). Also, in China there traditionally was a taboo on using characters from the ruler’s name, which was sometimes avoided by using existing variant forms or even coining new ones (Wang, 1997, 4)<sup>3</sup>. Furthermore, the physical quality of texts might vary, and OCR systems as well as preferences of human transcribers can have an impact on which variant characters are presented to us in digital versions of the texts. For example, in Figure 1 four characters are shown that represent the same word but are transcribed into two different, but very similar variant forms, which the *Dictionary of Chinese Character Variants* (DCCV) (Ministry of Education, R.O.C) lists as having overlapping but not identical usage. One case matches fine variations in the writing style of the original text while the other appears to be a transcription error. Thus, we anticipate a considerable amount of variability and noise in the data, and it is to be expected that there is no single normal form that “normalization” will result in.

In order to cope with these difficulties, an intelligent system for character normalization should ideally satisfy the following conditions:

1. It should be able to detect in which cases variants are completely interchangeable, and when there is a meaningful difference in their usage.
2. Using that information, when substituting characters to a more regular form, it should do so in the direction of higher differentiation, e.g. replacing 「*li* 歷」 with 「*li* 歷」 or 「*li* 曆」 depending on which word it represents in its specific context.

Towards the development of such a system, in this study, we have extracted a dataset of variant characters in context from a corpus of texts in two editions. In two experiments, we have tested whether contextual word embeddings can be used to train models to predict variant characters.

<sup>3</sup>We would like to thank one of the anonymous reviewers for pointing out the importance of considering this taboo when studying the usage of variant characters.

## 2 Related work

Given the fact that many quasi-variant characters are distinguished from each other by being preferred for specific words or word senses, we expect the problem to be highly similar to word sense disambiguation. For Classical Chinese, Shu et al. (2021) and Pan et al. (2022) have recently used BERT for this task, with some success, although results for some characters were mixed. Our approach of using a parallel corpus has already been successfully applied for learning word sense disambiguation, using alignments of translated sentences (Ng et al., 2003).

Wang et al. (2023) have developed a dataset of loangraphs, i.e. characters used to write a word that is commonly written with another character, and used BERT embeddings to detect them and predict the more usual character for writing the word in question. Many variant characters originate from loangraphs (Qiu et al., 2000, 371-372), and the tasks share the problem of having to decide whether a character should be replaced by another character, so the study is very similar to the subject of this study. However, the authors use a hand-annotated dataset, which compared to ours, has the advantage of higher accuracy, and greater coverage of rare loangraph usage. On the other hand, the number of samples for each type of character is quite limited in comparison to our automatically derived dataset, and since there is no systematic annotation of an entire corpus, it is impossible to quantify how widespread the phenomenon is, and how the usage of loangraphs differs between texts.

A somewhat comparable task for modern Chinese is conversion from simplified to traditional characters, as one simplified character often replaces several traditional ones, such as 「*li* 历」 replacing both 「*li* 歷」 and 「*li* 曆」. Hence, machine learning techniques that take context into account have been investigated for this task (Pang and Yao, 2015). The problem of substituting one character with another, more common character is shared with spelling correction, for which BERT has also been used (Wu et al., 2023). An important difference to modern languages is of course that for Classical Chinese, there is no uniformly accepted normative authority, so it is not *a-priori* clear which character is “correct” in a given context.

For Western languages, normalization of historic spelling variations has been intensively stud-

ied. [Bollmann \(2019\)](#) gives an overview over different techniques, including machine learning approaches. [Jurish \(2010\)](#) and among others more recently [Makarov and Clematide \(2020\)](#) introduce techniques to take context into account to differentiate words, similar to what is attempted here. However, a key difference between alphabetical languages and Chinese is that techniques for the former often rely on edit distances between words, which is not directly transferable to Chinese characters.

### 3 Building a dataset

To the best of our knowledge, no comprehensive annotated corpus to train and test a system for variant normalization exists. However, there is a readily available data source which has high potential: works for which different prints or handwritten editions are digitally available. Since the usage of variant characters can vary considerably between several editions of the same text, aligning them allows for the mining of variant characters. Crucially, using an algorithm that will be described in detail below, we were able to automatically extract instances of variant characters that are used concurrently in one edition but correspond to only a single variant in another, giving potential cases of quasi-variant characters differentiated by one writer but not the other.

#### 3.1 A corpus of parallel editions

All texts used in this study were obtained from two collections of pre-modern Chinese works, the *Wenyuange* copy of the *Siku quanshu* (SKQS), compiled in the late 18th century, and the 1919 edition of the *Sibu congkan* (SBCK), digital versions of which were sourced from the Kanseki repository ([Wittern, 2016](#)). In the repository, there are 286 works with editions from both collections.<sup>4</sup> For our purposes, an important distinction between the two collections is that the editions in the SKQS were produced as the result of an organised editing process over some 15 years in the 18th century ([Guy, 1987](#), 67-120), whereas the SBCK consists of photographic reproductions of older editions from different periods of time, prioritizing early prints where available ([Cui and Wang, 2011](#)). Nevertheless, a comparison of different historical copies of the SKQS has revealed considerable

<sup>4</sup>According to the catalogue of the repository, there should be another 30 parallel editions, but our script failed to retrieve them.

Period	Num. of works	Num. of chars.
Zhou (1046 BC-256 BC)	8	467 505
Qin (221 BC-206 BC)	2	353 858
Han (202 BC-220 AD)	32	2 644 549
Three Kingdoms (220-280)	6	537 497
Jin (265-420)	4	411 061
Northern and Southern Dynasties (420-589)	11	1 665 058
Sui (581-618)	2	84 228
Tang (618-907)	62	7 694 484
Song (960-1279)	87	17 754 465
Yuan (1271-1368)	29	5 938 237
Ming (1368-1644)	10	3 008 885
Qing (1636-1912)	7	1 622 437
Other	8	859 096

Table 1: Composition of the corpus by period assigned in the Kanseki repository, with dates from [Wilkinson \(2018, 4-5\)](#) and length in characters in the SKQS version after truncation.

freedom in the choice of variant characters, which might be attributed to preferences of scribes ([Lan, 2015](#), 49). Hence, the combination of both should give a good overview of variant character usage by different editors or scribes from different times.

Table 1 shows the composition of the corpus by time of origin of the works as recorded in the Kanseki repository. Of course, the editions of the works contained in the repository will often be later.<sup>5</sup> As can be seen, although both collections contain many ancient works, they are in no way exclusively composed of works in Classical Chinese in the strict sense, i.e. the written language of China before ca. 0 AD. Instead, they also contain numerous works from medieval and late imperial China. We expect that the choice of variant characters is often more strongly influenced by the copyists than the original authors of documents, and since the earliest extant editions of ancient texts are often not that ancient, understanding writing conventions of later times is highly relevant to our

<sup>5</sup>Given the high degree of intertextuality present in the corpus, for any given work, significant parts of the textual content might not actually originate from the period assigned in the repository. However, it should at least give a rough approximation.

understanding of ancient texts. Hence, we did not exclude any material based on the time of origin of the work, and have tested as part of the second experiment below whether learning conventions for variant character usage transfers between documents from different time periods.

From the raw text files obtained from the Kanseki repository, all metadata was removed, and all characters that are not Chinese characters deleted. In order to limit the influence of extraordinarily long documents, the length of each text was truncated to 500 000 characters. Afterwards, an optimal global alignment for each pair of editions of the same work was computed using Hirschberg’s algorithm, using the implementation from the Python package `sequence-align` (Kensho Technologies LLC), with gap and mismatch penalty both at  $-1$ , and match score at  $1$ .

### 3.2 Searching for quasi-variant characters

Subsequently, each pair of aligned sequences was searched for potential instances of quasi-variant characters using an algorithm that looks for instances of a single character in one edition corresponding to more than one character in the other edition, applying some frequency thresholds to avoid noise. We exclude cases where more than one of the differentiated characters occurs in both editions (with a small margin of error of a single occurrence), because this indicates either noise or intentional but divergent differentiation by both writers. While including these cases would be interesting for a future study, it was decided to err on the side of caution here and not consider them, reducing the amount of noise in the dataset.

For an aligned pair of sequences  $x = x_1, x_2, \dots, x_n$  and  $y = y_1, y_2, \dots, y_n$ , the algorithm proceeds by the following steps:<sup>6</sup>

1. Let  $C \leftarrow \{c \mid c \neq \square \wedge |\{i \mid x_i = c\}| \geq 100\}$ , the set of all characters that are not the gap character  $\square$  and that occur at least 100 times in  $x$ .
2. For each  $c \in C$  and each  $d \neq \square$ , let  $S_{c,d} \leftarrow \{i \mid 11 \leq i \leq n - 10 \wedge x_i = c \wedge y_i = d \wedge \sum_{j=i-10}^{i+10} \delta_{x_j, y_j} > 10\}$ , the set of all indices where  $c$  is aligned to  $d$  (which might or might not be equal to  $c$ ) in  $y$ , and where at least half the characters in a 21 character

span are equal between the two editions to avoid passages with alignment errors ( $\delta_{x_j, y_j}$  denotes the Kronecker delta taking the value 1 if  $x_j = y_j$  and 0 otherwise).

3. For each  $c$  and  $d$  such that  $|S_{c,d}| < 20$ , let  $S_{c,d} \leftarrow \emptyset$ , deleting substitutions without sufficient support.
4. Return as candidates for quasi-variant characters all  $c, d_1, d_2, \dots, d_k$  with  $k \geq 2$  and the respective indices  $S_{c,d_i}$  such that the  $d_1, \dots, d_k$  are exactly those characters  $d$  for which  $S_{c,d}$  is not empty, and such that at most one of the  $d_1, \dots, d_2$  occurs more than once in  $x$ .

For example, when running the algorithm on the *Xunzi*, an ancient philosophical text, with the SKQS edition as sequence  $x$  and the SBCK edition as sequence  $y$ , we start by collecting in  $C$  a list of characters that occur at least 100 times in the SQKS edition, giving in this case 283 different characters.

Next, in step two, for all the locations where one of these 283 characters occurs in the SQKS edition and where in the surrounding context, a reasonably good alignment was computed by Hirschberg’s algorithm, the two characters in the two editions are recorded in  $S$ . For example, in the *Xunzi*, after this step,  $S_{\text{疆}, \text{疆}}$  contains 241 indices, indicating that for that number of occurrences of 「*qiang/jiang* 疆」 in the SKQS edition, the parallel passages in the SBCK edition have the same character. In  $S_{\text{疆}, \text{強}}$ , there are another 66 indices of passages where the SBCK has 「*qiang/jiang* 強」 instead. This pattern of non-substitution and substitution is potentially relevant for our purposes, as the DCCV lists 「*qiang/jiang* 疆」 as a variant form of 「*qiang/jiang* 強」, but also has a separate entry for it. On the other hand,  $S_{\text{疆}, \text{能}}$  also contains one entry, which in this case corresponds to a specific difference in a single passage between the two editions, which is not relevant for our study.

Hence, in the third step, entries like those in  $S_{\text{疆}, \text{能}}$  with less than 20 indices are deleted from  $S$ .

Finally, in the fourth step, it is checked for which characters from edition  $x$  alignments to more than one character in edition  $y$  are recorded in  $S$ , and whether these characters also occur in  $x$  itself. For the *Xunzi*, at this stage, there are only eleven characters left for which  $S$  contains alignments to more than one character. Out of these, seven are cases where a character in the SKQS edition is aligned to

<sup>6</sup>The implementation of the algorithm as well as all other code used in this paper can be accessed at <https://github.com/notiho/variants>.



two characters in the SBCK, both of which are also used in the SQKS edition. For example, the two visually highly similar variant forms 「*de* 德」 and 「*de* 德」 are both used in both editions. Hence, the indices contained in  $S_{\text{德,德}}$  and  $S_{\text{德,德}}$  are not returned by the algorithm. On the other hand, 「*qiang/jiang* 強」 does not occur in the SKQS edition of the *Xunzi*. Thus, 「*qiang/jiang* 疆」 and its alignments to either itself or 「*qiang/jiang* 強」 are reported as one of the candidates from this invocation of the algorithm.

In general, candidates returned by the algorithm consist of one character that is differentiated into multiple characters in the other edition. In the following, a candidate  $c, d_1, d_2, \dots, d_k$  reported by the algorithm will be referred to as a *substitution profile*  $c \leftrightarrow d_1, d_2, \dots, d_k$ , and the occurrences corresponding to it as samples of that substitution profile from the respective document. Note that a substitution profile may be attested in multiple pairs of editions, but that the samples are specific to each pair.

The algorithm is run on all aligned pairs in both directions, giving 563 substitution profiles. These were filtered to remove all instances where the DCCV lists one of the characters on the right hand side only as a variant of the other character, suggesting that no meaningful difference can be found.<sup>7</sup> For these, an unconditional normalization approach is sufficient. The remaining 108 profiles originate from 103 of the aligned documents, showing as a first result that using more than one variant form of a character is a widespread phenomenon in the corpus.

Table 2 shows four examples from the dataset. The upper two examples display a meaningful distinction between 「*li* 歷」 and 「*li* 曆」, while the lower two are pulled from an edition that arbitrarily uses either 「*mu* 母」 or 「*mu/wu* 毋」 to write “*mu* mother”.

The number of samples per substitution profile ranges from 41 to 5139 (mean 562.6, sd 830.3). On average, each substitution profile is found in

<sup>7</sup>Variants not found in the dictionary, which usually correspond to minor graphical alterations, were also removed. Another two profiles were removed as noise resulting from a difference in how the chapter (*juan*) number is stated in the beginning of each text file. The full unfiltered list can be found in the [supplementary material](#). The filtered version is shown in Appendix A. After inspection of the results, it was further decided to normalize the minor graphical alterations 「*li* 歷」 to 「*li* 歷」 and 「*li* 曆」 to 「*li* 曆」. This allows us to focus on the interesting semantic difference between 「*li* 歷」 and 「*li* 曆」 in the following.

2.2 different documents (sd 2.5), with the highest number of documents for a single profile reaching 14.<sup>8</sup> For some of the profiles, one variant form is highly dominant, accounting for 96.9% of all samples in the most extreme case (mean 71.1%, sd 14.7).

Note that some of the substitution profiles do not consist of variant characters according to the dictionary. For example, we found a profile 留  $\leftrightarrow$  留留, where 「*wan* 留」 is listed as a variant of 「*wan* 畹」 and not 「*liu* 留」. Since they are visually highly similar, this could be an artefact introduced by the digitalization process, for which normalisation is also desirable. The DCCV also has some variant characters with separate entries without noting any difference in usage. For example, in the profile 爾  $\leftrightarrow$  尔爾, 「*er* 尔」 is listed as a variant form of 「*er* 爾」, but also has its own entry, which however only states that it is the same as 「*er* 爾」. Since the first experiment described below is specifically designed to test which profiles represent or do not represent meaningful differentiations in usage, there is no need to remove these cases *a-priori*.

### 3.3 Contextual embeddings

For the 109 profiles found to be potential cases of quasi-variant characters differentiated in one edition but not in the other, contextualised BERT (Devlin et al., 2019) embeddings were collected, which have shown to be useful for a wide variety of tasks (Liu et al., 2019). Specifically, the model from Wang and Ren (2022) was used.<sup>9</sup> Compared to other BERT-family models for Classical Chinese, it has a relatively large vocabulary size of 38 208, making it especially useful for studying variant characters, some of which are quite rare.<sup>10</sup>

For the purposes of the study, we are interested in whether for the substitution profiles, the differentiation on the right hand side is meaningful. We

<sup>8</sup>The profiles with the highest document frequencies highlight the importance of taboo characters, as two of the top-five profiles, 歷  $\leftrightarrow$  曆歷 and 歷  $\leftrightarrow$  曆歷, both involve the character 「*li* 曆」, which was part of the personal name of the Qianlong emperor, under whose reign the SKQS was compiled, and whose name thus had to be avoided by the writers at the time (Wang, 1997, 276).

<sup>9</sup>Obtained from <https://huggingface.co/Jihuai/bert-ancient-chinese>.

<sup>10</sup>In fact, out of the left hand sides of the substitution profiles investigated, which are input into the model, only three characters, 「*chuang* 窻」, 「*chi* 勑」, 「*mao* 貞」, were absent from the vocabulary. Even for these cases, the model still has the context available, so it is in principle capable of computing useful embeddings.

Profile	Edition	Text
厯 ↔ 曆歷	SKQS	非明 厯 理不足與共事
厯 ↔ 曆歷	SBCK	非明 曆 理不足與共事
<b>Translation</b>		If someone doesn't understand the principles of <b>calenders</b> it is not worth making common cause with them.
厯 ↔ 曆歷	SKQS	鄮山昌上人 厯 游諸方獨為此懼
厯 ↔ 曆歷	SBCK	鄮山昌上人 歷 游諸方獨爲此懼
<b>Translation</b>		Chang Shangren from Maoshan has <b>experienced</b> travelling in all the different directions, but was only ever worried over this.
母 ↔ 毋母	SKQS	母 年七十遠在絕域不知死生
母 ↔ 毋母	SBCK	毋 年七十遠在絕域不知死生
<b>Translation</b>		[My] 70 years old <b>mother</b> is far away in an inaccessible place, and [I] don't know whether she is alive or dead.
母 ↔ 毋母	SKQS	父 母 妻子徙日南
母 ↔ 毋母	SBCK	父 母 妻子徙日南
<b>Translation</b>		[Their] fathers, <b>mothers</b> , wives and children were banished to Rinan.

Table 2: Four examples from the dataset, showing passages with relevant context from editions of two works, belonging to two profiles. The relevant characters are highlighted in red in the original text and our translations.

take this to mean that it is in some way predetermined through the context it occurs in. Hence, when there is a meaningful difference, the model should be able to predict the variant used in the edition corresponding to the right hand side of the substitution profile having only seen the undifferentiated version from the left hand side edition. Accordingly, for each substitution profile  $c \leftrightarrow d_1, d_2, \dots, d_k$  only the passages corresponding to the left hand side of the profiles were input into the BERT model. Specifically, for each occurrence of a  $c$  substituted by one of the  $d_1, \dots, d_k$ , the  $c$ , alongside with 200 characters each to the left and right, or less if the end of the document was reached before that, were extracted. The passages were then input the model. Since embeddings produced by different layers can have significantly different performance on various tasks (Liu et al., 2019), the output of all twelve hidden layers was collected to test which gives the best results.

## 4 Experiments

### 4.1 Can conventions in single documents be learned?

In the first experiment, it was tested which substitution profiles in which documents correspond to meaningful differentiations, and which are arbitrary. Since many substitution profiles are attested in more than one document, and it could be the case

that for the same profile, substitutions are purely noise in one document, but meaningful in another, each pair of editions of documents was tested separately. For this purpose, we have fitted a logistic regression on the contextual embeddings computed from the non-differentiated editions, separately for each unique combination of substitution profile and document. If the resulting model is capable of predicting which of the differentiated variants should occur in a particular position, this indicates that the choice is in some way determined, and the differentiation meaningful for that particular set of variant characters in that particular edition.

For evaluation, ten-fold cross-validation was used, that is, for each substitution profile found in each document, the available samples were randomly partitioned into ten parts, and each part held out as test data for a model trained on the remaining nine parts. Following among others Shi et al. (2016), logistic regression was used to fit models on the contextual embeddings. In particular, we used the `sklearn` package (Pedregosa et al., 2011), with L2 regularization and softmax loss for those profiles with more than two alternative variants. After training, the R package `caret` (Kuhn, 2008) was used to test whether the model's predictions on the test set are significantly better than a naive predictor that always predicts the most fre-

	Above naive	Total
<b>Combinations of profile and document</b>	77 (32.4%)	283 (100%)
<b>Profile (at least one document above naive)</b>	45 (41.7%)	108 (100%)

Table 3: Counts of unique combinations of profile and document for which significantly better accuracy compared to a naive classifier was achieved (first row), and of profiles for which this was the case for at least one document (second row).

quent class, at a significance level of 0.05, adjusted for multiple testing with a Bonferroni correction.

The experiment is run twelve times, using the different hidden layers as input. The highest number of combinations of document and profile with prediction significantly better than a naive classifier was achieved when using the output of the final hidden layer, where 77 cases could be found, compared with 69 for the second best, the second-to-last hidden layer. This agrees with the intuition that the problem of predicting the precise variant used in a particular position is highly similar to the masking problem BERT is trained with, in contrast to most other tasks where embeddings taken from middle layers generalize better (Liu et al., 2019).

The results of the experiment, a summary of which is shown in Table 3, indicate that meaningful differentiation of variants is less common than free alteration of variants, even after having filtered out variants that are always interchangeable according to the DCCV as described in Section 3.2. Only for a minority of unique combinations of profile and document the model learns to predict samples significantly better than naively predicting the most frequent class.<sup>11</sup>

Interestingly, the model is able to predict some variants which we would expect to be completely interchangeable based on the DCCV, such as 「er 尔」 and 「er 爾」 described above, albeit only for a single document. A manual investigation of that document, the *Tai ping yulan* reveals that indeed, one of the editions consistently writes the surname “Erzhu” as 「erzhu 尔朱」, but the name of a well-known gloss dictionary, the *Erya*, as 「erya 爾雅」,

<sup>11</sup>A complete list documenting for how many documents this was the case for each profile can be found in Appendix A.

Profile	Pairs above naive classifier / all pairs
厯 ↔ 曆歷	112/132
勅 ↔ 勅勅	0/20
明 ↔ 明明	0/20
歷 ↔ 曆歷	9/12
聲 ↔ 声聲	3/12
于 ↔ 于於	0/2
巳 ↔ 己己	0/2
歷 ↔ 厯歷	0/2
苔 ↔ 答苔	0/2
解 ↔ 解解	0/2
須 ↔ 湏須	0/2
魯 ↔ 嚕魯	1/2

Table 4: Number of directed pairs for which a model trained on the first document was able to achieve performance significantly better than a naive classifier, by substitution profile.

whereas the other edition uses 「er 爾」 for both. Thus, the method has successfully revealed a distinction not found in the dictionary.

The accuracy achieved by the model is difficult to compare between different profiles and documents. For those combinations of profile and document where the accuracy is significantly better than the naive predictor, it ranges from 51.2% to 100% (mean 89.2%, sd 8.7).

#### 4.2 Do conventions transfer between documents?

The first experiment has shown that in principle, a simple logistic model is able to learn to predict differentiated variant characters from contextual embeddings taken from an edition that does not differentiate the variants. However, it was only tested whether this is possible for individual pairs of editions of documents. Hence, the logistic regression could have learned to overfit the conventions of an individual writer, which would not be useful for normalizing other texts. Thus, in a second experiment, we tested whether what was learned on one pair of editions of a document ( $u, v$ ) can be applied to another pair of editions of a different document ( $x, y$ ) that exhibits the same substitution profile.

For this purpose, all profiles were selected where in the first experiment, the model was able to learn to predict variants for more than one document. This was the case for only 12 profiles, which

are listed in Table 4. Then, for each directed pair<sup>12</sup> of documents with above naive classifier performance, each consisting in turn of a pair of aligned editions, a model was fitted with the same basic setup as in the first experiment, using all samples from the first document as training data, and all samples from the second document for testing. That is, for documents  $a$  and  $b$  with aligned editions  $(u, v)$  and  $(x, y)$  respectively, the model is trained to predict the variants in  $v$  based on embeddings taken from  $u$ , and it is test on predicting variants in  $y$  based on embeddings from  $x$ . Finally, it was again tested whether the model has significantly higher accuracy than a naive classifier that always predicts the most frequent variant in the target document, at a significance level of 0.05 with Bonferroni correction. The counts of pairs where this was the case are also shown in Table 4.

As can be seen in the table, for most profiles, a convention learned on one document does not generalize to other documents in most cases. In fact, the only profiles where for a majority of directed pairs, a model trained on one document was successful in predicting variants in the target document were the two profiles having 「*li* 曆」 and 「*li* 歷」 on the right hand side. Other than that, only the profiles 聲 ↔ 声声 and 魯 ↔ 嚙魯 had successful cross-training cases.

This result suggests that for the other profiles, idiosyncrasies rather than universal norms are more frequently found in the corpus. Of course, training on single documents means the model is exposed to only one type of content. To stay with an example from above, although we have only a single document for it, having learned to write the surname Erzhu with 「*er* 尔」 can't be successfully applied to a document that does not mention a person of that name. And even if it does contain that name, it would not necessarily agree in that choice of variant, as historically, there was no general consensus to write that name with 「*er* 尔」. Furthermore, there is also the possibility that for some documents, the model has simply learned to predict patterns in artefacts that are introduced by the digitalization process, which also should not transfer to other documents.

In a similar vein, a manual investigation of the two documents with successful transfer for the profile 聲 ↔ 声声 shows that they share a strong pref-

erence for writing the name of tones, e.g. “*qusheng* departing tone”, with the simplified form 「*sheng* 声」, another convention we do not expect to be widely adopted.

For the only group where a high degree of transferability could be observed, i.e. the two profiles 歷 ↔ 曆歷 and 歷 ↔ 曆歷, time of origin of the works doesn't appear to have an effect on transferability. Using the dating information in the form of dynasties provided by the Kanseki repository, a chi-squared test shows no dependency between documents originating from the same time period and above naive predictor performance of the model ( $\chi^2 = 0.5232$ ,  $df = 1$ ,  $p = .4695$ ). Out of 94 pairs from different dynasties, 81 (86.2%) transferred successfully, whereas for pairs from the same dynasty, it was 40 (80%) out of 50. We take the result to indicate that conventions regarding the use of 「*li* 曆」 and 「*li* 歷」 were quite stable over time. Further research is needed to determine how this relates to the time of origin of editions instead of works.

Accuracy for the models of the same group calculated for each directed pair ranges from 71.8% to 97.8% (mean 88.4%, sd 6.3). A preliminary experiment suggests that accuracy can be much improved by training on more than one document, with mean per-document accuracy for the same set of documents reaching 99.5% (sd 0.9) when dividing the documents randomly into ten parts, using one part for testing and the others for training. We leave it to further studies to investigate how this might be further improved upon.

## 5 Conclusions

In this article, we have demonstrated the general viability of using parallel editions and contextual embeddings for context-aware variant character normalization for Classical Chinese, by showing that a simple logistic model can be trained to predict which of more than one differentiated variants could replace a character in a given context. At the same time, our analysis has also revealed that meaningful variation of variant characters is quite a rare phenomenon, while in the digital editions surveyed, alteration between variant characters without meaningful difference is ubiquitous. This confirms the need for some form of variant normalization. In this regard, the failure of the model to learn to distinguish variants can actually be highly useful, because it can increase confi-

<sup>12</sup>For two documents  $a$  and  $b$ , both  $(a, b)$  and  $(b, a)$  are considered distinct directed pairs.



dence that for those cases, a simple list based normalization approach does not run the risk of losing information.

For those cases where the model was able to learn a differentiation, the results of the second experiment indicate that idiosyncratic usage of variant characters is quite common in the corpus. Training a model on the conventions used by one writer of one edition of a document does often not generalize to other documents. Taken together with the high overall number of variant characters, this confirms that copyists had considerable freedom in choosing variant characters, and highlights the importance of considering the transmission process when reading received versions of ancient texts.

In terms of the two design goals for an intelligent system for variant normalization stated in the introduction, we have achieved more progress towards the first goal. As we have seen with the example of 「*er* 尔」 and 「*er* 爾」, the system has shown itself capable of discovering deliberate variation in a case where we would not expect it to occur based on consulting a dictionary. It could be a worthwhile endeavour to rerun the experiments with the full list of substitution profiles, i.e. without removing instances that are completely interchangeable according to the DCCV, to see how widespread such cases are.

Towards the second goal of normalizing variants towards specialised forms, we have made significant progress only for a single case, 「*li* 曆」 and 「*li* 歷」. In differentiating these two characters, our simple approach that did not require any manually annotated data achieved high accuracy. Since the second experiment has shown an apparent lack of uniform conventions in the usage of many variant characters, further endeavours in this direction will first need to decide which conventions to adopt.

## 6 Limitations

Since we did not systematically compare the original manuscripts or prints with the digitalized editions, for some visually similar variants we do not know whether they are merely the result of inconsistencies in the digitalization process.

Due to the lack of a manually annotated dataset, we do not know how good the recall of our approach of extracting quasi-variant characters from an aligned corpus of parallel editions is. Since the algorithm that computes the list of candidates con-

tains some filters to reduce noise, it might miss cases where a variant only occurs with very low frequency.

The approach we took towards determining whether the variation of variant forms is meaningful or not can only detect differentiations that the BERT model is aware of, and that are encoded in a simple enough way for a logistic model trained on a limited set of data to extract them.

For the cases where the model was able to learn to predict variant characters, we do not know what factors the decisions are based on, and whether a human would find them meaningful.

## References

- Marcel Bollmann. 2019. [A large-scale comparison of historical text normalization systems](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3885–3898, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jianli Cui and Yun Wang. 2011. 《四部丛刊》编纂考略 (Brief Account of an Investigation of the Compilation of the “Sibu congkan”). *Shandong tushuguan xuekan*, 6:102–103.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- R. Kent Guy. 1987. *The Emperor’s Four Treasuries: Scholars and the State in the Late Ch’ien-Lung Era*. Number 129 in Harvard East Asian Monographs. Harvard University Asia Center.
- Bryan Jurish. 2010. More than words: Using token context to improve canonicalization of historical German. *Journal for Language Technology and Computational Linguistics*, 25(1):23–39.
- Kensho Technologies LLC. [Sequence-align: Efficient implementations of Needleman-Wunsch and other sequence alignment algorithms in Rust with Python bindings](#). [https://github.com/kensho-technologies/sequence\\_align](https://github.com/kensho-technologies/sequence_align).
- Max Kuhn. 2008. [Building predictive models in R using the caret package](#). *Journal of statistical software*, 28:1–26.
- Wen-Chin Lan. 2015. [The Collation of Three Versions of Front Annotations of the Siku Quanshu: Based on](#)

- 365 Pieces of Front Annotations. *Journal of Library and Information Studies*, 13(1):33–68.
- Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.
- Peter Makarov and Simon Clematide. 2020. Semi-supervised contextual historical text normalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7284–7295, Online. Association for Computational Linguistics.
- Ministry of Education, R.O.C. Dictionary of Chinese Character Variants. <https://dict.variants.moe.edu.tw/>. Accessed: 3.05.2024.
- Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 455–462, Sapporo, Japan. Association for Computational Linguistics.
- Xiaomeng Pan, Hongfei Wang, Teruaki Oka, and Mamoru Komachi. 2022. Zuo Zhuan Ancient Chinese Dataset for Word Sense Disambiguation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 129–135, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.
- Zhenjun Pang and Tianfang Yao. 2015. Chinese Bilateral Translation between Simplified and Complex-Character Texts based on Conversion Table and Context. In *14th Chinese National Conference on Computational Linguistics*.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Xigui Qiu, Gilbert L. Mattos, and Jerry Norman. 2000. *Chinese Writing*. Number 4 in Early China Special Monograph Series. Society for the Study of Early China, Berkely.
- Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.
- Lei Shu, Yiluan Guo, Huiping Wang, Xuetao Zhang, and Renfen Hu. 2021. 古汉语词义标注语料库的构建及应用研究 (the construction and application of Ancient Chinese corpus with word sense annotation). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 549–563, Huhhot, China. Chinese Information Processing Society of China.
- Pengyu Wang and Zhichen Ren. 2022. The uncertainty-based retrieval framework for Ancient Chinese CWS and POS. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 164–168, Marseille, France. European Language Resources Association.
- Yankun Wang, editor. 1997. 历代避讳字汇典 (*Collection of Taboo Characters of Past Dynasties*). Zhongzhou guji chubanshe, Zhengzhou.
- Zhaoji Wang, Shirui Zhang, Xuetao Zhang, and Renfen Hu. 2023. The Construction and Application of an Ancient Chinese Language Resource on Tongjiazi. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 535–546.
- Endymion Wilkinson. 2018. *Chinese History - A New Manual*, 5. edition edition. Harvard University Asia Center, Cambridge, Massachusetts.
- Christian Wittern. 2016. Kanseki Repository. *CIEAS Research Report 2015*, Special issue:1–80.
- Hongqiu Wu, Shaohua Zhang, Yuchen Zhang, and Hai Zhao. 2023. Rethinking masked language modeling for Chinese spelling correction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10743–10756, Toronto, Canada. Association for Computational Linguistics.

## A List of substitution profiles and results of first experiment

Profile	Docs. above naive classifier / docs.		
明 ↔ 明明	5/14	歷 ↔ 曆歷	12/13
得 ↔ 得得	0/12	歷 ↔ 歷歷	2/9
聲 ↔ 声声	4/9	萬 ↔ 万万	1/7
解 ↔ 解解	2/7	於 ↔ 于於	1/6
歷 ↔ 曆歷	4/6	爾 ↔ 尔爾	1/6
勅 ↔ 勅勅	5/5	玉 ↔ 玉玉	0/5
等 ↔ 等等	1/5	須 ↔ 須須	2/5
丘 ↔ 丘邱	1/4	京 ↔ 京京	0/4
于 ↔ 于於	2/3	已 ↔ 已已	1/3
幸 ↔ 幸幸	0/3	爾 ↔ 尔爾	0/3
盡 ↔ 尽盡	0/3	總 ↔ 摠摠	0/3
遷 ↔ 迁遷	0/3	體 ↔ 体體	1/3

茲 ↔ 兹兹	0/2	厭 ↔ 厭厭	0/2
厯 ↔ 厯厯	1/2	巳 ↔ 己巳	2/2
文 ↔ 文文	0/2	最 ↔ 冦最	0/2
母 ↔ 母母	0/2	筆 ↔ 筆筆	1/2
篇 ↔ 篇篇	1/2	總 ↔ 總總	0/2
荅 ↔ 荅荅	2/2	閒 ↔ 閑閒	0/2
魯 ↔ 魯魯	2/2	貞 ↔ 貌貌	1/1
窻 ↔ 牕窻	0/1	于 ↔ 于於	0/1
亦 ↔ 亦尔	0/1	仙 ↔ 仙僊	0/1
以 ↔ 叭以	0/1	伏 ↔ 伏伏	0/1
元 ↔ 元玄	1/1	充 ↔ 充克	0/1
克 ↔ 克克	0/1	全 ↔ 全訂	1/1
勅 ↔ 勅勅	1/1	勢 ↔ 勢執	0/1
十 ↔ 十卅卅	1/1	合 ↔ 合瑪	1/1
同 ↔ 仝同	0/1	名 ↔ 名構	1/1
名 ↔ 名玄	1/1	在 ↔ 在狂	0/1
多 ↔ 多朶	1/1	己 ↔ 己巳	0/1
弘 ↔ 宏弘	0/1	彊 ↔ 強彊	1/1
憐 ↔ 怜憐	0/1	摠 ↔ 摠摠總	0/1
支 ↔ 支支	0/1	明 ↔ 明明明	1/1
厯 ↔ 厯歷	0/1	望 ↔ 望望	0/1
某 ↔ 厶某	1/1	校 ↔ 校校	0/1
機 ↔ 机機	0/1	檢 ↔ 檢檢	1/1
歸 ↔ 歸歸皈	0/1	注 ↔ 注註	0/1
無 ↔ 无無	0/1	然 ↔ 然狀	0/1
燕 ↔ 燕鷺	0/1	爲 ↔ 為謂	0/1
爾 ↔ 兒尔	1/1	爾 ↔ 尔尔爾	0/1
窓 ↔ 忞窓	0/1	留 ↔ 留留	0/1
痕 ↔ 痕痕	0/1	兒 ↔ 貌貌	1/1
窓 ↔ 忞窓窓	0/1	窓 ↔ 窓窓	0/1
荅 ↔ 荅荅	1/1	總 ↔ 摠摠摠	0/1
脫 ↔ 托脫	1/1	與 ↔ 歟與	1/1
舊 ↔ 旧舊	0/1	苟 ↔ 苟苟	0/1
茂 ↔ 茂茂	1/1	草 ↔ 艸草	0/1
謂 ↔ 爲謂	0/1	貌 ↔ 貞兒	1/1
貌 ↔ 貞兒貌	0/1	貌 ↔ 貌貌	0/1
貌 ↔ 貞兒	1/1	遊 ↔ 游遊	0/1
醫 ↔ 醫醫	0/1	釋 ↔ 釋釋	0/1
野 ↔ 埜野	0/1	鍼 ↔ 針鍼	1/1
閑 ↔ 閑閒	0/1	閑 ↔ 閒閑	1/1
體 ↔ 体躰體	0/1	體 ↔ 体體體	0/1
體 ↔ 躰體	0/1	勅 ↔ 勅勅	0/1